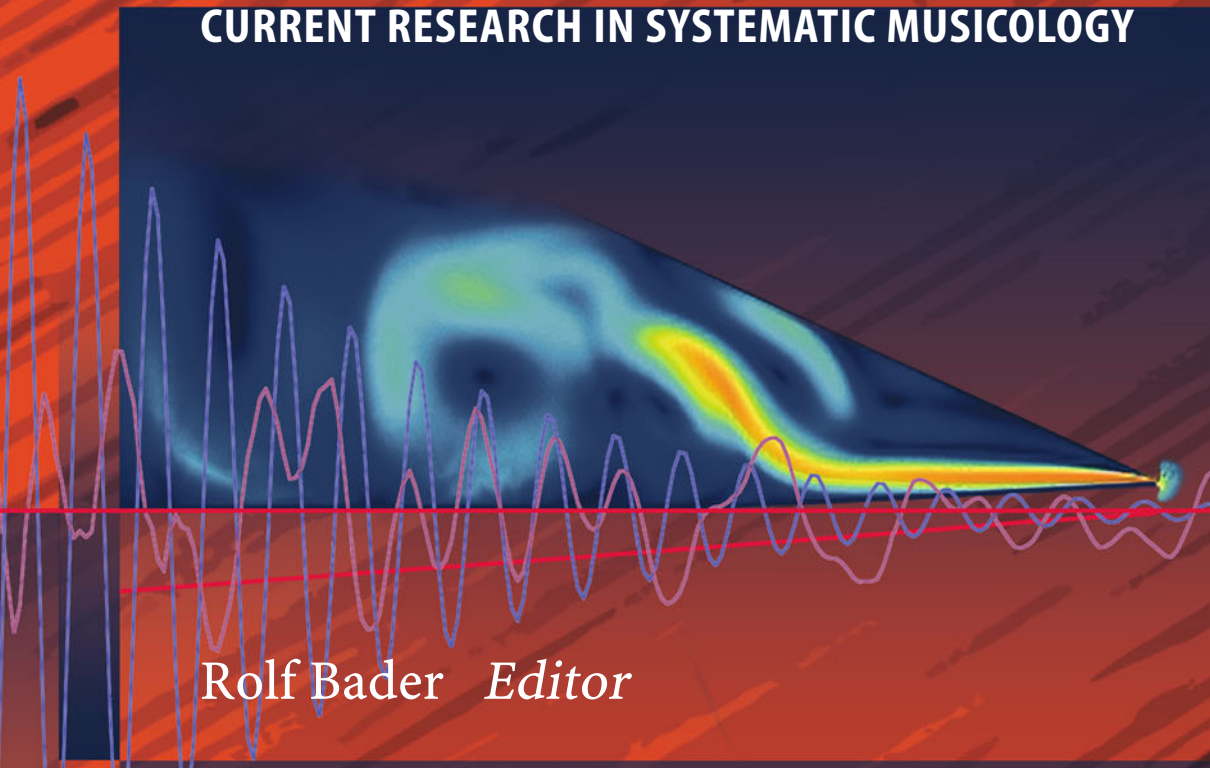
CURRENT RESEARCH IN SYSTEMATIC MUSICOLOGY



Rolf Bader *Editor*

Sound—Perception—Performance

Springer

Current Research in Systematic Musicology

Volume 1

For further volumes:
http://www.springer.com/series/11684

Rolf Bader
Editor

Sound—Perception—Performance

Springer

Editor
Rolf Bader
Institute of Musicology
University of Hamburg
Hamburg
Germany

ISBN 978-3-319-00106-7 ISBN 978-3-319-00107-4 (eBook)
DOI 10.1007/978-3-319-00107-4
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013933578

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Part III Applications

Introduction

Musical performance is based on a variety of conditions, parameters, and systems which allow differentiated articulation, movements, or emotions. This volume is about to discuss basic concepts of performance in the field of Musical Acoustics, Music Psychology, and Music Theory. It also presents recent advances in modern performance environments, algorithms, or hard- and software. The focus is on understanding performance on a basic level of production and perception of musical features, relevant for musicians, composers, or engaged listeners. It suggests systems to understand the framework on which performance in all kinds of music around the world is based on. Therefore, it is asking core questions of Systematic Musicology, namely an understanding of musical performance on a level holding for music generally. The concepts, methods, and applications discussed by the authors are recent advances in the field and cover the wide range of thoughts, experiments, or soft- and hardware used to understand or enhance musical performance practice. The three sections *Production and Perception Models*, *Neurocogition and Evolution*, and *Applications* reflect this approach.

The volume starts suggesting musical instruments and music perception to be based on synchronization, therefore to be synergetic systems. Reviewing the literature suggesting this finding, Rolf Bader discusses nonlinearities of musical instruments, which are often the base for articulation and the musically important features. Furthermore, only the complexity of musical instruments is the reason for their harmonic tone production. Musical performance produced by complex sound production is nearly always based on this complex behavior in terms of synchronization and initial transient production. In terms of Music Psychology, the literature is reviewed about concepts and findings of self-organized systems of perception and production. The paper suggests an impulse pattern formulation (IPF) to hold for all musical instruments as synchronized systems.

Following a complementary reasoning for musical performance, perception, and production, Karl Friston is suggesting a free-energy principle. His model is based on the notion of minimizing surprise, establishing a stable and coherent state after an initial transient phase of tone production, perception, or motion in an interactive way. The difference between the income of a percept and the expected event leads to a readjustment of the system to minimize surprise and end in a stable

state. The idea is exemplified with the production and the perception of a bird song. The system is very flexible as it works as a physical production as well as a perception framework, where motion feedback or changes of physical parameters during perception can be used or left out, leading to estimations about the salience of several performance practices, like gestures or emotional states.

In a review paper, Albrecht Schneider discusses the history of sound analysis with respect to musical acoustics, music perception, and transcription. With an in-depth discussion of the evolution of sound analysis algorithms, methods, and tools he finds many problems still unsolved in a continuous struggle for understanding and producing musical sounds. Starting from a melographic notation of pitch by the notion of periodicity, sound color is discussed in terms of its acoustical and psychological aspects from Chladni and Helmholtz to Stumpf and the notion of 'Ausdehnung' (extension). Formants, as known from speech, have also often been found with musical instruments, still in slightly different forms, not yet understood today and also discussed in the paper, as well as transient sound behavior, most interesting for musical expression and performance.

In the field of music theory, Rolf Inge Godøy is suggesting a basic framework for music experience and performance, consisting of distinct elements he calls Quantals. Following ideas of chunking and concatenation, he finds music happening on three time scales, a micro, meso, and macro scale. Incorporating ideas for sonic objects to be impulsive, sustained or iterative which are also discrete events, the idea of an impulsive nature of musical events is also found on an experimental level with musical gestures, in musical acoustics, and in perception. Especially in terms of body movement, he finds an unequal distribution of attention and effort, where motion is split into key-postures connected by continuous movements. The idea of impulse-driven chunking is proposed, and future research of mathematical formulations is suggested.

The second section about Neurocognition and Evolution starts with a review paper about the foundation of musical emotions by evolutionary aspects, as discussed by Altenmüller, Kopiez, and Grewe. Reviewing the different views of the foundation of music based on language or the survival of the fittest, the paper also discusses archeological artifacts as early as about 35,000 BC to point to a cultural usage of sound in the Neolithic age. Reviewing the literature about music and emotion, in a second part the paper discusses the notion of musical chill, the effect of strong emotional reaction to musical pieces. The neurocognitive findings in this field are presented, and a model is suggested to explain the physiological as well as psychological foundation of the phenomenon of chill.

Focusing on the motion aspect of musical performance, Kölsch and Maidhof present the state-of-the-art of neurocognitive findings for perception–action mediation, starting from the common coding scheme to models of pre-motor area-based music perception. They also discuss neural correlates of music production as used in musical performances. Aspects of anticipation of musical events, the role of mirror neurons, differences between trained versus untrained listeners, the integration of emotional reactions to music, or influences of visual

stimuli are discussed in detail, often using pianists as subjects. Perception errors are found suitable to indicate perception of performance and action data. The investigations clearly show a strong connection between perception and action, although the subject is still strongly under debate.

Särkämö, Tervaniemi, and Huotilainen then focus on the therapeutic use of music in terms of neurocognition, where disorders like amusia, autism, depression, schizophrenia, or strokes are treated. Also diseases like Parkinson or the loss of speech may be approached by music therapy on a neurocognitive basis toward improvement of performance, both on haptic or motion, as well as on the speech production side. Additionally, music is used in everyday life as a mood enhancing inspiration, which also needs explanation on a neural level. The paper gives a review on the subject, starting with the healthy brain only then to cover disorders and discuss methods of music therapy and rehabilitation. Although many aspects are still unknown, several approaches are already successful and promising for the future.

Another important aspect of music perception is syntax, phrasing, and contour of musical pieces, which Neuhaus addresses again in terms of neural correlates. After a discussion about the literature in terms of Gestalt psychology, ideas very close and even developed with musical examples from the start, the paper develops a theory of music segmentation based on mismatch-negativity EEG experiments performed on musical phrases. It appears that several of the Gestalt phenomena are found and are similar to language, while others deviate from the syntax used in speech communication. Still, similar brain regions are responsible for understanding syntax of music and language, which also holds for the neural correlates associated with segmentation and phrasing. Therefore, musical performance on a syntactic level can be formulated on a neural level.

The section about applications to musical performance starts with the presentation of a virtual reality for room acoustic simulations built as ‘the Cave’ at the RWTH Aachen, presented by Vorländer, Pelzer, and Wefers. The application allows performance of virtual musicians in a virtual concert hall, which can be built on the scratch by architects, who right away can listen to music performed in the environment they only just design. The system is based on the room acoustic model of ray tracing, estimating the impulse response of the room in real time. Also in real time is the convolution with virtual performers on stage, which are built as complex radiators. The listener can enjoy a virtual performance in a virtual performance space. This spectacular environment is able to predict and perform the acoustics of concert halls and therefore is a valuable tool for architects.

Modern performance spaces include complex sound systems, which may be ambisonic or surround. The latest development of such systems is the wave-field synthesis presented as an application at the University of Applied Sciences in Hamburg by Fohl. The system is able to reproduce sound spaces, and therefore also virtual concert halls, classrooms, or any kind of artificial acoustic environment. The paper discusses the basic principle of wave-field synthesis, next to a detailed discussion of the system used. Also a motion tracking system is presented to retrieve the position of a person in the sound filed. It then focuses on applications of the

system as built at the institute, which are gesture based, simulating concert or classrooms, and discussing rendering software modifications. The system is in use in a lab environment, still nowadays, wave-field synthesis is also used by composers and discos, too.

Understanding musical structure and texture is a major performance task for listeners. Building environments which perform this understanding successfully has always forwarded our understanding of music in an analysis-by-synthesis methodology. Braasch presents a virtual musician with which one can freely improvise. The system understands the performance of the co-musicians to then play sounds suiting the overall performance. The basis for this understanding is a Hidden Markov Model, analyzing the performance to calculate transition probabilities between performance stages as hidden layers. After this understanding process, the hidden layer structure can then be used to perform music in association to the co-musicians playing along. The performance of the model allows a musically satisfying and interesting performance.

The performance of traditional musical instruments has lead to numerous modifications of the geometry and materials of these instruments. Mores investigates prominent experimental violins as found in the Hannenforth collection of musical instruments. After an introduction to the acoustics of the violin with its basic properties, he compares the impulse responses of several violins in the collection with a Stradivarius violin judged as a high quality example. Most violins, like the 1820 Channot or the Zoller bottle-shaped violin, show different sound holes, both in shape as well as in distribution, while others, like the Philomele or the 1836 Howell, or most prominent the 'grammophone' Stroh violin have considerable different body shapes, with corresponding spectral changes, improving, or worsening the violin sound.

Also algorithms have been proposed to analyze sounds and therefore musical performances to quite an extend. Schneider and Mores discuss the use of several kernels of Fourier transforms in terms of their usefulness to musical sound analysis. After a discussion of the basic problem of the time/frequency uncertainty principle, examples of 3D-Fourier, gammatone filter bank, or autoregressive models are presented. The paper is then about to discuss the Fourier-time transform (FTT) as a method proposed to come as close as possible to human perception, discussing the math and perception tasks. It proceeds to a discussion about the relation between the FTT and Wavelet transforms, exemplified by an organ pipe sound. The paper concludes that the FTT model to be useful with some restrictions and proposes Wavelets or advanced methods for musical sound analysis.

Musical performance strongly depends on the musical instruments used. Physical modeling methods are now able to produce sounds in real time, using whole-body geometries on a field-programmable gate array (FPGA) hardware, which calculates massively parallel. Pfeifle and Bader present a performance tool for controlling such virtual instruments, like banjos, violins, or pianos, using a software tool. Therefore, the FPGA hardware is implemented on a board with a PCIe interface, inserted in a standard personal computer. The interchange of

real-time calculated sound data from the FPGA to the controller, as well as the flow of performance controller data to the physical model is discussed in detail. The system is able to change e.g., the geometry of instruments while playing and therefore offers new kinds of possibilities of performance.

Another way to speed-up performance using physical modeling of musical instruments is the use of advanced algorithms for solving the differential equation system. Pfeifle presents the idea of pseudo-Fourier techniques, already proposed in the 80th, to such physical modeling solutions. Therefore, the iterative process of calculating new displacements and velocities of a vibrating body from the previous state is no longer performed in the spatial dimension but rather in the Fourier domain of this space. Then the convolutions involved in the finite-difference model becomes a multiplication which speeds-up computation time tremendously. Although still real-time performance is not achieved on a standard PC, this method is suitable to speed-up tests of models in software like Matlab or Mathematica.

Performance of virtual instruments also need sophisticated controllers varying the parameters of the sound generating algorithms. Rosa-Pujazón, Barbancho, Tardón, and Barbancho present an overview about sensors and sensor techniques used in the field. Then the paper discusses motion tracking system for musical gestures, both from the hardware and the software side, where the recorded data are reduced to retrieve useful information for controlling. Different applications for the system are discussed, including descending and linear prediction of positions. Different gestures are retrieved by the system used for musical parameter changes. A real-time motion-based composition is shown, where a user can compose using gestures. The system therefore performs music solely by recorded camera gestures without the need to touch any musical instrument.

The papers presented show the high complexity of musical performance and the need for sophisticated methods, algorithms, and hard- and software to understand and to perform music in all its aspects. Our sensibility to slight sound changes which are meaningful to us is therefore the basis of the richness of music and musical performance. This sensibility makes the art complex and makes the field so differentiated and fascinating. In all fields presented, research need to proceed to come to a point where its fine structure is able to meet musical perception and music performance we know and enjoy.

Hamburg, January 2013 Rolf Bader

About the Authors

Eckart Altenmüller holds a Masters degree in Classical flute, and a MD degree in Neurology and Neurophysiology. Since 1994 he is chair and director of the Institute of Music Physiology and Musicians' Medicine at the Hannover University of Music, Drama, and Media. He continues research into the neurobiology of emotions and into movement disorders in musicians as well as motor, auditory, and sensory learning. From 2005 to 2011 he was President of the German Society of Music Physiology and Musicians' Medicine.

Rolf Bader is Professor for Systematic Musicology at the University of Hamburg, Germany. He received a Master, Ph.D., and Habilitation in the field about topics of Musical Acoustics, especially Physical Modeling of Musical Instruments, Musical Signal Processing, and Music Psychology and Music Theory. His teaching includes lectures about Physical Modeling of Instruments as a visiting scholar at the University of Stanford, California. His fieldwork as an Ethnomusicologist focus mainly on Southeast Asia, where he recorded and collected material in Bali, Nepal, Thailand, or Cambodia. He published several books on Musical Acoustics and Music Psychology, both as an author and editor. As a musician he studied Jazz and Classical guitar, Jazz and classical composition, violin, piano, and several non-Western instruments, recorded CDs and played in concert halls, clubs, and streets.

Ana M. Barbancho received the degree in telecommunications engineering and the Ph.D. degree from University of Málaga, Málaga, Spain, in 2000 and 2006, respectively. In 2001, she also received the degree in solfeo teaching from the Málaga Conservatoire of Music. Since 2000, she has been with the Department of Communications Engineering, University of Málaga, as an Assistant and then Associate Professor. Her research interests include musical acoustics, digital signal processing, new educational methods, and mobile communications. Dr. Barbancho was awarded with the "Second National University Prize to the Best Scholar 1999/2000" by the Ministry of Education of Spain in 2000 and with the "Extraordinary Ph.D. Thesis Prize" of ETSI Telecomunicación of University of Málaga in 2007.

Isabel Barbancho received the degree in telecommunications engineering and the Ph.D. degree from University of Málaga (UMA), Málaga, Spain, in 1993 and 1998, respectively. In 1994, she also received her degree in piano teaching from the Málaga Conservatoire of Music. Since 1994, she has been with the Department of Communications Engineering, UMA, as an Assistant and then Associate Professor. She has been the main researcher in several research projects on polyphonic transcription, optical music recognition, music information retrieval, and intelligent content management. Her research interests include musical acoustics, signal processing, multimedia applications, audio content analysis, and serious games. Dr. Barbancho received the Severo Ochoa Award in Science and Technology, Ateneo de Málaga-UMA in 2009 and the Málaga de Investigación 2011 Award from Academia Malagueña de Ciencias y la Real Academia de Bellas Artes de San Telmo.

Rolf Inge Godøy is professor of music theory at the Department of Musicology, University of Oslo. His main research interest is in phenomenological approaches to music theory, meaning exploring mental images of musical sound and corresponding body motions and acoustic features.

Oliver Grewe holds a diploma in biology and a master in musicology. He received his Ph.D. in neurosciences at the Institute of Music Physiology and Musicians' Medicine at the Hannover University of Music, Drama, and Media. After his postdoc at the same institute he switched to science management and works currently as a program director for biomedicine at the Volkswagen Foundation in Hannover.

Reinhard Kopiez received a degree in classical guitar, and a Masters and Ph.D. in musicology. He is professor of music psychology at the Hanover University of Music, Drama and Media, Germany and head of the Hanover Music Lab. From 2009 to 2012 he was president of ESCOM. His latest journal publications concern psychological research on music performance, historiometric analyses of Clara Schumann's repertoire, and music and emotion. Together with A. C. Lehmann and H. Bruhn he edited the German standard handbook on music psychology (Musikpsychologie. Das neue Handbuch, 2008, Rowohlt). From 2013 he is Editor of the Journal Musicae Scientiae.

Robert Mores is Professor at the University of Applied Sciences in Hamburg, where he teaches Telecommunications, Digital Signal Processing, and Acoustics. He received a Diplome in Electrical Engineering in 1988 and a Ph.D. in Computer Science from the De Montfort University Leicester, UK, in 1994. His research is partially devoted to musical acoustics and he plays the violin and the cello.

Christiane Neuhaus worked for 7 years as a postdoctoral research fellow at the Max Planck Institute for Human Cognitive and Brain Sciences in Leipzig. She completed her habilitation thesis in 2011 and recently had a one-year temporary position as a full-time professor of systematic musicology at the University of Hamburg.

Alejandro Rosa-Pujazón was born in Málaga, Spain. He received the B.E. degree in telecommunication engineering from the University of Málaga, Málaga, Spain, in 2006, and the M.Tech. degree in telecommunication engineering from the University of Málaga, Málaga, Spain, in 2008. In 2008, he joined the Department of Electrical Technology, University of Málaga, as a Research Assistant. Since February 2012, he has been with the Department of Communication Engineering, University of Málaga, where he is currently working as a Research Assistant while performing Ph.D. studies. His current research interests include signal processing, human–computer interfaces, and interaction with music.

Lorenzo J. Tardón received the degree in telecommunications engineering from University of Valladolid, Valladolid, Spain, in 1995 and the Ph.D. degree from Polytechnic University of Madrid, Madrid, Spain, in 1999. In 1999 he worked for ISDEFE on air traffic control systems at Madrid-Barajas Airport and for Lucent Microelectronics on systems management. Since November 1999, he has been with the Department of Communications Engineering, University of Málaga, Málaga, Spain. He is currently the Head of the Application of Information and Communications Technologies (ATIC) research group. He has worked as main researcher in several projects on music analysis. His research interests include digital signal processing, serious games, audio signal processing, and pattern recognition. Dr. Tardon received the Málaga de Investigación 2011 Award from Academia Malagueña de Ciencias y la Real Academia de Bellas Artes de San Telmo.

Michael Vorländer is Professor at the Institute of Technical Acoustics at the RWTH Aachen University. The institute is interested in the physical and technical aspects of sound and in the effects of sound on humans. Main topics are: Virtual Reality as Simulation and Auralisation, Computational Acoustics, Room and Building Acoustics, Binaural Hearing, Audiology, Psychoacoustics, Electroacoustics, Digital Technology, Acoustical Measurement Technique, Vibroacoustics, and Automotive Acoustics. Physical and technical acoustics is usually regarded as a subtopic of physics or engineering. Still, acousticians often face questions about the perception of sound. In teaching and research, the institute works on a large variety of different aspects which are frequently found in the daily work of acousticians.

Part I
Production and Perception Models

Synchronization and Self-Organization as Basis of Musical Performance, Sound Production, and Perception

Rolf Bader

1 Introduction

This chapter is about to show that nearly all processes in music, may they be physical in terms of musical acoustics, may they be psychological in terms of psychoacoustics, perception, and music production are highly complex and non-linear in nature. It further argues that the simple, harmonic or linear output of these systems is only caused by self-organization of subsystems which fuse to a global system by nonlinear coupling of these subsystems. The linear approaches known in acoustics and psychology may be a first approximation to many problem. Still when examining the systems closer, many problems occur, and the simple models can no longer explain the phenomena appearing in music.

This view is contradicting traditional assumptions. In ancient music theory, the simple numerical relations of Pure Tone or Pythagorean temperament are also found in cosmic dimensions, as well as in nature in general. So e.g. in his dialog *Timaeus* Plato derives the cosmos from powers of two and three like x^2 and x^3 (Platon et al. 1997). With x = 1, 2, 3, 4, the numbers 1, 4, 9, 16 and 1, 8, 27, 64 are built, from which all musical intervals, like the octave 2:1, the fifth 3:2, the fourth 4:3, etc. and even the whole tone step 9:8 can be derived. Although not explicitly related to music by Plato, this cosmology is generally assumed to be derived from musical harmonies. Explicitly, this relation is found in the *Dream of Scipio Africanus* by Cicero, the earliest source describing the relation of the planets with harmonic numbers. This spirit is continued in Renaissance times in the Quadrivium, consisting of geometry, arithmetic, astronomy, and music, maybe most prominent in the *Harmonia Mundi* of Johannes Kepler (Kepler and Caspar 1997). There he derives his music theory not using the traditional one-dimensional string of a monochord, but by a two-dimensional geometrical representation of polygons and 'stars' which allows him semantic differentiations, e.g. between the minor an the major third. In his earlier

R. Bader (✉)
Institute of Musicology, University of Hamburg, Hamburg, Germany
e-mail: R_Bader@t-online.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_1,

writing *Mysterium Cosmographicum* he also refers to the harmonic relations of the solar system, still there assuming that the distances between the planets are like those of the five perfect Platonic bodies inscribed one into another (Kepler et al. 1981). Interestingly, he also justifies this finding by a dream or a dream-like vision he had lying below a tree, similar to the dream of Scipio. Still this idea turned out to be physically wrong, the planets do not show such distant relations.

The fact that music theory is often not in accordance to music practice has often been mentioned before. Still it is reasonably assumed that the harmonic relations between tones produced on a string or an air column are caused by the linearity of these systems, which therefore follow simple differential equations, the string equation or the Helmholtz equation. Their solutions result in harmonic relations, and therefore the harmonies present with most musical instruments are taken to be a natural result of the moving bodies. It is also assumed that this holds with perception. Tonal fusion, the cycle of fifth, timbre perception, or body movements in music production are mostly explained using simple, linear models, and are therefore taken to result from bottom–up processes. Only higher musical features like semantic, synaesthetic, or cognitive perception of music theory is often assumed to imply also top–down processes.

Still, as discussed below, this simple explanation of musical systems do not hold. So in a way the ancient idea of music as an audible output of simple relations in nature is misleading, and harmony is only an output of the synchronization of otherwise inharmonic motions. Perception of harmony and simple rhythm production is only the output of self-organized systems of nonlinear coupled subsystems. Without these couplings and nonlinearities, the output of the systems would not be harmonic at all most of the time, but mistuned, out of straight rhythmic relations, or not fusing to a whole, a musical Gestalt.

The chapter is organized in three sections. First, musical acoustics is discussed in terms of simple examples of musical instruments, showing that most of them are self-organized systems. As clear experimental evidence is facing mathematical formulations able to explain the findings, this section is about to discuss musical instruments as self-organized systems intrinsically. The second section is a mathematical framework to simplify the basic concept of musical instruments as self-organized systems, an Impulse Pattern Formulation, which is able to explain basic features of musical instruments, especially their initial transients, or make instrument families comparable. The model, as most of the synchronization models discussed below, is able to get from basic physical parameters to global phenotypes within one single step. Self-organizing models often do not think of the traditional bottom–up/top–down interaction, but view the system as a whole, and therefore explain sound production and perception not as a several-stage hierarchical workflow but as one dynamic system leading to a musical Gestalt immediately. In a third section about music psychology, evidence for self-organization and synchronization for several psychoacoustic features are given. As we cannot look into the neuronal networks of the brain in real time to a great extend, the evidence here is one of showing that the nonlinear models of perception are often much better able to explain features in perception and music production than linear models can cover.

2 Musical Acoustics

Nonlinearities in general are often found with musical instruments. Some of them are only add-ons of a basically linear behaviour of the vibrator. So e.g. the brassiness of trumpets and trombones is caused by a shock-wave formation of the travelling wave-front in the tube, which steepens along its way (Hirschberg et al. 1996). This change of the wave-front geometry means an increase of the amplitudes of higher overtones. The effect only appears with loud tones, as only there the sound pressure level in the tube is so strong that the air no longer behaves linear.

Another example are longitudinal waves in strings with the piano, guitar, or harpsichord (Beurmann and Schneider 2009; Bank and Sujbert 2005). Contrary to transverse waves which are the displacement of the string normal to its length, in the plucking or striking direction, longitudinal waves are the compression or expansion of the string along its length. These waves are normally about 10–15 times faster than transverse waves, and therefore have quite high frequencies. Under certain circumstances, when the bridge of guitars or keyboards are constructed in a way to transfer energy from the in-plane direction of the string, and therefore from the longitudinal waves into the bending direction of the soundboard, the longitudinal waves are also radiated and contribute to the overall sound of these instruments. Due to their energy in the high frequency domain they contribute much to the brightness of these instruments.

Also the noise appearing with flutes is caused by chaotic, nonlinear behavior of the air flow of a player onto the labium of the flute (Howe 1975). There a self-sustained oscillation takes place, producing a periodic sound (Coltman 1968a, b), [for a review see (Fabre et al. 2012)]. Still as the air is changing direction from into the flute to outside of it into the surrounding air in a periodical manner, this fast change in direction is causing vortices in the surrounding of the embouchure hole which are radiated as noise. This noise is an add-on to the periodic sound of the flute caused by the nonlinearities in the flow equations and is very characteristic to the flute.

Yet another example may be the brightness of radiating plates under tension, like is the back, and sometimes also the top plate of the guitar (Bader 2005a) or the soundboard of the piano (Mamou-Mani et al. 2008). Such a plate shows an increase of its eigenvalues compared to a plane plate, most noticeably heard with the singing saw. Still when the plate is only radiating the forced oscillations of a string, the increase of these eigenmodes may not be audible, as the strings are the cause of the sound. Still additionally, in plates two waveforms are present at the same time, transversal and longitudinal waves, very similar to the ones discussed above with strings. In a curved plate these wave types are coupling one to another to quite an extend. Energy from the high-frequency longitudinal, or in-plane waves flows into the transversal waves, which are normally much lower in frequency. Still only the transversal waves are radiated to a considerable amount, as the radiating surface is large. The radiating surface of the in-plane waves is only the

small boundary around the corners of the plate. Therefore, without any coupling, the longitudinal waves in the plate are not radiated and do not contribute to the sound. Still due to the curvature of the plate, the in-plane waves are radiated via the transversal waves into which they couple and transfer energy. Therefore a curved plate is brighter in sound.

Many other examples of nonlinearities in musical instruments sound production adding a characteristic component to the sound could be mentioned. Most of them contribute some and often very much brightness to the sound, some add noise or characteristic frequencies, amplitude fluctuations or the like. One might state that very often the really interesting aspects of musical instruments are caused by the nonlinear components of their sound production.

Still this chapter is about to discuss nonlinearities and synchronization not only as an add-on feature to generally linear musical instrument sound production, but as their crucial part. The chapter argues that nearly all musical instruments are self-organized systems driven by strong and characteristic nonlinearities. Only because of these nonlinearities synchronization appears with these instruments at all, and only because of this synchronization musical instruments show such precise harmonic overtone structure. If the systems were purely linear, most of the instruments, like violins, guitars, flutes, or saxophones would

- not have harmonic overtone structures,
- not show such different playing regimes, from noise production over harmonic series to bifurcation, and quasi-chaotic regimes as is present within musical instrument performances, and
- not show regions of such high stability of sound production over a wide range of control parameter changes, like blowing pressure of clarinets or plucking strength of string instruments.

The field of synchronization of musical instruments could also be viewed in the framework of Synergetics (Haken 1990). Within this formulation a multitude of nonlinear coupled subsystems may lead to a slaving of these subsystems by one of them. So within the struggle between the subsystems, one may win and slave the others, force them to go with its frequency, amplitude, or change in parameter. In such a case the system is acting as one synergetic unit. A very simple example is one of two subsystems f and g with temporal derivatives $\dot{f}$ and $\dot{g}$ respectively, acting upon themselves via strengths a and b and coupling onto another in a nonlinear way via coupling constants k_1 and k_2 like

$$\dot{f} = af + k_1 g \tag{1}$$

$$\dot{g} = bg + k_2 f^3 \ . \tag{2}$$

In case a $\gg$ b the second equation simplifies, and we have

$$\dot{f} = af + k_1 g \tag{3}$$

$$\dot{g} = k_2 f^3 \ . \tag{4}$$

Then the change in g only depends on the state of f and no longer on the state of g. Therefore the system g is moving along with the system f and is slaved by it. This reasoning was already proposed by (Aschoff 1936) when discussing saxophones, where the reed is forced to vibrate with the frequency of the tube, necessary to play pitches on the instrument. Aschoff argues that this is caused by the higher damping of the reed compared to the air in the tube (Aschoff 1936). So in our example, if f is the air column and g is the reed, a $\gg$ b holds and the reed g is forced to vibrate with the air column f.

There are several reasons for such a slaving. To the well-established ones

- the system which is less damped takes over the stronger damped system, and
- the system with higher eigenvalues takes over the system with lower eigenvalue,

we will argue below that a third one holds for some musical instruments,

- the system which is lower in spatial dimension takes over the one with higher dimensions, so e.g. a one-dimensional system takes over a two- or three-dimensional system.

Of course this list of reasons for a take-over is not complete, depending on the instrument, other mechanisms may be possible, too, like e.g. sets of initial conditions or special kinds of nonlinearities.

To illustrate the point, first a simple linear system is presented. Then frameworks for nonlinear, synchronizing systems are discussed, two already present in the literature and a new one, an Impulse Pattern Formulation (IPF), which is discussed in some detail. Then the guitar and the organ pipe as examples for synchronized systems are discussed in their basic behaviour. As initial transients are crucial parts of musical instruments, for the guitar as an example it is found that the basic behaviour of the instruments within its initial transient can be simulated using the IPF. The violin string/bow coupling as another synergetic, self-organizing system is not discussed here, as it is much better known, and because of lack of space. [For a detailed description of the violin and of other systems in terms of their physical parameter settings as well as for the IPF solutions see (Bader 2013)].

To get an idea of how complex musical instrument behaviour can be in terms of the slaving principle, Table 1 shows the generator and resonator parts for several musical instruments. In a traditional view the generator is supplying energy which is then radiated into the surrounding air by the radiator. Still it is interesting to see that not in all cases the system vibrates with the frequencies of this generator. With string instruments it is the generator who determines the pitch, with wind and percussion instruments it is the generator. Indeed this is crucial for a regular performance, a guitarist wants to play pitches by placing his fingers on the fretboard, a saxophone player wants to change the pitch by choosing a fingering, a valve combination. Still with organ pipes or with the violin string/bow interaction,

Table 1 List of musical instruments with generator–resonator coupling. The radiated frequency of the guitar and the violin is determined by the generator, still with the saxophone, trumpet, or percussion instrument, the resonator determines the radiated pitch. With organs and with the bow/string model both, the generator or the resonator may determine the pitch, depending on the control parameters

Instrument	Generator	Resonator	Frequency determined by
Guitar	Strings	Body	Generator
Violin	Strings	Body	Generator
Saxophone	Reed	Tube	Resonator
Trumpet	Lips	Tube	Resonator
Percussion	Mallet	Bar/membrane	Resonator
Organ	Labium	Tube	Generator/resonator
Violin bowing	Bow	String	Generator/resonator

the pitch may be produced by both coupled subsystems, depending on the control parameter. In a strict sense, also with saxophones or trumpets, the pitch may be caused by the generator, e.g. when playing above the cut-off frequency, where the pitch is only determined by the blowing pressure of the player, no longer by a valve combination.

So the idea of a coupled system of mostly two subsystems, generator and resonator, is perfectly fitting into the synergetic framework, where one of the subsystems slaves the other and forces it to vibrate with its pitch. Of course the reasons are mostly more complicated as displayed by the simple equation system discussed above.

Still it is interesting to note that with many musical instruments the eigenfrequencies of the slaved system are most often no longer present in the sound, although they may be sill present within the initial transient of a played tone, like is the case with the eigenfrequencies of the guitar body modes. Still then these frequencies are no longer present in the sound, although the ‘generator’, the string, is no longer a real generator after the initial transient, as then the whole system is vibrating and both, the guitar body and the string are vibrating and coupling one to the other. This can also be verified with a simple experiment. Of course the guitar body can also be made a generator by knocking on it. Then a knocking sound is heard only very short to then end in a harmonic sound of the strings which have been driven by the body in a sympathetic manner. So although the body has been driven, still the strings win the game and force the body into their frequencies. Although this simple experiment is not astonishing at first, it still challenges the traditional generator–resonator model and asks for another explanation.

2.1 Linear Example: The String

To approach the problem, first a linear example is discussed, the vibrating string. The differential equation of the string is

$$c^2 \frac{\partial y^2}{\partial x^2} = \frac{\partial y^2}{\partial t^2} \ , \tag{5}$$

with displacement y and speed of sound c. If the string with length l has boundary conditions like y(0) = y(l) = 0, so both ends are fixed, Eq. 5 can be solved using the d'Alambert solution like

$$y(x,t) = f(\omega t + kx) + f(\omega t - kx) \ , \tag{6}$$

with angular frequency ω and wave vector k. This leads to a harmonic overtone series for the string like

$$y(x,t) = \sum_{n=0}^{n=\infty} A_i e^{tnk_0 x} e^{tn\omega_0 t} \ , \tag{7}$$

with amplitudes A_i and fundamental frequency ω_0 and fundamental wave vector k_0. A similar equation is present with the vibrating tube of air, as present with saxophones, clarinets, etc. Here the differential equation is

$$c^2 \frac{\partial p^2}{\partial x^2} = \frac{\partial p^2}{\partial t^2} \ , \tag{8}$$

with pressure p. Again when assuming simple boundary conditions of a tube open at both ends with p(0) = p(l) = 0 we arrive at the same d'Alambert solution, now for the pressure like

$$p(x,t) = e^{t\omega t} f(\omega t + kx) + f(\omega t - kx) \ . \tag{9}$$

The frequencies in the tube again follow a harmonic overtone series like.

$$p(x,t) = \sum_{n=0}^{n=\infty} A_i e^{tnk_0 x} e^{tn\omega_0 t} \ , \tag{10}$$

2.2 Nonlinear Approach: Mode Coupling

Still this is a rude simplification when considering a stringed instrument where the string is coupled to a soundboard, simply because an energy transfer can only take place at a point where the string is still moving. If the string would rigidly be fixes at point where the string couples to the body, no energy transfer could happen and therefore the bridge would not move. So the boundary conditions of the string are not idle in practice and therefore the linear approach, although very useful at first, is not perfectly correct. Indeed strings continue beyond the bridge and the nut and are moving there, too. So the whole system is more difficult and therefore the modes on the string must be more complex, too.

A more radical example is the tube like with saxophones or clarinets. These tubes are not straight but have a horn at one end with a more or less sharp flaring. The flaring of this horn determines which frequencies may pass the open end and are radiated into the air to a listener, where it appears that the shaper the flaring, the better the radiation of high frequencies and the less radiation of lower ones. Therefore a trumpet with its sharp flaring sounds much brighter than a clarinet. Still this flaring has a trade-off. At this end of the tube, the different frequencies behave differently, depending on the end-correction of the tube. Lower frequencies tend to reach out beyond the length of the tube much more than higher frequencies to form standing waves. This means that the tube has a different length for each frequency which is travelling in it. For lower frequencies the tube is much longer than for higher ones. Still the harmonic overtone series of tubes discussed above in the section of the linear approach assumes that the length l of the tube is the same, no matter which frequency is present. But if this length is a function of frequency, the spectrum produced by such a horn is no longer harmonic, indeed it will become inharmonic very fast. Therefore we would expect the sound of a trumpet or a clarinet not to be harmonic at all but would find the instrument sound maybe like the sustained sound of a percussion instrument.

Still we find the overtone structures of stringed instruments and wind instruments to be nearly perfectly harmonic, even for high partials. So there must be a mechanism to force these inharmonic overtone structures back into harmonic ones. This is indeed achieved by the nonlinear coupling of the tube to a 'generator' system, a reed with saxophones and clarinets, a double reed with the oboe or the bassoon, or the players' lips with trumpets and trombones. This coupling works in such a way that the modes of the instrument, which are inharmonic at first, couple one to another.

For this mode coupling to appear, the modes of the system must not be orthonormal one to another any more, as is the case with the solution of the linear d'Alambert solution. Two modes y_1 and y_2 are orthonormal if their convolution integral becomes zero like

$$\int y_1(x)y_2(x)dx = 0 \ . \tag{11}$$

This means that they cannot interact. Still if the modes appearing in musical instruments are slightly different from the simple modes, so e.g. if they are prolonged or shortened a bit, this integral is no longer perfectly zero like

$$\int y_1(x)y_2(x)dx \neq 0 \ . \tag{12}$$

Therefore these modes will interact to a certain extend. As discussed above, the modes appearing in strings and in the air columns of saxophones, clarinets and the like are slightly or even quite a bit different in length. Therefore the integral is different from zero, the modes interact, they transfer energy one to another.

Still this energy transfer is not the crucial point for mode coupling. It would only mean that one modes becomes a bit louder while another one is reducing its amplitudes. Still modes not only have an amplitude and a frequency, they also have a phase which they travel through. Taking the air column as example, the pressure of each modes takes the value of 0–2 π, goes through a whole sine wave. So a mode in the air column may be a bit ahead or behind the other modes. If these modes were perfectly orthonomal one to another they would not interact and continue their phase relations undisturbed. Still as the modes are not perfectly orthonormal they interact and therefore exchange also their phase. Then the one behind in phase will tear the one in front a bit back in his direction and vice versa. So the modes will then control their phases in such a way to lock these phases to one common phase. Then mode coupling may become a mode locking which is the case with musical instruments and is the reason that these instruments do not sound inharmonic, as they would if such a mechanism would not exist (Fletcher 1978; Dubnov and Rodet 2002; Abel et al. 2006; Lottermoser 1983). Mode locking is forcing the modes into a common phase and therefore ensures a harmonic overtone structure.

This mode locking depends upon certain conditions, the amplitudes need to be strong enough, the modes must not be too far apart, and the coupling strength need to be strong enough. If these conditions are not met, musical instrument sounds fall apart and such instruments are not working properly anymore.

It is interesting to note that the core element of this mode locking is present at the generator of the system, as there the interaction of the modes is strongest. This also holds for strings, where most of the energy transfer happens around the bridge and nut. Although this mode locking is only one possible mathematical explanation for the slaving of the system, it is an important formulation, as it concentrates on a crucial point for musical instruments, namely that they play in harmonic overtones and do so only because of this nonlinear effect. If such nonlinearities would not be present, we would not be able to produce pitches with these instruments at all.

2.3 Nonlinear Approach: Generator–Resonator Coupling

As discussed above, the standard model of most musical instruments is one of a generator–resonator coupling (Fabre et al. 2012). Indeed, most of the generators are nonlinear by nature. With reed instruments, the generator is the reed, inserted in a mouthpiece and driven by an air flow into this mouthpiece. With double-reed instruments, like the oboe, the back-pipe, the crumhorn, or zurna-like instruments, two reeds are attached one to another, also driven by the airflow of the player. These reeds act like valves, opening and closing, and therefore allow more or less airflow through the reeds or through the small gap between the reed and the mouthpiece. The reed rests in an equilibrium position without any airflow. Then, if the player starts blowing, pressure is acting upon the reeds surfaces, its lower one in the mouth of the player and its upper one in the mouthpiece. The air stream

travelling into the mouthpiece is strongly changing direction in there, because of the enlargement of the cross section, the gap between the mouthpiece and the reed is about 1 mm during resting, the mouthpiece is several centimeters wide. So the flow starts changing its direction in the mouthpiece in an complex way forming vortices and ending in a turbulent mixture (da Silva et al. 2008; Bader 2008). This turbulent mixture acts like a heavy stop to the flow. Additionally, the pressure in the mouthpiece is much lower than the one in the mouth. Therefore, the pressure acting on the lower side of the reed is much stronger than the one on the upper side. This pressure gradient over the thickness of the reed acts as a force on it, the reed starts moving and closing like a valve. Still, this closing means that the small gap between the reed and the mouthpiece of about 1 mm is decreasing. But a decrease in this gap also means a decrease in the flow allowed through this gap. Therefore the flow changes its strength and the pressure in the mouthpiece changes. This change in pressure has a kind of Gauss impulse shape in time. It is travelling with the speed of sound through the tube, is reflected at the horn (the end of the tube), and travels back to the mouthpiece. There it is interacting with the reed, opening the gap of the reed, the valve opens again. This leads to an increase of flow of air through the gap once more and the whole process starts all over. As this process will repeat, a periodicity T is present. The length of this periodicity is determined mainly by the time, the impulse needs to travel down the tube and comes back again. If the player is applying a certain fingering, a combination of open and closed valves along the tube, the tube length changes, and therefore the periodicity is determined by the fingering of the player. Of course, the fundamental frequency, or played pitch is $f = 1/T$. So in other words, the player can change the pitch of the instrument by changing valve combinations. It is interesting to see that sometimes the valve is not basically closed and opens again when the impulse comes back, but behaves vice versa, is mainly open and closes when a back-travelling impulse hits the reed again. This happens, if the impulse travelling along the tube is an underpressure impulse rather than an overpressure one.

From this behavior of reed instruments, the nonlinearity of the reed as valve becomes obvious. When plotting the air flow through the reed as a function of the pressure applied by the player, a highly nonlinear function appears (Dalmont et al. 2005). With low playing pressure the flow increases rather linearly, still closing a bit and therefore reducing the flow. So there must be a point, where an increase of blowing pressure will no longer lead to an increase of flow, but to a decrease, as then the valve already allows only a small flow. If the pressure is still increased, the reed will close even more, then reducing the air flow allowed through the reed-mouthpiece gap. Indeed, with very high flow, theoretically no flow can happen at all, the valve is perfectly closed. This is a highly nonlinear behaviour of the reed and therefore, if taken as a generator, the reed is a nonlinear generator, while the tub, the impulse travels through is still acting linear. So we have a nonlinear generator and a linear resonator.

A similar behavior is present with the violin bow-string interaction (Duffour and Woodhouse 2004a, b; Woodhouse and Schumacher 1995). Its nonlinear nature appears right away if one considers the states of the bow friction. Either the bow

sticks to the string, acting on it with a sticking friction, or it is sliding along the string with a sliding friction impact. These two states are clearly separated one from another, and plotting them would lead to a sudden jump in the function at the sticking-slipping boundary. Therefore, the bow-string interaction is also highly nonlinear, and therefore again a nonlinear generator is present, which is the bow. When we assume the string as linear (with its restrictions discussed above) we again have a nonlinear generator and a linear resonator.

Now starting from such an assumption, a model of a delay-line was proposed by (McIntyre and Woodhouse 1983). The nonlinear generator is given as a function of the velocity of the system. This leads to a state at the generator, changing over time. As each discrete time point, this state is fed into a delay line, so e.g. with the violin bow-string model the impact of the bow is travelling along the string, in both directions, to the bridge and to the nut. Still the delay-line is also feeding back a previous state of the system which has already travelled along the string back to the generator, changing the generators' state, too. This delay-line and interaction model is a simple generalization of such systems and is easily implemented using a convolution of the generator function with the delay-line. It is capable of reproducing complex bow-string or reed-tube interactions, also producing initial transients.

2.4 Impulse Pattern Formulation

Another approach, previously proposed by the author, is an Impulse Pattern Formulation (IPF). It is based on the fact that most musical instruments work with impulses, short wave-fronts or Gauss-impulse shaped displacements or accelerators. As appears from the discussion above, this holds for the reed-tube interaction, where a short impulse is travelling along the tube. Indeed, when simulating blown instruments with a synthesizer, as a first, rough estimation a rectangular impulse is used. It also holds for the bow-string interaction, as the sudden change from sticking to slipping is a short impulse travelling along the string. Other examples are guitars or pianos. A plucked string shows a travelling wave like a trapezoid shape circling around the string. Still a string has nearly no radiation, as the air is rather flowing around the string than radiating into the surrounding air. The string is an acoustical short circuit. The only radiation appears with the soundboard, top plate, etc., which are structures of large areas. So only the energy transferred from the string to the soundboard is radiated. Also, only vibration energy is radiated, a static displacement does not vibrate. So only the temporally changing component of the string at the bridge is transferring vibrational energy to the soundboard and contributes to the radiated sound. Still with a trapezoid shape travelling around the string, the temporal change at the bridge is nearly zero most of the time. Only when the tip of the trapezoid shape is travelling around the bridge, being reflected there, a considerable temporal change of force is acting on the bridge. This time span is pretty short compared to the complete periodicity T of the string. So with

guitars or pianos, the string is rather 'knocking' on the soundboard than performing a continuous movement. At each periodicity T one of these knocks appears, therefore preserving this periodicity. Still again, the energy transfer is impulse-like and not continuous.

This impulse-like behavior of musical instruments also holds for brass instruments, where the lips act like a valve, very much like with reed instruments, although the valve is an opening rather than a closing one. With percussion instruments the impulse nature is obvious right away. Flutes and organ pipes may show a sinusoidal behavior with low frequencies, still an impulse pattern appears with higher frequencies. E.g. transverse flutes experimentally show a very short time span when the flow is sucked into the tube, while most of the time it is flowing over the embouchure hole (Coltman 1968b). This short in-flow then produces a short impulse into the flutes' tube.

So if the impulse nature of most musical instruments is obvious, it is straightforward to formulate a model based on travelling impulses or wave-fronts. The model proposed tries to hold for all musical instruments. Therefore it is formulated in the most general way. It therefore at first lacks of many characteristic features of these instruments, which are taken into consideration explicitly by bottom-up methods like e.g. Finite-Element (FEM) or Finite-Difference (FDM) time domain solutions, where a set of coupled differential equations governing the motion of the instruments is performed. Still it is shown, especially in the section of the guitar, that this model is capable of reproducing the basic nature of even such complex parts as initial transients compared to a FDM model for the different guitar parts to an astonishingly high precision. As the model proposed is a self-organizing one, a basic feature of such models seem to hold here, too, namely that these synchronization models are often able to connect very basic parameters with high-level phenotypic system behaviour with only one operational step. This again leads to further assumptions about the nature of musical instruments, as well as about musical perception, maybe evolutionarily built in such a way to fit perfectly the physical nature of musical instrument or general sound producing systems.

To be as general as possible, at first a system parameter g is defined, which will become a physical parameter only later. So g could become a periodicity of the system—changing from an unstable periodicity at the initial transients of the sounds to a stable one during the quasi steady-state of musical tone production. It may also be the amplitude of a sound, or even a combination of both. Furthermore, the model assumes a standpoint. This could be the string of a guitar, which acts upon the guitar top plate, which again acts upon the inclosed air or the ribs and the back plate. All these parts then act back to the string. The string will then find the impulse it sent out previously as coming back to it, still most often with a certain damping. So it is reasonable to have a parameter α, which accounts for the reduction of energy, the impulse experienced during its travel through the system. Still the standpoint may also be the top plate or any other component of a musical instrument, as will be shown below. The model is discussed here briefly, for more details see (Bader 2013).

The system g is assumed to act upon itself with a damping α like

$$-\frac{\partial \bar{g}}{\partial t} = \frac{1}{\alpha}\bar{g} \ . \tag{13}$$

The system state g changes over time. Still to be most general and allow delayed as well as instantaneous interaction, this most general formulation uses a bar as symbol $\bar{g}$, denoting that the kind of temporal interaction is not yet decided.

Next we include damping, which is a strong factor with all musical instruments. The damping is introduced as an exponential function, where the change of the system is the exponent like

$$e^{-\frac{\partial \bar{g}}{\partial t}} = \frac{1}{\alpha}\bar{g} \ . \tag{14}$$

This equation can be taken as an instantaneous interaction, e.g. a reed coupling to itself simply because of its stiffness. Then the system state $\bar{g}$ would be at the same time point on both sides of the equation. This leads to a system behaviour of a damped 'vibration', which of course lacks of the details of the motion and only describes the basic system behaviour, as discussed above.

Still when we have the generator/resonator model, the system state leaves the generator and is reflected by a resonator, so acts back delayed, and therefore there is a present state g and a future state g+, acting back on the system like

$$e^{-(g_+ - g)/\Delta t} = \frac{1}{\alpha}g \ , \tag{15}$$

If we solve for g+, we get

$$g_+ = g - \ln\frac{1}{\alpha}g \ . \tag{16}$$

This equation is an iterative one, similar to those used in programming languages, where the future state g+ is calculated from the present state g. Although this equation is derived from a reasoning of a musical instrument with impulse nature, coupling a generator to a resonator and applying damping, its mathematical form is like that of a logistic map. This map is very well studied and shows stable system behaviour for low parameter values of $1/\alpha$, only then to bifurcate several times to end up in chaotic motion with high values, as shown in Fig. 1. This is very reasonable for musical instruments, e.g. for reed instruments. With normal playing pressure, a stable oscillation appears. Still with low blowing pressure, only noise is produced. Indeed, at the boundary between these two states, saxophones or clarinets may play multiphonics, several harmonic overtones series at the same time. In performance of contemporary jazz or classical music these multiphonics are often used. They can also be produced by complex valve combinations, where textbooks have been published showing the fingering with and the respective multiphonics produced, up to 5–6 tones at once (Krassnitzer 2002). Also, from the standpoint of nonlinear dynamics, the sudden phase change of the reed, and also of brass instruments, is a clear indication of a nonlinear system which may self-organize, to then produce a simple harmonic output.

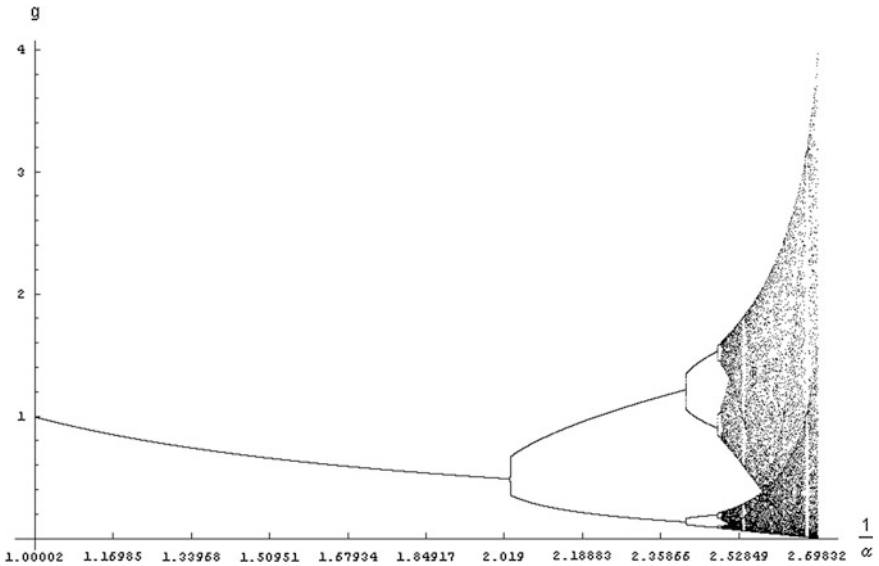


Fig. 1 Bifurcation scenario of the logistic map Eq. 16 for increasing values of $1/\alpha$. So e.g. with saxophones, low blowing pressure would be on the *right side*, while higher blowing pressure is located on the left. Then, the noise produced with low blowing pressure is the chaotic behaviour on the *right*, while the stable oscillation is the constant value on the *left*. In between bifurcations occur, equivalent to multiphonics which can be played at the boundary of the sudden phase shift from noise to stable harmonics

Still there are musical instruments which are more complex. So e.g. the guitar string as a generator distributes energy to the top plate. This top plate drives the air inside the instrument, as well as the ribs and the neck. The air and the ribs drive the back plate, which in the end acts back to the string via the ribs, the air, and the top plate. This complex set of interactions may be simplified as impulses coming back to the string from these parts with different time delays. Each part is then acting back to the initial system at different time points in the future. As discussed above, the choice of this viewpoint is arbitrary, and therefore the system may be examined from any possible standpoint. In terms of mathematics, only the strength of the back impulses would need to be adjusted.

We first derive the equation for three parts, the viewpoint and two reflecting parts. It is then trivial to write the equation of an arbitrary amount of reflection points. When we again use the system variable $\bar{g}$ in its temporal neutral form in analogy to Eq. 13 we can write

$$\frac{\partial \bar{g}}{dt} = -\alpha \bar{g} - \beta \bar{g} \ . \tag{17}$$

When again using an exponential damping we can rewrite Eq. 17 like

$$\bar{g}\ e^{\bar{g}} = -\alpha e^{\bar{g}} - \beta e^{\bar{g}} \ . \tag{18}$$

As $\bar{g}$ is only general, we must not divide this equation by $e^{\bar{g}}$, as the different $\bar{g}$ may be on different time points. Then we can write

$$\bar{g} = -\alpha e^{\bar{g}-\bar{g}} - \beta e^{\bar{g}-\bar{g}} \ . \tag{19}$$

Inserting reasonable time points for the different parts we have

$$-\frac{1}{\alpha}\bar{g} + \frac{\beta}{\alpha} e^{\bar{g}-\bar{g}} = e^{\bar{g}-\bar{g}} \ . \tag{20}$$

Here, g is the present system state, g+ the next and g− the previous state. This equation can be generalized for n+1 reflecting points like

$$g(t_+) = g(t) - \ln\left[\frac{1}{\alpha} g(t) - \sum_{k=1}^{n} \frac{\beta_k}{\alpha} e^{g(t)-g(t_{-k})}\right] \ . \tag{21}$$

Where $g(t_{-k})$ is the k$^{\text{th}}$ delayed reflection.

For understanding musical instruments with several reflection points, it is interesting to test the model for an increasing amount of reflection points, assuming that the parts more far away will have a larger damping and therefore reflect less than parts which are closer to the generator. Also quite strong damping is assumed, as is the case with wood if which most musical instruments with several reflection points are built of. Table 2 sums the results when iterating the system for a value of $1/\alpha$ from 1 to 2.7, the point of the first bifurcation with $n = 1$, the case shown in Fig. 2. Indeed, the maximum possible $1/\alpha$ for a stable, non-bifurcated behaviour of the system is $(1/\alpha)_{max} = 2.71$, as expected.

If additional reflection points are added up to $n = 4$ with parameters displayed in Table 2 it is interesting to see that the point of maximum stability is enlarged up to $(1/\alpha)_{max} = 3.41$ for $n = 4$. This seems contrary to intuition, as one might expect the system to become more unstable with more reflection points. Still the opposite is true, the system is stabilized through additional reflection points. This is because then the outgoing impulse is distributed much more, and therefore the impulse acting back to the generator is temporally spread and therefore not able to force the generator in a bifurcation. In other words, the one-dimensional string with its stable strong impulse is more capable of forcing a multi-delayed,

Table 2 Maximum value of stability $(1/\alpha)_{max}$ for different amount of reflection points n with respective reflection coefficients β_1–β_3 iterated for $1 < 1/\alpha < 2.7$. The increasing amount of reflection points stabilizes the system from $(1/\alpha)_{max} = 2.71$ for $n = 1$ to $(1/\alpha)_{max} = 3.41$ for $n = 4$

n	$1/\alpha$	β_1	β_2	β_3	$(1/\alpha)_{max}$
1	$1 \rightarrow 2.7$	0	0	0	2.71
2	$1 \rightarrow 2.7$	.2	0	0	2.74
3	$1 \rightarrow 2.7$	.2	.1	0	3.25
4	$1 \rightarrow 2.7$	.2	.1	.05	3.41

geometrically complex structure, because the straight impulse of the string is blurred by the multiple reflections of this complex geometry. All the smaller impulses returning to the string from the body parts are then no longer able to disturb the string with such ease as only one strong compact reflection can. So the large stability stringed instruments show in terms of their string-body coupling after its initial transient seems to be caused by the difference of geometry of the interacting parts, the one-dimensional string is capable of forcing the three-dimensional body to go with its frequencies only because of this difference in dimensionality.

2.5 Example: Guitar String—Body Coupling

As an example, the coupling of a guitar string onto its body is discussed. From previous physical modelling results of a finite-difference model of the whole guitar (Bader 2005b), the amount of reflection of the several guitar body parts back to the string can be estimated as shown in Table 3. We rewrite the impulse pattern equation with damping variables α, β, γ, and δ like

$$g(t_+) = g(t) - \ln\left[\frac{1}{\alpha}g(t) - \frac{\beta}{\alpha}e^{g(t)-g(t_{-1})} - \frac{\gamma}{\alpha}e^{g(t)-g(t_{-2})} - \frac{\delta}{\alpha}e^{g(t)-g(t_{-3})}\right]. \quad (22)$$

So e.g. if the system variable is chosen to be the top plate, the impulses sent out by it to the other parts of the instrument, strings, inclosed air, ribs, and back plate are considered. Then the damping variables represent the strength of the other parts to act back to the top plate, normalized by the first, strongest back impulse. The values are chosen according to results from the FDM solution of the guitar.

Figure 2 compares the results of the initial transient of an open e-string tone plucked on the virtual FDM whole-body model on the right side of the figures with

Table 3 Damping coefficients α, β, γ, and δ for strings, top plate, back plate, ribs, and inclosed air for those parts as viewpoints for the system variable g for the classical guitar

Viewpoint on part	α	β	γ	δ
String	Top plate	Inclosed air	Ribs	Back plate
	.7	.1	.05	.05
Top plate	Strings	Inclosed air	Ribs	Back plate
	.5	.2	.15	.05
Back plate	Ribs	Inclosed air	Top plate	Strings
	.5	.01	.2	.19
Ribs	Top plate	Back plate	Inclosed air	Strings
	.5	.09	.01	.3
Inclosed air	Top plate	Back plate	Ribs	Strings
	.36	.18	.18	.18

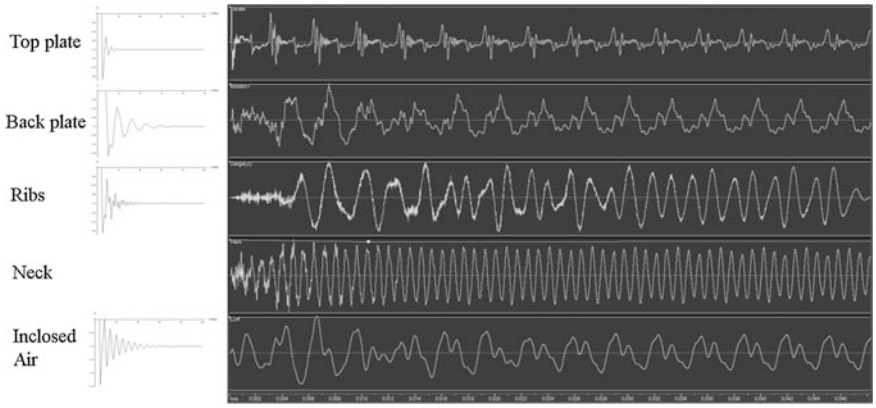


Fig. 2 Comparison between *Right* the radiation of the different guitar parts top plate, back plate, ribs, neck, and inclosed air for the initial transient of an open e-string tone plucked virtually on a whole-body finite-difference solution of the classical guitar, and *Left* initial transients of the impulse pattern equation for these parts as viewpoints of the calculation. The general behaviour of the initial transients in terms of duration and complexity clearly appears. Note that the FDM results are radiated pressure over time, while the IPF are values of g, and therefore show a stable or unstable system behaviour. So the convergence of the IPF plots means an establishment of a stable oscillation in the time domain. The basic behaviour of the different parts during the initial transient clearly appear. The *top plate* has the shortest transient phase and reaches a stable oscillation within one or two impulse cycles, the periodicities of the string. This is well known from guitars, the *top plate* is mostly responsible for the attack of the tone, it should react as fast as possible. This is one of the basic quality characteristics of musical instruments, fast attack, as otherwise fast playing is not possible, the tones would blur one into another. The IPF clearly shows such a fast attack, it converges nearly instantaneously to a constant value of g, which means a stable oscillation of the system. This also clearly appears with the FDM time series on the *right*

the results of the IPF on the left side. The time series of the FDM are integrated over the area with respect to a virtual microphone position 1 m in front of the respective part to model radiation. Note that the FDM results are pressure values over time, while the IPF results are periodicity. So when the IPF values converge to a value, as all of them do, a stable oscillation is reached. In other words, the initial transient is over then. So next to the duration of the initial transient, the IPF values show the struggling of the system during this transient phase, its change in periodicity, amount of chaoticity, or complexity. So next to a stability analysis shown above, the IPF is able to display the initial transient of musical instrument tones. It does so in terms of the system variable g, interpreted as periodicity and/or amplitude of the system. So the IPF does not result in a time series at first, it results in a basic description of the initial transient.

The inclosed air at the bottom of Fig. 3 is the slowest to converge, has the longest initial transient. Still this transient is not very complex, mainly the periodicity is oscillating with decreasing oscillation amplitude, as can be seen in the IPF for the inclosed air. The same behaviour holds for the time series of the FDM

shown on the right, the system needs quite some time to establish a stable oscillation, still the struggling of the time series until this stable-state is not very complex, without many high frequency components, a rather smooth curve.

The ribs, on the other side show a much more complex initial transient. In the FDM the amplitude raises only after some time, when the air inside the instrument drives the ribs in their transverse direction, so that they are able to radiate at all. Then the behaviour shows more complex vibration, pointing to more high frequencies [for details see (Bader 2005a)]. The IPF also shows a much more complex initial transient compared to the air.

The back plate again takes some time to come in, both in the FDM and the IPF, its periodicity oscillation, displayed in the IPF, is slower than that of the air, which may point to a faster adaptation to the frequencies of the string, as the eigenvalues of the back plate is closer to the 330 Hz of the string fundamental compared to the about 100 Hz of the Helmholtz motion of the inclosed air.

So it appears that the basic behaviour of the initial transients of the different guitar parts are very close between the FDM simulation and the IPF. As discussed above, both methods have a radical different starting point, the FDM is bottom up, solving the differential equations for the complex geometry of the instrument, while the IPF is top-down, only assuming reflections of outgoing impulses by the other parts with a certain damping. Still clearly both result in about the same for initial transient length and complexity of the different guitar parts. As the FDM has the advantage of going into details of the time series, it is often difficult to understand the reason for this behaviour or even come to general conclusions about families of similar systems. On the other side, the IPF lacks of the details of the time series, still it is able to give insight into the nature of initial transients, the reasons for their length and their periodicity oscillations. It is also able to understand why some systems, like blown instruments, are driven out of their stable oscillation regime so easily, while complex structured instruments, like guitars or pianos, are so stable in this respect.

It is also possible to produce time series from the IPF system variable g development over time, e.g. by inserting a basic waveform, sawtooth, rectangle, etc. with respective change of periodicity. Although this is beyond the scope of this chapter, it is interesting to note that the resulting sounds clearly show typical initial transients, the spitting of a trumpet or trombone, or the scratchy sound of violin initial transients. It is further interesting to see, that the same IPF initial transient time series can be used to model e.g. a trumpet and a violin string-bow system, as both are one-dimensional, and therefore the same for the IPF equation. Still the resulting sounds, and especially their initial transients, sound one time like a trumpet spitting, the other like a violin scratching. So the IPF is also able to compare different instrument families, like blown and bowed instruments, using the same impulse pattern reflection model.

2.6 Example: The Organ Pipe

Flue organ pipes, as well as flutes are labium instruments. An air flow from the players mouth or from an orfice hits a cutting edge. There it may flow above the edge or below it. With the flute, above the edge means the flow goes in a direction outside the flute into the surrounding air, while flowing below the edge means flowing inside the flutes' tube. With flue organ pipes the picture is similar, the flow goes into the direction inside the tube or flows outside the instrument.

The problem of the tone production of flues and flue organ pipes is still under debate (Fabre and Hirschberg 2000). Historically, this cutting edge tone production, also known as self-sustained oscillation, was considered a linear phenomenon. In a discussion between Helmholtz (1863) and Rayleigh (1894), different views considered the pressure of the flow acting on the air going inside or outside respectively as the reason for tone production. While Helmholtz found it to be a monopole source, Rayleigh argues in favour of a dipole sound production. The same debate was going on post-war between (Cremer and Ising 1967), supporting the inside tube flow tone production thesis, and (Coltman 1968a), who found the flow mainly outside the instrument. The difference may also be that Cremer and Ising used a very large organ pipe, while Coltman was investigating the rather small flute. (Elder 1973) fused the positions arguing for a combination of both, the Rayleigh picture. With these models, the edge tone generator was considered in analogy to an electrical oscillating circuit, the pipe, being driven by an air flow, which supports the energy necessary to maintain the oscillation.

Still with such a linear model many other phenomena of the flute and flue organ pipes cannot be understood, as the sudden phase transition from noise to a stable oscillation from a certain pressure threshold on, an oscillation with a stable pitch over a large range of blowing pressure, the sudden jumps to higher harmonics, while overblowing the flute, or the appearance of multiphonics, tones with several pitches, bifurcations at pressures at the boundaries of the state changes between noise, tone production, and higher harmonics. All these findings are very similar to the ones found with the bow-string model of the violin or with reed instruments and strongly point to a self-organized system. For such behaviour to appear, nonlinearities need to be present in the system.

The reasoning of the edge tone generator to be a self-organized system is based on experimental findings (Kaykayoglu and Rockwell 1986a, b), physical modeling results (da Silva et al. 2008; Bader 2008), and a discussion of the Navier-Stokes (NS) equation. A typical oscillation around a cutting edge is shown in Fig. 3 as an experimental result of a water flow. Here, no tube is attached. Such a situation is different from flutes and flue organ pipes. Coltman criticized such experiments stating that we try to understand the flute by a system we do not understand either. Indeed, the labium of flues may only therefore oscillate because a tube is attached. In Fig. 3, this oscillation appears solely. Still this is also the case with flue organ pipes. When the pipe foot is deattached from the pipes tube, one may simply blow into the foot, as would do the wind system of the organ, to produces the pitch of

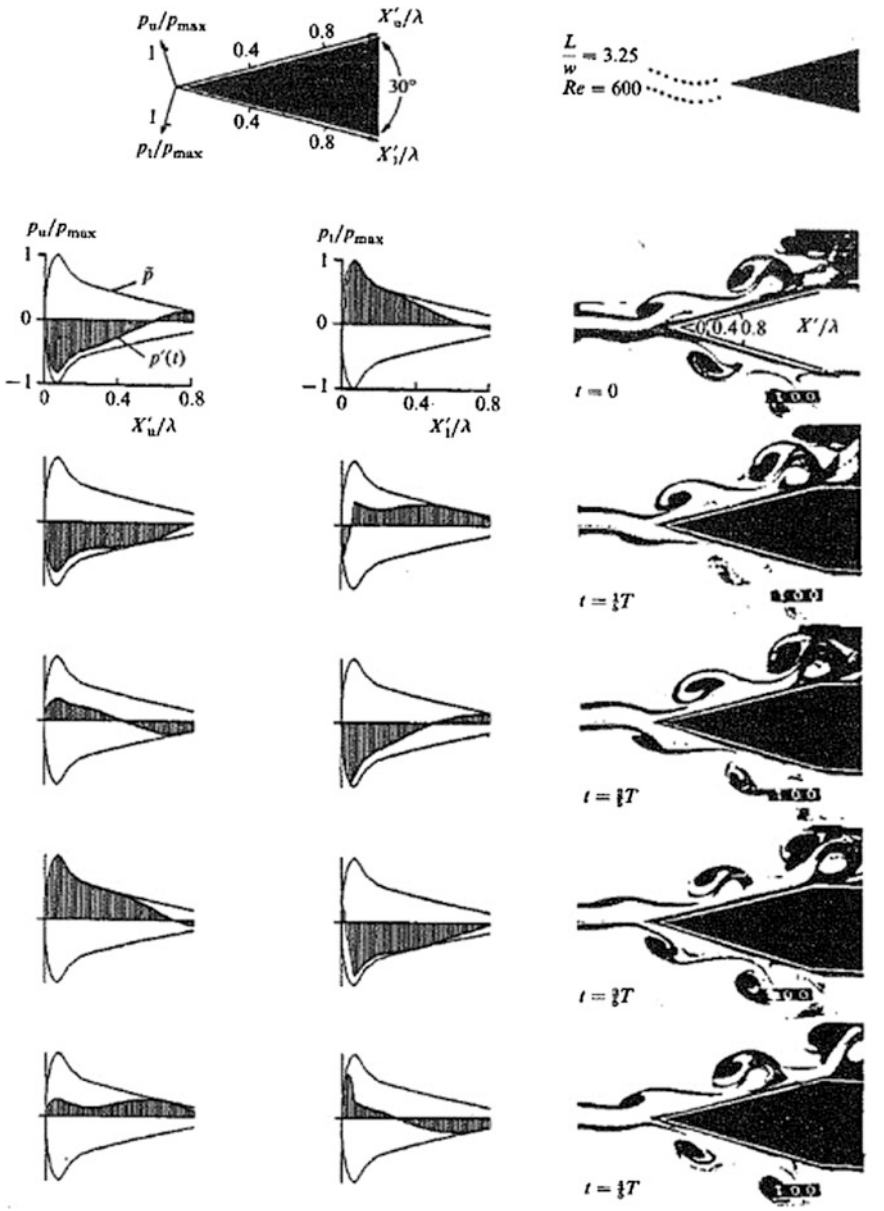


Fig. 3 Pressure distribution (*Left* above and *Middle* below the *edge vertical line*) and *Right* velocity field for a measured edge tone oscillation without an attached tube at five different time points t. The different vortices of the flow field are much more elaborated than the pressure, which has a much more smooth distribution. *From* Kaykayoglu and Rockwell 1986a, b

the pipe itself. Still with flutes this may be different, where the labium itself is not so easily able to oscillate the way it would with the pipe. The reason for this difference is again pointing to a complex nature of the system, as discussed above with the generator–resonator picture. With organ pipes the labium produces the pitch, with flues this pitch is determined by the length of the tube, which is necessary to play melodies.

So discussing a self-sustained oscillation without a tube is reasonable for musical instruments, at least for the flue organ pipe. Still the reason for presenting exactly this experimental results is the comparison between the velocity field on the right of the figure and the pressure fields above and below the labium middle line on the left and middle column respectively. Next to a very precise description of the vortex field, showing the first and the second vortex behind the laminar flow very clear, this velocity field is very differentiated, the vortices are highly distributed and differentiated. On the other hand, the pressure fields are very smooth and do not show a complex distribution. Next to many other interesting findings of this experiment, e.g. the spectral distribution of the vortex street, for the reasoning here, this is enough to proceed to the physical model of a flute.

Another important finding of the experimental data of Kakayoglu and Rockwell is the pressure field around the labium. It shows a very strong gradient at the first vortex, shown in the left an middle column as maximum pressure envelope. The pressure is nearly zero at the tip of the labium (very left) and reaches its maximum very fast. This strong pressure gradient is a sound source, which is identified with the self-sustained oscillator as present right behind the tip of the labium.

Still, to understand this model, the differential equations used there need to be briefly considered, which is basically the Navier-Stokes equation

$$\partial_t \, u_i + u_j \, \partial_j \, u_i = -\frac{1}{\rho} \, \partial_i \, p + \nu \nabla^2 \, u_i \text{ with i=1,2,3 .} \tag{23}$$

This two-dimensional version of NS has flows u_i in the two directions i = 1, 2, which are differentiated with respect to both directions j = 1, 2. The subscripts with the differential symbols indicate an Einstein sum convention, so the differentiation is over both directions, where the terms are summed. The flow differentials are balanced by the pressure p differential on the rhs with density ϱ, and the damping terms of a second derivative with respect to time and viscosity ν. This version is incompressible, which means that the changes in the flow sum up to zero like $\nabla u = 0$.

Obviously, there is a nonlinear terms in the equation, $u_j \partial_j u_i$, leading to an unstable behaviour and turbulent flow. Although turbulence is a highly complex field, still not fully understood, the $k - \epsilon$ model proposed by Kolmogorov 1941 and its different formulations [see e.g. (Durbin and Pettersson 2001)] is highly intuitive and fits astonishingly well in many experimental situations. Although a detailed mathematical description is beyond the scope of this chapter, the model is used in a Finite-Element simulation of a flute (Bader 2005b) which is part of the reasoning proposed here. The main idea of the k$-\epsilon$ model is the development of vortices in a flow, which is laminar at first and then becomes turbulent. Kolmogorov assumes an

energy flow from the laminar flow to the first, large vortices, from these to smaller vortices, and so on until a smallest vortex is built. As the rotation speed of vortices increase with decreasing vortex sizes, the damping of smaller vortices is considerable and so a smallest vortex exists, as even smaller ones would be dissipated immediately. This is an important point for understanding turbulence. As each larger vortex produces many smaller ones, a cascade of vortices, or eddies, is produced in a turbulent flow. Now as the damping increases with many small eddies, a highly turbulent flow means a very strong damping. This appears in the classical experiment of turbulent flow, where at first a laminar flow with a medium speed is flowing through a simple tube. When increasing the speed of this flow, from a Raynolds number of about Re $>$ 6,500 the flow becomes suddenly turbulent (sudden phase change from laminar to turbulent). Indeed, the flow therefore also suddenly stops within the tube, its mean velocity becomes nearly zero because of this turbulence. So the Kolmogorov assumption of increased damping caused by the cascade of vortices seems reasonable.

Following this reasoning, the k$-\epsilon$ model does not consider all the small vertices appearing in a turbulent flow. It need not to do so, as they are included in the general idea of energy flow through the eddies. Therefore, the flow u of the NS equation is substituted by $\bar{u}$, which now is divided into two flows, one laminar, mean flow U and one turbulent flow u like

$$\hat{u} = U + u \ . \tag{24}$$

Substituting this into the NS gives the Raynold-Averaged Navier-Stokes (RANS) equation

$$\partial_t \ U_i + U_j \ \partial_j \ U_i = -\frac{1}{\rho} \ \partial_i \ p + \nu \nabla^2 \ U_i - \partial_j \ \overline{u_j \ u_i} \ . \tag{25}$$

The RANS is very similar to the NS except for one additional term on the rhs $[u_j u_i]$, summing all terms which produce, transport, redistribute, and dissipate energy from the mean flow to the turbulence. To deal with the many new variables introduced in this term, the model defines a turbulent energy k and a turbulent dissipation ϵ, resulting in two new equations for these terms describing their time and spatial differentiations [for details see (Bader 2005b)]. Therefore, turbulence is modelled without displaying the details of the vortices, therefore resulting in an understanding of the main process in terms of energy flow and dissipation. In other words, the new term in the RANS equation can also be viewed as a second damping term to the NS equation, representing turbulent damping, which is considerable as discussed above.

Figure 4 shows results of two Finite-Element simulations of a flute, where a player is blowing to the labium. The flute is the horizontal tube, the box above the flute corresponds to a section of the air outside the flute to display the flow not entering the tube. The streamlines are flow velocity, the background is the pressure distribution. The top plot shows the results of a NS simulation, the bottom plot represents the RANS results. Experimentally, only about 3 % the players lips goes

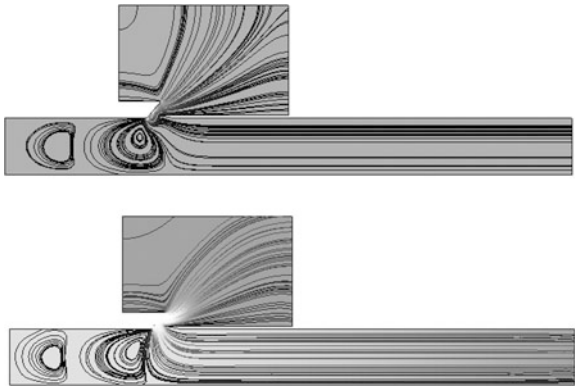


Fig. 4 Finite-Element simulations of blowing on the labium of a flute, stream-lines are velocity, background is pressure. The *box* above the horizontal tube is a small part of the surrounding air above the labium to display the flow behaviour when the flow is not entering the tube. The simulation uses *Top* the NS equation and *Bottom* The Raynolds-Averaged Navier-Stokes equation. With the NS model, about 50 % of the flow goes into the tube, with the RANS model it is only about 3.5 %, corresponding to experimental results (Coltman 1968a, b)

into the tube, a very inefficient tone production system. The NS simulation leads to a 50/50 split of flow into and out of the tube, while the RANS results in about 3.5 % energy going into the tube. The difference is caused by the last term of the RANS equation which is an additional damping caused by any region, where the flow changes direction, which is especially the case at the region around the labium. Additionally, again the flow streamlines are much more differentiated than is the pressure distribution, the background in the figure. Note, that both models did not take turbulence into account with respect to its fine structure. All small eddies which may be present around the labium are not displayed, and must not be displayed, as they are already present in the RANS model definition. Still, the elaborated velocity and smooth pressure distributions were also found with the experimental results of (Kaykayoglu and Rockwell 1986a) discussed above.

Therefore we can summarize the results from experimental findings, physical modelling, and the NS as well as the RANS equations like

- the labium is a bi-stable geometry, where the flow may easily change from above to below the labium,
- the NS equation is nonlinear with respect to the flow term $u_j \partial_j u_i$,
- the NS equation is linear with respect to the pressure term, and
- the velocity flow distribution is much more complex than the pressure field.

Additionally it appears that

- sound production happens at regions of strong velocity changes, which cause a strong pressure gradient here. The point of the strongest pressure change is the first vortex, as experimentally shown (Kaykayoglu and Rockwell 1986a, b).

- The interaction between the pressure and the laminar flow from the orfice is present all over the laminar flow region.

From these findings we can formulate a model of tone production in musical instruments with a labium coupled to a tube. Returning to the global picture of two interacting systems, we have as a tone production mainly the first vortex above or below the labium. These vortices appear because of the nonlinear nature of the NS equation. Due to the very strong damping within turbulence, the laminar flow decrease its velocity very strongly, leading to a sudden stop in the flow. This stop corresponds also to a strong gradient in the pressure at this point. Therefore, the first vortex is acting as a sound source. As this vortex is very small, sound production can be viewed as point-like. These point-like pressure change then travel into the tube. The tube, on the other side, acts back onto the labium over a wide range all along the laminar flow from the orfice to the edge. As the pressure distribution is very smooth, it can influence this laminar flow to a great extend. This is further enhanced by the nonlinearity of the NS equation enabling vortex production easily. The bi-stable nature of the labium then allows these vortices to flow in two distinctively different directions, into or outside the tube. Therefore, a highly nonlinear generator, the NS-labium-vortex system is acting on a linear one, the tube. The tube is then reacting onto the laminar flow via its smooth distribution very easily.

Therefore, the linear system is forcing the nonlinear one to go with its frequencies, like is the case with the flute. The organ pipe, which may oscillate with the tubes' frequency on its own, may still be disturbed by a tube attached to it with considerable different resonance frequency, leading to bifurcations and noise. There are very few experimental studies here, as musically this is not part of performance practice. So the edge tone generator or labium system is again a self-organizing one, coupling a nonlinear and a linear system. It is therefore not surprising that is shows all behaviour typical for such a system, as sudden phase changes from noise to a stable pitch, bifurcations, etc.

2.7 Summary

To discuss other musical instruments as self-organized oscillators is beyond the scope of this chapter. Still many experimental findings and models point into this direction. The singing voice is most often considered to be a van der Pol oscillator, so a nonlinear model [see e.g. (Steinecke and Herzel 1995)]. Therefore it shows all characteristic features of self-organized systems, like sudden phase changes from noise to a stable oscillation, hysteresis loops, or bifurcations, e.g. used in Tuvan throat undertone singing in the style of *kagyraa*. With organs, the effect of synchronization of two or more pipes via the surrounding air is found (Mitnahmeeffekt) (Abel et al. 2006; Lottermoser 1983). Mode couplings in bars, similar to those found in strings and with tubes of complex flaring have been

discussed (Legge and Fletcher 1987). Cymbals are identified to be strongly nonlinear, showing beautiful bifurcations when played with increasing amplitude (Touzê and Chaigne 2000). The tam-tam shows mode coupling to higher harmonics, which lead to a sound starting from a nearly sinusoidal fundamental, only to increase energy up to very high frequencies, bursting out (Rossing and Fletcher 1983). Also friction instruments, like the singing saw, singing glass, friction wooden blocks, like the *lounuet* of New-Ireland, or the *harmonia* built by Chladni all are systems very similar to that of the bow-string model and therefore self-organizing. Many other examples could be discussed, too.

So in the end, musical instruments not only have some additional nonlinearities to fresh up sounds, they are self-organizing and synchronizing systems by their very nature. Only because of this self-organizing nature they have

- such a very precise harmonic overtone structure,
- have such a large range of possible articulations, or
- have such characteristic initial transients enriching musical performance.

Now, as musical instruments are self-organizing systems, it is interesting to consider self-organization as a principle in music psychology, too. The next section, within the scope of a chapter, discusses findings in this direction. It appears that the models developed there are often much more elaborated than linear models and therefore able to explain complex perception and performance actions linear models cannot fully account for.

3 Music Psychology

Music perception in most cases is not linear at all, and examples of nonlinear relations between a stimulus and a perception are many. In loudness perception, the curves of equal loudness depend are frequency and furthermore depend on the overall loudness levels, where the dB weighting dBA, dBB, or dBC are needed to compensate for this nonlinearity (Stevens 1961; Florentine et al. 2011). The perception of roughness (Daniel 2008; Schneider et al. 2009) depends on the critical bandwidth physiologically present on the basilar membrane (Seever 2008) is are also the reason for masking, a nonlinear effect e.g. used for MP3 coding (Vaseghi 2007). In rhythm perception, phase-transitions occur at about 60 beats per minutes (BPM) and at about 240 BPM (Repp 2003). Below 60 BPM the perception and production becomes increasingly difficult and subjects tend to count double time. At 240 BPM the situation is vice-versa, as the beats are too fast, known e.g. from Bee-Bop tunes. Timbre perception is very complex and not yet understood completely, as many features like brightness, fluctuations, noise at initial transients, etc. interact (Herra et al. 2003). Indeed, all senses perceive logarithmically, and so a nonlinear relation is present right from the beginning (Keidel 1975).

Another feature of music perception making the situation complex is that the parameters all interact, and so in a state-space, the axis are not orthonormal one to

another (Garner 1974). A simple and well known example is the interrelation between pitch and loudness. If a sine wave, producing a pitch sensation is increased in volume, and so in loudness, at the same time, the pitch is raising perceptually without changing its physical frequency. In loudness perception, partials of a sound do not add linearly to a loudness percept, even when they are in different critical bands (Green 1988). This feature is even stronger with spectra consisting of harmonic partials compared to spectra with an inharmonic overtone structure (Drennan and Watson 2001). This makes an analysis of the components difficult, and complex models of confound percepts are needed.

Interaction between the different musical features may lead to sudden phase changes in perception. When combining rhythm and timbre in a groove, the beat may change by adding one instrument sound on the rhythmic level where before an off-beat was perceived (Bader and Markuse 1994). In certain pitch registers and during the steady-state of a sound, two instruments may not be distinguishable and perception changes between e.g. a violin and a saxophone, back and forth, all of a sudden (Reuter 1995). In a chord progression of a cadence, playing the solving chord at a rhythmically unstressed time point, the meter perception may change because of the chord content (Lerdahl and Jackendoff 1983). The situation may become even more difficult when higher semantic levels are incorporated.

The link between percept and physiological localization in the ear, the auditory pathway, or the brain has not yet lead to a clear picture for most of the features. Still a global understanding is still one of a reduction of complexity from the basilar membrane in the inner ear (Oertel 2002) to an understanding of music in the neocortex (Poeppel 2012). Although this is a very rough description, many features point to such a model. So e.g. the nerve cells at the stereocilia on the basilar membrane fire with a frequency of up to 4 kHz by interlocking nerve firing rates. Still the highest frequency measured in the primary auditory cortex AI is about 200 Hz (Bendor 2011). Traditionally, the brain stem, ganglia cells, or auditory pathway were associated with a more automatic, unconscious treatment of sensory data, while the neocortex, the secondary auditory cortex, the Broca's area analysing musical syntax (Koelsch 2012), the supplementary motor cortex (Levitin 2006), strongly connected to music perception, and the prefrontal lobe, active with only very simple rhythm perception to perform pattern analysis (Thaut 2005), all have been considered the associative part of perception. Here, subjects are perceiving music consciously and are able to deal with the income more or less freely. This picture of an automatic, unconscious lower-level and a associative, conscious higher-level treatment of acoustic stimuli is complemented by a two-way interaction, a bottom-up way using afferent nerve fibres and a top-down way of efferent nerves acting back to the very low level of the inner ear again (Ryugo et al. 2011). So conscious decision making in hearing music also tunes the ear mostly by inhibition of certain nerve fibres. Traditionally, the low-level parts are taken as physiology, the higher level part as psychology. Historically, this lead C. Stumpf to the conclusion that tonal fusion, the perception of one pitch although many partials are present, is not a psychological but a physiological one (Stumpf 1883/1890). Still this reasoning runs into problems with a monist view of

consciousness (Damasio 2006). Although Descartes claimed a difference between body and mind and so promoted a dualistic view, most neurocognitive researchers today are monistic and claim that percepts are neural activity. So telling the boundary between physiology and psychology would need a dualistic view. Still on the other hand, promoting a monistic standpoint would lead to the question, where conscious perception ends, so e.g. if it would be also present in the solar plexus, where also neural nets are found.

Nevertheless, this picture of a bottom-up/top-down perception has been challenged by models of neural networks (Todd and Loy 1991; Arbi 2003). The highly self-organizing nature of these networks resulting in emergent behaviour is obvious. Also the amount of information processing needed for a linear model of perception, performing feature extraction by data reduction as used in most models also in the field of Music Information Retrieval (MIR) is demanding [see e.g. (Klapuri 2006)]. Taking in mind the high efficiency found in many biological systems, perception based on self-organization is much more efficient (Smith and Lewicki 2006). This reasoning is old and already found in music theory very early with H. Riemann who supported an idea of Lotze, the economy of hearing, as the basis for tonality perception, the fact that we extract a tonality like e.g C-Major from the chord progression C-F-G-C, or the tonality e♭ minor from e♭-a♭-b-e♭ [for a deeper discussion see (Bader et al. 2009)]. Furthermore, many linear bottom-up models fail to show sudden phase-changes in perception, like the ones discussed above. Therefore it is reasonable to assume neural maps of self-organized and synchronized nature as basis of many perceptual features in music.

Only to mention a few, below several examples are presented, discussing past and present work in this field. Many more of these are found in the literature. With music perception, the reasoning is experimentally much less approved than with musical instruments, as measuring neural networks with all interactions is only partially possible and therefore the physiological proofs are scarce [see e.g. (Bandyopadhyay et al. 2010; Bendor 2011)]. Still perceptual models have been proposed which are able to produce the features found in music perception most often much more detailed and more close to music than linear models. Still often they are mathematically more demanding. As this is only an overview, like with the musical acoustics section above, only the findings are given and the reader may consult the research papers for more details. Although not on music, also synergetic models of the brain (Haken 2002) and psychological phenomena in general(Haken and Schiepek 2006) have been proposed.

3.1 Timbre

Timbre has often been found to be a multidimensional space. In listening tests, comparing different instrument sounds pairwise by their perceptual difference, a two- or three-dimensional embedding space was found [see e.g. (Herra et al. 2003)

or (Bader 2013) for a review]. The axis of this perceptual space can then be associated with physical parameters. The most important of these parameters are

- pitch
- brightness
- synchronicity of phases during the initial transient
- spectral fluctuations
- inharmonic components within the initial transient
- relation between even and odd harmonics.

The first one is not a timbre parameter. Still sounds of different pitches have been used in such experiments, too, where it is striking to see that this feature overrules all timbre features. So it is reasonable to assume pitch as more prominent in terms of attention. Indeed, increasing the pitch will also increase brightness, which is the first and most important timbre feature. Brightness is also most prominent when it comes to instrument identification during the steady-state of the sound. If the initial transient is present, this is most often used for instrument identification, where inharmonic components and the phase relations of the harmonic partials are perceptually the most salient ones. It is also interesting to see that there are different perceptual strategies of the two tasks, comparison and identification. Also, next to the main components mentioned here, there are numerous different features with special sounds, artificial, hybrid instruments, etc. So it is reasonable to assume that perception is much more complex and adaptive to the sounds presented.

To test the findings of multidimensionality with timbre perception, (Caclin et al. 2006, 2008) investigated the orthogonality of these parameters in terms of neuronal coding. If the features, like brightness or flux are treated in different neuronal networks, their content would be independent one from another. Then changing one of the parameters would only change the perception of this parameter and leave the perception of the other parameters unchanged. As a result, the perception would need to be additive, where all parameters added will result in the perceptions in an additive way, too.

Caclin used the paradigm of (Garner 1974). He compared stimuli varying by two parameters, so e.g. varying pitch and loudness, and uses the reaction time of subjects to estimate if the relation is separable, and orthogonal, or if it is integral, confound. When examining pitch, three combinations are possible. Either the loudness increases as pitch increases, or it stays the same, or it even goes the other direction, decreasing with increasing pitch. If the change in loudness does not effect the reaction time of subjects on pitch variation, the dimension is separable, if it does change the reaction time, the two variables are integral, they depend one on the other perceptually.

Indeed, the perception of dimensions are not always additive with the stimuli used in (Caclin et al. 2008). In a mismatch negativity (MMN) EEG experiment, varying the attack time and the relation between even and odd harmonics was found to be additive at the fronto-central electrode. Still when including brightness, the results were no longer additive, pointing to an integral perception.

Self-organizing neural networks have been proposed to perceive musical timbre, too, like in (Cosi et al. 1994; Toiviainen 1992; Kostek 2005), or (Feiten and Günzel 1994). The Kohonen type of neural nets was used. Here, features of the sounds used are extracted. Each neuron on a two-dimensional map is a feature list. The network is trained by comparing the features of the sounds with the feature list of each neuron. The neuron which comes closest to the feature list of the one sound is changed in direction of the sound features slightly. After many of these tunings the network shows fields which correspond closely to different sounds. Therefore, when testing the network with new sounds similar to the ones used for training, the neurons of feature lists close to the training sounds will correlate strongest to the new sounds.

It is important to note that these networks are trained by the sounds using many trails, which is similar to a training of subjects. Then the schemata of these sounds are learned and similar sounds are categorized according to the schemata of the trained network. Next to possible applications of such nets, the basic principle of learning appears with these models and succeed to built up regions similar to certain sounds.

3.2 Rhythm Production

The production of rhythm when playing with two hands, like drummers, pianists, or xylophone players do, has features associated with a synergetic system. This appears when formalizing such movements to a simple isochronous movement of the index fingers of the two hands, one on the beat, the other off-beat. In slow tempo such a movement is simple. Still when increasing speed, all of a sudden both fingers do no longer move in a beat/off-beat manner but tap both at the same time. From a standpoint of synchronization this is a sudden phase change. Furthermore, if this regime is reached, the fingers continue moving this way, even if the tempo is slowed down. It takes considerable concentration to return to the beat/off-beat regime. This phenomena is a hysteresis loop, where a system behaves differently when moving forward than backward, in our case speeding up or slowing down. Additionally, at the point of phase change, the speed of the fingers is strongly disturbed and normally slows down considerably. As this happens at the critical point of phase-change, this phenomenon is known as a critical slowing down.

So three features of such a simple beat/off-beat movement of two hands show complex behaviour while is stable during normal tempo. This is typical for synergetic systems. The stability of the system in a regime is due to a nonlinear and complex system behaviour, forcing the system to a stable solution over a wide range of the control parameter, which was the bowing pressure above and is the tempo in this case. The change of the control parameter above a critical point then leads to a break-down of the simple behaviour and suddenly the system behaves completely different.

Now with musical instruments we are able to look at the physical system and therefore find the reason for this behaviour. With the performance of musicians this is not so straightforward, as we cannot look into the neurons of the musician when playing. So the reasoning here, as already discussed above, is one of phenotype. Still what we can do is to propose a model which is able to explain the phenomena, filling in the neurological details in the future, when we may be able to measure and analyse all neurons directly. Such a model was porposed by (Haken et al. 1985) and is known as the Haken/Kelso/Bunz (HKB) model [For sake of space we do not discuss the linear models here, for a review see (Repp 2011)].

The model suggests a behaviour, which can mathematically be described like

$$V = -a\cos\phi - b\cos 2\phi \ , \tag{26}$$

where $0 < \phi < 2\pi$ is the phase relation of the movement, so for $\phi = \pi$ they move in a beat/off-beat manner, for $\phi = 0$ they are in parallel. a and b are parameters changing during performance. Figure 5 shows the potential when plotting Eq. 26 for different relations b/a. The figure shows the development of the system when speeding up the tapping movement (left to right, top to bottom). At first, the movement is at $\phi = \pi$, indicated by the small ball. When decreasing b/a, the potential changes, still the small ball stays at the local minimum at $\phi = \pi$. But at

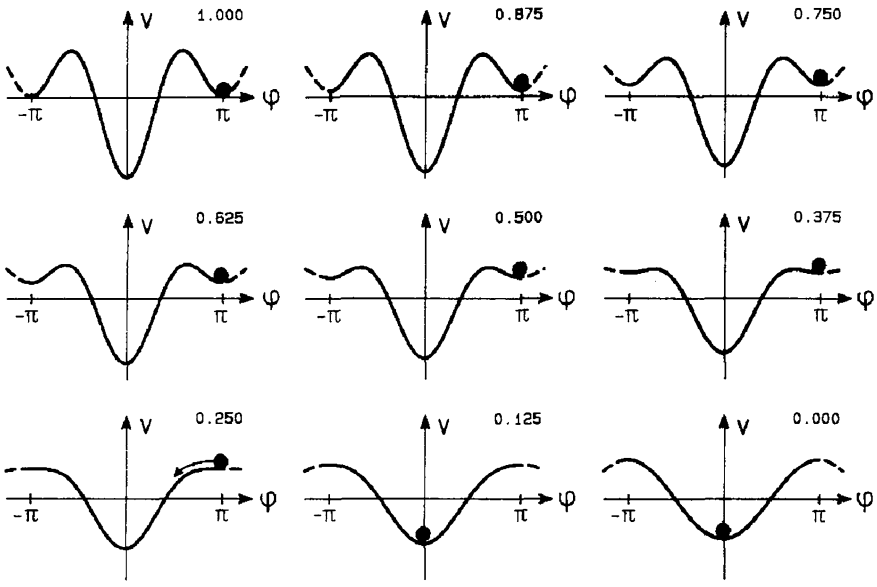


Fig. 5 Potential V/a vs. finger movement phase relation ϕ of the Haken, Kelso, & Bunz model of rhythm production for different values of b/a, representing the speed of the movement, producing bistable potentials up to b/a > 0.25. The ball indicates the phase state of the fingers which move antiparallel for b/a > 0.25. Therefore the dot is at $\phi = \pi$. Above, the *dot* suddenly falls into the well and the phase is $\phi = 0$ for a parallel movement. *From* Haken et al. (1985)

b/a = 0.25 the potential well has no longer such a local minimum and the ball will fall into the global minimum at $\phi = 0$, hands move in parallel. So a sudden phase change appears in such a model.

This is the behaviour one is seeking for in a mathematical formulation. The HKB model is then looking for an underlying differential equation which will result in such a behaviour. Although the whole model is too complex to discuss here in detail, it appears that a kind of van der Pol equation will do, namely

$$\frac{\partial^2 x}{\partial t^2} + (\epsilon_1(x^2 - r_0^2) + \epsilon_2(\frac{\partial x}{\partial t} - \omega^2 r_0^2))\frac{\partial x}{\partial t} + ax = 0 \ . \tag{27}$$

Here, the system variable x appears as second derivative and with the constant a leads to an oscillation. Additionally, a first-order differentiation with respect to time serves as a damping. Still this damping depends on the system variable x, too. As x may be positive or negative, the damping may also be an energy supply, which is the core of the van der Pol equation. In this model, this depends on the amplitude of finger movement r, its frequency ω, and coupling constants ϵ_1 and ϵ_2. Furthermore the van der Pol damping once depends on the system variable, once on its velocity. Then for the phase of oscillation one gets

$$\frac{\partial \phi}{\partial t} = -\frac{2}{2(\omega^2 + \gamma^2)r}((\alpha r - 6\delta r^3)\sin\phi + 3\delta r^3 \sin 2\phi) \ . \tag{28}$$

This solution has the same form as Eq. 26 we wanted to model, where additional constants appear due to a lengthy calculation [for details see (Bader 2013)].

So it is reasonable to assume rhythm production to be caused by a van der Pol equation, a nonlinear equation, where the damping of the system changes with its system parameter itself. Such a model is able to produce the sudden phase change, a hysteresis loop, as well as the critical slowing down of the performance at the phase change.

Another approach is to discuss polyrhythms due to a nonlinear coupled system of oscillators known as the sine-cycle map like in (Haken et al. 1995). This map is an iterative process and can be written as

$$\Theta_{n+1} = \Theta_n + \Omega - (K/(2\pi)\sin(2\pi\Theta_n) \ . \tag{29}$$

The system state Θ at the next time point depends on the previous system state, a coupling $\Omega = \omega_1/\omega_2$ of the two oscillator frequencies ω_1 and ω_2, and the phase of the state with a coupling constant K. A winding number $W(K, \Omega) = (\Theta_2 - \Theta_1)/n$ is defined for the two oscillators for $n \to \infty$. During the iteration it appears that the system can only stay stable, so at a constant winding number, for certain relations between the oscillation frequencies and certain coupling numbers. In Fig. 6 the stable regions are printed in black. d'Arnold tongues appear, favouring simple relations of the two oscillations like 1:2, 1:3, or 2:3. Relations which are more complex appear not to be stable, which is known from rhythm performance. Indeed, this system is also able to cover learning processes by changing the

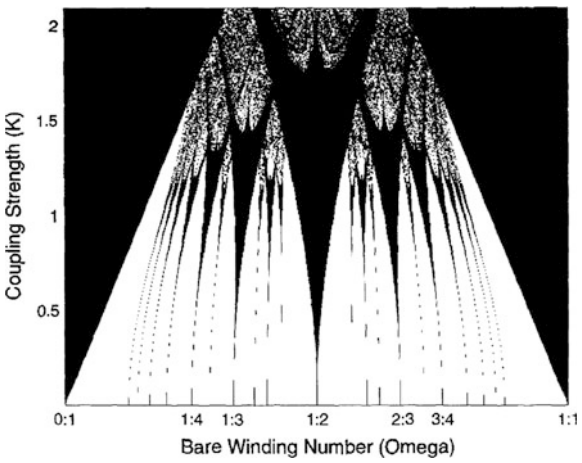


Fig. 6 Sine-circle map of the iteration $\Theta_{n+1} = \Theta_n + \Omega - (K/(2\pi)\sin(2\pi\Theta_n)$ showing stable oscillations (*black*) and unstable oscillations (*white*) for different frequency relations $\Omega = \omega_1/\omega_2$ vs. different coupling strengths K. The black so-called d'Arnold tongues narrow for smaller K which in the model is associated with high tapping frequency making high speed performance much more unstable. *From* Haken et al. 1995

coupling strength, where more complex rhythms can be learned. This is an important feature, as often music shows slight deviations from simple relations, which are performed with a high consistency, like the unstable 3/4 m in a Waltz or the microrhythmic phrasing of swing triplets.

3.3 Tonality

The Riemann idea of tonality as based on the economy of hearing has been discussed above already. Tonality arises because it is more economic to relate tones and consider a common ground for all chords and melodies, the tonality. Still several levels of abstraction point to a complex nature of this referencing. Tonality may be heard while only presenting a single melody or only chords. It may be defined by a cadence or by modal playing. In the history of western art music, in Renaissance polyphony vertical melodic thinking, still the intervals of the voices are crucial, as e.g. in the rules of Johann Fux *gradus ad parnassum*. In the Vienna classic period, music was composed in a homophonic way as a vertical thinking in chord progressions with one main melodic line. There, tonality is established mostly by the chord progressions.

We may consider tonality as a simple statistical process, where the amount of intervals decide about tonality, which also has successfully been performed [see e.g. (Krumhansl 1990)]. Still interval counting is a global view on a musical piece, where all intervals are considered. The establishment of tonality on the other hand is a temporal process, as listeners will decide about it while listening to a piece and not only at the end of it. Alternatively, we may consider tonality and tonal centers as an emergent property araising from combinations of melodies, chords, cadences, and phrases. Again we cannot look into the brain on a neuronal level to decide from a physiological standpoint. So again models have been proposed to cover tonality, tonal centers, or phrases using neural networks as self-organized systems.

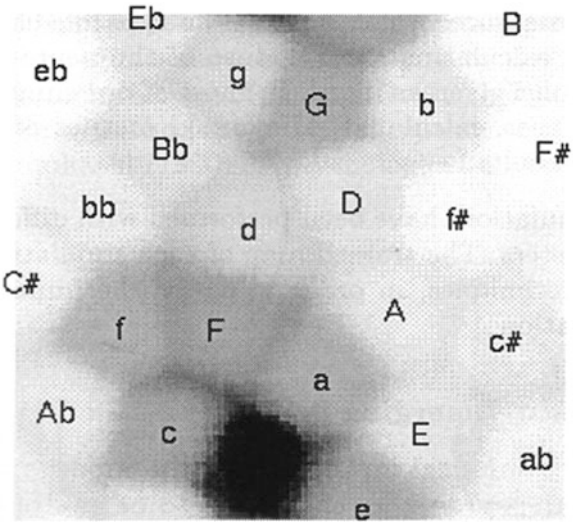


Fig. 7 SOM map of tonal centers clearly showing the cycle of fifth for major (*capital letters*) and minor (*lower case letter*) chords. The map reacts to a *C* major chord (*darkest point*) correctly, but also to similar chords like *F* major, *F* minor, *c* minor, *a* minor, *e* minor, and *G* major with less activation. Again, this map is continuous, connecting left to right and bottom to top. *From* Leman and Carreras (1997)

Two basic approaches have been proposed here, networks of the Kohonen type and connectionist models. The Kohonen nets have been discussed briefly above. One example is the model proposed by Leman and Carreras (1997). The network was trained using major and minor chords, where again the feature vector of the neuron most close to a training chord was slightly changed in the direction of the chord. This was repeated many times for all chords, resulting in a map which indeed shows the tonal relations known from the cycle of fifth, as shown in Fig. 7. The map is cyclic, means that the top is connected with the bottom and the left to the right side. The chords G–D–A–D are in one line, which is cyclic repeated at the bottom with the chord C, which is black in this case, followed by F, B♭, E♭, etc. It is interesting to see that the minor chords are close to the major chords (capital letters). The case shown in the figure is while testing the network with a C—major chord. Black and grey shaping represents the correlation strength between the net and the test chord. Indeed, the C-chord is most strongly met, still chords in the neighbourhood are also slightly correlating, interestingly those in close tonal relations to C—major, like F—major, a—minor, etc.

Another approach is the connectionist approach. Here several layers of neurons are connected in a hierarchical manner, mostly one layer of income neurons, one resulting layer, and one hidden layer in between. The network is trained by stimulating the income neurons which again stimulate the hidden layer depending on weighting functions. These weightings are changed according the reaction of the network. The rules of weighting adaptation are manifold, depending on the model. Often inhibition is performed, where ‘the winner takes it all’ establishes perceptual centers.

In Fig. 8 one example of a neural net learning musical phrases is shown (Gjerdingen 1990). The network is trained by early Mozart pieces and is then able to identify phrases. In the figure two layers are shown, a middle layer, representing

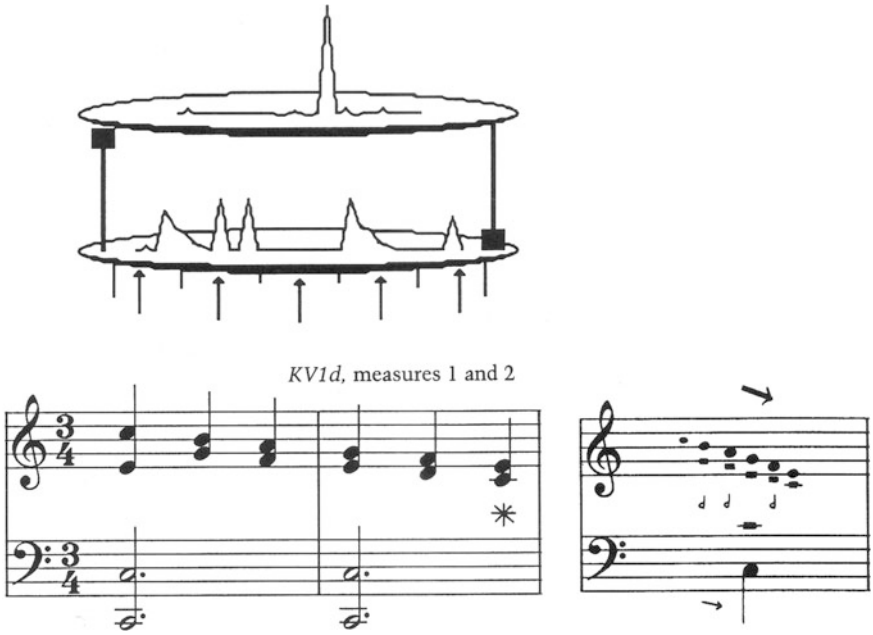


Fig. 8 Mozart KV1d, measures 1 and 2 as original score transposed from F to C (*bottom left*), neural net activation pattern (*top*), and score representation of this activation pattern (*bottom right*), where activation strength is displayed as note size. The asterisk shows the time point at which the activation appears. *From* Gjerdingen (1990)

the events over the two-bar phrase and the top layer. In the net, the different reaction strength of the network is shown which is again plotted on the right side where the size of the notes represent the different weights. This can be taken as Gestalt of the phrase. Then, on the highest level, this phrase causes only one neuron to react, therefore an identification of the phrase by the network is present.

4 Conclusions

This review only contains a small amount of approaches for synchronized, self-organized, or synergetic models of music production, perception, and performance. It appears that the features used in music performance are mainly found with the special features found in these formulations, in terms of articulation, deviations, or synchronization.

Nonlinearities are strongly discussed with musical instruments. The pre-stress and bending of piano soundboards is crucial for the instruments' sound (Mamou-Mani et al. 2008). Longitudinal waves in strings enhance the brightness of the sounds considerably (Bank and Sujbert 2005). Shock waves in trumpets or trombones are the cause of their brassiness (Hirschberg et al. 1996). Synchronization of organ pipes

appear when pipes are close in space and frequency (Abel et al. 2006). Self-sustained oscillations of flute tone production is a bi-stable process including turbulence (Bader 2013). Forced oscillations of string/body couplings cause an enhanced resonance especially for low frequencies to avoid dead spots on the instrument (Bader 2010). Indeed, no instrument does work without nonlinearities. Still this is not enough to understand musical instruments. They appear to be self-organized systems only producing harmonic overtone structures and musical tones at all because of this synchronizing processes. As an additional result, the crucial initial transients of instruments are caused by the struggle between the subsystems during the slaving process. So indeed, the ancient picture of music as based on harmonic relations is not found with musical instruments, contrary, the cause of harmonicity is a strong nonlinear nature of the vibrating systems.

This also seems to hold on the perception side, although with the restriction mentioned that we cannot look into the neural networks of musicians while playing. Another recent approach to this, also presented in this volume, is a free-energy principle (Friston 2010). This model is in the line of Bayes models of hidden-layer structures, like Hidden-Markov models, which are also hot topics in music analysis and production (see also (Nakano et al. 2001) or Braasch in this volume). Efficient neural coding of auditory perception is found with the cat auditory system (Smith and Lewicki 2006). Models of stochastical information processing find dependencies of perception on the deviation of the stimulus rather than on its mean (Rao et al. 2002). Nonlinear distortion in the cochlea is known and discussed for quite some time (Young et al. 2005). Synchronization of brain regions in the 40 Hz regime or by perception tasks like expectation are found on a neuronal basis (Shamir et al. 2009). Oscillatory neurons are strongly discussed in speech processing (Wang 2010). The afferent/efferent coupling in the auditory pathway is also a candidate for such synchronizing behaviour.

Still also many musicians and composers work with fractals or emergent properties of music, e.g. G. Ligeti in his piano sonatas (Bader 2013). Self-similarity has been proposed as a compositional principle in pieces of J.S. Bach (Hofstadter 1980). Algorithms of contemporary music production, like Physical Modeling (Bader and Hansen 2008), use many nonlinearities to produce interesting sounds and come close to real instrument sounds.

So nonlinearities in musical acoustics cause most instruments to produce sounds and harmonic overtone structures at all. With perception the models are often much closer to reality than linear models. It is worth considering these phenomena more closely in the future, both, in terms of production and perception.

References

Abel, M., Bergweiler, S., & Gerhard-Multhaupt, R. (2006). Synchronization of organ pipes: Experimental observations and modeling. *The Journal of the Acoustical Society of America*, *119*, 2467–2475.

Arbi, M. A. (Ed.). (2003). *The handbook of brain theory and neural networks* (2nd ed.). Cambridge: MIT Press.

Aschhoff, V. (1936). Experimentelle Untersuchungen an einer Klarinette. [Experimental investigations of a clarinet]. *Akustische Zeitschrift*, *1*, 77–93.

Bader, R. (2013). Musical acoustics and music psychology. In: Nonlinearities and synchronization. Heidelberg: Springer Series in Systematic Musicology (in print).

Bader, R. (2010). Theoretical framework for initial transient and steady-state frequency amplitudes of musical instruments as coupled subsystems. In: *Proceedings of 20th international symposium on music acoustics (ISMA)* (pp. 1–8), Sydney and Katoomba.

Bader, R., Diezt, M.-K., Elvers, P, Elias, M., & Tolkien, L. (2009). Foundation of a syllogistic music theory. In: Bader, R. (Ed.), *Musical acoustics, neurocognition and psychology of music. Hamburger Jahrbuch fr Musikwissenschaft* (Vol. 25 pp. 177–196).

Bader, R., & Hansen, U. (2008). Acoustical analysis and modeling of musical instruments using modern signal processing methods. In D. Havelock, M. Vorlnder, & S. Kuwano (Eds.), *Handbook of signal processing in acoustics* (pp. 219–247). Berlin: Springer.

Bader, R. (2008). Individual reed characteristics due to changed damping using coupled flow-structure and time-dependent geometry changing finite-element calculation. *Proceedings Forum Acusticum joined with American Acoustical Society Paris*, *08*, 3405–3410.

Bader, R. (2005). *Computational mechanics of the classical Guitar*. Berlin: Springer.

Bader, R. (2005). Turbulent k − ε model of flute-like musical instrument sound production. In: E. Lutton & J. Lévy-Véhel (Eds.), *Fractals in Engineering. New trends in theory and applications* (pp. 109–122). New York: Springer.

Bader, R. & Markuse, B. (1994). Perception and analyzing methods of Groove in popular music. In: *Systematische Musikwissenschaft II/1* (pp. 145–153).

Bandyopadhyay, S., Shamma, S. A., & Kanold, P. O. (2010). Dichotomy of functional organization in the mouse auditory cortex. *Nature Neuroscience*, *13*(3), 361–370.

Bank, B., & Sujbert, L. (2005). Generation of longitudinal vibrations in piano strings: From physics to sound synthesis. *The Journal of the Acoustical Society of America*, *117*(4), 2268–2278.

Bender, D. (2011). *Temporal processing in primate auditory cortex*. Germany: Lambert Academic Publishing.

Beurmann, A. & Schneider, A. (2009). Acoustics and sound of the harpsichord. Another case study. In: Bader, R. (Ed. / Hrsg.), *Musical acoustics, neurocognition and psychology of music, current research at the Institute of Musicology, University of Hamburg. Hamburg Yearbook of Musicology* (p. 25). Frankfurt: Peter Lang Verlag.

Caclin, A., Smith, B. K., & Giard, M.-H. (2008). Interactive processing of timbre space dimensions: An exploration with event-related potentials. *Journal of Cognitive Neuroscience*, *20*, 49–64.

Caclin, A., Brattico, E., Tervaniemi, M., Näänen, R., Morlet, D., Giard, M.-H., et al. (2006). Separate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience*, *18*, 1959–1972.

Coltman, J. W. (1968). Sounding mechanism of the Flute and Organ Pipe. *The Journal of the Acoustical Society of America*, *44*(4), 983–992.

Coltman, J. W. (1968). Acoustics of the Flute. In: *Physics today* (pp. 25–32).

Cosi, P., De Poli, G., & Lauzzana, G. (1994). Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research*, *23*, 71–98.

Cremer, L., & Ising, H. (1967). Die selbsterregten Schwingungen von Orgelpfeifen. [The self-sustained vibrations of organ pipes.]. *Acustica*, *19*, 143–153.

Dalmont, J.-P., Gilbert, J., Kergomard, J., & Ollivier, S. (2005). An analytical prediction of the oscillation and extinction thresholds of a clarinet. *The Journal of the Acoustical Society of America*, *118*(5), 3294–3305.

Damasio, A. R. (2006). *Decartes' error: Emotion*. Vintage: Reason and the Human Brain.

Daniel, P. (2008). Psychoacoustical roughness. In: D. Havelock, S. Kuwano, & Vorländer, M. (Eds.), *Springer handbook of signal processing in acoustics* (pp. 263–274). Berlin: Springer.

Drennan, W. R., & Watson, Ch S. (2001). Sources of variation in profile analysis. II. Component spacing, dynamic changes, and roving level. *Journal of the Acoustical Society of America*, *110*(5), 2498–2504.

Duffour, P., & Woodhouse, J. (2004). Instability of systems with a frictional point contact: Part 1, basic Modelling. *Journal of Sound and Vibration*, *271*, 365–390.

Duffour, P., & Woodhouse, J. (2004). Instability of systems with a frictional point contact: Part 2, model extensions J. of. *Sound and Vibration*, *271*, 391–410.

Durbin, P. A., & Pettersson, R. (2001). *Statistical theory and modeling for turbulent flows*. New York: Wiley.

da Silva, A. R., Scavone, G. P., & van Salstijn, M. (2008). Numerical simulations of fluid-structure interactions in single-reed mouthpieces. *The Journal of the Acoustical Society of America*, *122*(3), 1798–1809.

Dubnov, Sh, & Rodet, X. (2003). Investigation of phase coupling phenomena in sustained portion of musical instruments sound. *The Journal of the Acoustical Society of America*, *113*(1), 348–359.

Elder, S. A. (1973). On the mechanism of sound production in organ pipes. *The Journal of the Acoustical Society of America*, *54*(6), 1554–1564.

Fabre, B., Gilbert, J., Hirschberg, A., & Pelorson, X. (2012). Aeroacoustics of musical instruments. *Annual Review of Fluid Mechanics*, *44*, 1–25.

Fabre, B., & Hirschberg, A. (2000). Physical modeling of flue instruments: A review of lumped models. *Acta Acustica United with Acustica*, *86*, 599–610.

Feiten, B., & Günzel, S. (1994). Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, *18*(3), 53–65.

Fletcher, N. H. (1978). Mode locking in nonlinearly excited inharmonic musical oscillators. *The Journal of the Acoustical Society of America*, *64*, 1566–1569.

Florentine, M., Popper, A. N., & Fay, R. R. (2011). *Loudness* (p. 37). Berlin: Springer Handbook of Auditory Research.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

Garner, W. R. (1974). *The processing of information and structure*. New York: Wiley.

Gjerdingen, R. O. (1990). Categorization of musical patterns by selforganizing neuronlike networks. *Music Perception*, *8*, 339–370.

Green, D. M. (1988). *Profile analysis: Auditory intensity discrimination*. New York: Oxford University Press.

Haken, H. (1990). *Synergetics* (3rd ed.). Berlin: Springer.

Haken, H. (2002). *Brain dynamics*. Berlin: Springer.

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*, 347–356.

Haken, H., Peper, C. E., Beek, P. J., & Daffertshofer, A. (1995). A model for phase transitions in human hand movements during multifrequency tapping. *Physica D*, *90*, 179–196.

Haken, H., & Schiepek, G. (2006). *Synergetik in der Psychologie [Synergetics in psychology]*. Seattle: Hogrefe.

von Helmholtz, H. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage fr die Theorie der Musik. [On the sensation of tone as a physiological basis for the theory of music]*. Braunschweig: Vieweg.

Heerra, P., Peeteres, G., & Dubnow, S. (2003). Automatic classification of musical instrument sounds. *Journal of New Music Research*, *32*, 3–21.

Hirschberg, A., Gilbert, J., Msallam, R., & Wijnands, A. P. J. (1996). Shock waves in trombones. *The Journal of the Acoustical Society of America*, *99*, 1754–1758.

Hofstadter, D. (1980). *Gödel* . Escher. Bach. An eternal golden braid: Vintage Books.

Howe, M. S. (1975). Contributions to the theory of aerodynamic sound, with application to excell jet noise and the theory of the Flute. *Journal of Fluid Mechanics*, *71*(4), 625–673.

Lerdahl, F., & Jackendoff, R. (1983). *Generative theory of tonal music*. Cambridge: Cambridge University Press.

Kaykayoglu, R., & Rockwell, D. (1986). Unstable jet-edge interaction. Part 1. Instantaneous pressure fields at a single frequency. *Journal of Fluid Mechanics*, *169*, 125–149.
Kaykayoglu, R., & Rockwell, D. (1986). Unstable jet-edge interaction Part 2: Multiple frequency pressure fields. *Journal of Fluid Mechanics*, *169*, 151–172.
Kepler, J., (Author) & Caspar, M., (1997). *Weltharmonik [Harmonia Mundi, 1619]*. München: R. Oldenbourg Verlag.
Kepler, J., (Author), Duncan, A. M. (Transl.), Aiton, E. J. & Cohen, I. B., (Eds.). (1981). Mysterium cosmographicum. In: *The secret of the universe [1596]*. New York: Abaris Books.
Keidel, W. D. (1975). *Physiologie des Gehörs. [Physiology of the ear]*. Stuttgart: Thieme.
Klapuri, A. (2006). *Signal processing methods for music transcription*. Berlin: Springer.
Koelsch, S. (2012). *Brain and music*. New York: Wiley.
Kolmogovov, A. N. (1941). The local structure of turbulence in incompressible viscous fluid for vary large Reynolds number. *Dokl. Akad. Nauk SSSR*, *30*, 301–305.
Kostek, B. (2005). Perception-based data processing in acoustics. In: *Applications to music information retrieval and psychophysiology of hearing*. Berlin: Springer.
Krassnitzer, G. (2002). Multiphonics für Klarinette mit deutschem System und andere zeitgenössische Spielarten. [Multiphonics for clarinet with german system and other contemporary styles.] ed. ebenos Verlag Aachen.
Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford: Oxford University Press.
Legge, K. A., & Fletcher, N. H. (1987). Non-linear mode coupling in symmetrically kinked bars. *Jornal of Sound and Vibration*, *118*(1), 23–34.
Leman, M. & Carreras, F. (1997). Schema and gestalt: Testing the hypothesis of psychoneural isomorphism by computer simulation. In: M. Leman (Ed.), *Music, gestalt, and computing. Studies in cognitive and systematic musicology* (pp. 144–168). Berlin: Springer.
Levitin, D. J. (2006). *This is your brain on music*. New York: Dutton.
Lottermoser, W. (1983). *Orgeln, Kirchen und Akustik. [Organs, churches, and acoustics]*. Frankfurt: Verlag Erwin Bochinsky / Das Musikinstrument.
Mamou-Mani, A., Frelat, J., & Besnainou, C. (2008). Numerical simulation of a piano soundboard under downbearing. *The Journal of the Acoustical Society of America*, *123*, 2401–2406.
McIntyre, M. E., Schumacher, R. T., & Woodhouse, J. (1983). On the oscillations of musical instruments. *The Journal of the Acoustical Society of America*, *74*(5), 1325–1345.
Nakanoy, M., Le Rouxz, J., Kameokaz, H., Kitanoy, Y., Onoy, N., & Sagayama, Sh. (2011). Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms. In: *Proceedings of 2011 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA2011)* (pp. 325–328).
Oertel, D. (Ed.). (2002). *Integrated functions in the mammalian auditory pathway* (p. 15). Berlin: Springer Handbook of Auditory Research.
Platon, (Author), Cooper, J. M., & Hutchinson, D. S., (Eds). (1997). Platon. *Complete works*. Indianapolis: Hackett Publishing.
Poeppel, D. (Ed.). (2012). *The human auditory cortex* (p. 43). Berlin: Springer Handbook of Auditory Research.
Rao, R. P. N., Olshausen, B. A., & Lewicki, M. S. (2002). *Probabilistic models of the brain*. Perception and Neural Function: MIT Press.
Rayleigh, Lord. J. W. S. (1945). *The theory of sound (1894)*. New York: Reprint Dover.
Repp, B. H. (2011). Comfortable synchronization of cyclic drawing movements with a metronome. *Human Movement Science*, *30*, 18–39.
Repp, B. H. (2003). Rate limits in sensorimotor synchronization with auditory and visual sequences: The synchronization threshold and the benefits and costs of interval subdivision. *Journal of Motor Behavior*, *35*, 355–370.
Reuter, Ch. (1995). *Der Einschwingvorgan nichtperkussiver Musikinstrumente. [The initial transient of non-percussive musical instruments]*. Bern: Peter Lang Verlag.

Rossing, T. D., & Fletcher, N. H. (1983). Nonlinear vibrations in plates and gongs. *The Journal of the Acoustical Society of America*, *73*, 345–351.

Ryugo, D. K., Fay, R. R., & Popper, A. N. (Eds.). (2011). *Auditory and Vestibular Efferents* (p. 38). Berlin: Springer Series of Auditory Research.

Schneider, A., von Ruschkowski, A., & Bader, R. (2009) Klangliche Rauhigkeit, ihre Wahrnehmung und Messung. [Timbre rougness, its perception and measurement] In: Bader, R. (Ed.), *Musical acoustics, neurocognition and psychology of music* (Vol. 25 pp. 101–144). Germany: Hamburger Jahrbuch fr Musikwissenschaft.

Seever, B. U. (2008). Masking and critical bands. In D. Havelock, S. Kuwano, & M. Vorländer (Eds.), *Springer handbook of signal processing in acoustics* (pp. 229–240). Berlin: Springer.

Shamir, M., Ghitza, O., Epstein, S., & Kopell, N. (2009). Representation of time-varying stimuli by a network exhibiting oscillations on a faster time scale. *PLoS Computational Biology*, *5*(5), e1000370.

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, *439*(23), 978–982.

Steinecke, I., & Herzel, H. (1995). Birurcations in an asymmetric vocal fold model. *The Journal of the Acoustical Society of America*, *97*, 1874–1884.

Stevens, S. S. (1961). Procedure for calculation loudness: Mark VI. *The Journal of the Acoustical Society of America*, *33*, 1577.

Stumpf, C. (1883/1890) *Tonpsychologie*. Bd.1/2.

Thaut, M. (2005). *Rhythm, music, and the brain*. New York: Routledge.

Todd, P., & Loy, G. (Eds.). (1991). *Music and connectionism*. Cambridge: MIT Press.

Toiviainen, P. (1992). The organisation of timbres–A two-stage neural network model. In G. Widmer (Ed.), *Proceedings of the ECAI-92 workshop on AI and Music*. Vienna: Austrian Society for AI.

Touzê, C., & Chaigne, A. (2000). Lyapunov exponents from experimental time series: Application to cymbal vibrations. *Acta Acustica United with Acustica*, *86*, 557–567.

Young, E. D., Yu, J. J., & Reiss, L. A. J. (2005). Non-linearities and the representation of auditory spectra. In M. S. Malmierca & D. R. F. Irvine (Eds.), *Auditory spectral processing* (pp. 136–168). Amsterdam: Elsevier Academic Press.

Vaseghi, S. V. (2007). *Multimedia signal processing: Theory and applications in speech, music, and communications*. New York: Wiley.

Wang, X. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological Reviews*, *90*(3), 1195–1268.

Woodhouse, J., & Schumacher, R. T. (1995). The transient behaviour of models of bowed-string motion. *Chaos*, *5*, 509–523.

A Free Energy Formulation of Music Generation and Perception: Helmholtz Revisited

Karl J. Friston and Dominic A. Friston

1 Introduction

> It is the theory of the sensations of hearing to which the theory of music has to look for the foundation of its structure. (Helmholtz 1877, 4)

This chapter considers music from the point of view of its perception and how acoustic sensations are constructed into musical percepts. Our treatment follows the tradition established by Helmholtz that perception corresponds to inference about the causes of sensations. This therefore requires us to understand the nature of perceptual inference and the formal constraints that this inference places on the nature of music. The basic idea, developed in this chapter, is that music supports the prediction of the unpredictable and that this prediction fulfils a fundamental imperative that we are all compelled to pursue. Heuristically, those activities that we find pleasurable are no more, and no less, than the activities we choose to engage in. The very fact that we can indulge in the same sorts of behaviours repeatedly speaks to the remarkable fact that we are able to maintain a homeostatic exchange with our world—from a physiological to an aesthetic level. We will see later, that this remarkable ability rests upon an active sampling of the sensorium to minimise surprise and fulfil our predictions. In short, music provides the purist opportunity to do what we must do—the opportunity to predict. The opportunity is pure in the sense that musical constructs stand in intimate relation to auditory sensations, perhaps more than any other aesthetic construct:

K. J. Friston (✉)
Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK
e-mail: k.friston@ucl.ac.uk

D. A. Friston
Section of Anaesthetics, Pain Medicine and Intensive Care Imperial College London, 369 Fulham Road, London SW10 9NH, UK

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_2,

> Music stands in a much closer connection with pure sensation than any other art… In music, the sensations of tone are the material of the art. (Helmholtz 1877 2, 3)

A less heuristic version of this thesis can be motivated from the writings of Helmholtz (1866) on the perceptions in general: in brief, we will consider the brain as a Helmholtz machine (Dayan et al. 1995) that actively constructs predictions or explanations for sensory inputs using internal or generative models. This process of active prediction or inference rests upon predicting the causes of sensory input in a way that minimises prediction errors or surprises. If one generalises this notion of minimising surprise or prediction error to action, one obtains a fairly complete explanation for behaviour as the selective sampling of sensory input to ensure that it conforms to our predictions or expectations, as also articulated nicely by Pearce et al. (2010, 302):

> The ability to anticipate forthcoming events has clear evolutionary advantages, and predictive successes or failures often entail significant psychological and physiological consequences. In music perception, the confirmation and violation of expectations are critical to the communication of emotion and aesthetic effects of a composition.

In what follows, we will see that the ability to predict and anticipate is not just of evolutionary advantage, it is a hallmark of any self-organising biological system that endures in an inconstant and changing world. Active inference then puts prediction centre-stage in the action-perception cycle, to the extent it could be regarded as embodied inference: Embodiment in music production and generation is clearly an important formal constraint on the way music is perceived at many levels. For example, as noted by Helmholtz:

> The enigma which, about 2,500 years ago, Pythagoras proposed to science, which investigates the reasons of things, 'Why is consonance determined by the ratios of small whole numbers?' has been solved by the discovery that the ear resolves all complex sounds into pendular oscillations, according to the laws of sympathetic vibration, and it regards as harmonious only such excitements of nerves as continue without disturbance. (Helmholtz 1877, 279)

In other words, the sensory apparatus and neuronal infrastructure responsible for sensing and predicting auditory input places constraints on what we can predict and, according to the current thesis, what is perceived as musical. This is meant in the sense that colour perception is formally constrained by our (three wavelength selective) photoreceptors to lie in a low dimensional perceptual space—despite the fact that the wavelength composition of visual information arriving at the retina is infinite in its dimensionality. Not only is the perception of music constrained by the embodied brains that perceive it—the nature of music also conforms to embodied constraints on production, so music is "within the compass of executants":

> There is nothing in the nature of music itself to determine the pitch of the tonic of any composition…In short, the pitch of the tonic must be chosen so as to bring the compass of the tones of the piece within the compass of the executants, vocal or instrumental. (Helmholtz 1877, 310)

We will exploit this theme of embodied inference throughout this article; paying careful attention to the neuronal structures that can generate and predict music. This is particularly important for music perception that relies upon neuronal dynamics with deep hierarchical structure.

2 Music and Deep (Hierarchical) Structure

Why is music so compelling to listen to? We started with the premise that music affords the opportunity to predict the unpredictable. This predictable unpredictability rests upon the temporally extensive nature of music and its hierarchical dynamics. If we are biological machines that are built (have evolved) to predict, then the deepest most complicated predictions can only be elicited by stimuli that have a multi-layered (deep) hierarchical structure. Hierarchical causal structure is most evident in a separation of temporal scales, in which slower changes contextualise and prescribe faster changes in a recursive fashion. A non-musical example here would be a story (hours) that unfolds on the basis of a narrative (minutes), which entails semantics that emerge from prosody and syntax (seconds); where the semantics themselves depend upon phonological structures (milliseconds) and so on. Music represents a pure (perhaps the purist) example of deep hierarchical dynamical structure, whose prediction involves the resolution of surprise at multiple temporal scales and—by necessity—can only be accomplished by brains that can support similarly structured neuronal dynamics. We will see an example of this later using the production and recognition of bird songs that are composed by separating the temporal scales of hierarchical neuronal dynamics.

But why the prediction of the unpredictable? It is evident in many writings on music perception and appreciation, that the aesthetic qualities of music and its emotive aspects depend upon a resolution of unpredicted excursions or violations of what might have been predicted. It is the resolution of local violations that is afforded by music's hierarchical structure—and the hierarchical models predicting music. A simple example here would be the use of attractors from dynamical systems theory to produce and recognise musical structures—particularly, attractors that support deterministic chaos. Although this may sound fanciful, these (strange) attractors may play a central role in music and song for several reasons: first, the fact that they exhibit deterministic chaos means that the actual dynamics (say amplitude and frequency modulations as a function of time) are unpredictable from any initial conditions yet, at the same time, they evolve according to entirely deterministic rules which—once inferred by the brain—provide perfect predictions of what will happen next; namely, prediction of the unpredictable. We will see an example of this later, using simulated (bird) songs.

Second, the neuronal dynamics (central pattern generators) responsible for the production of musical stimuli, and—from the perspective of this chapter—their perception can be cast as attractors. Crucially, many of the fundamental aspects of music can be captured quite nicely by attractors with chaotic itinerancy (or related

mathematical images called heteroclinic cycles). This is important because it means that we can simulate or model music perception in a biologically plausible way. Furthermore, by hierarchically composing attractors with different time scales, one can model the perception of auditory objects or scenes with deep hierarchical structure. The last section provides a proof of concept of this approach to music perception, using dynamical attractor models of bird song to simulate both perceptual and neurophysiological responses, of the sort that are seen in real brains.

This chapter comprises four sections: In the first, we briefly review the literature on music and prediction with a special emphasis on the neuroscience of music—as it relates to unconscious inference. The second section introduces an abstract and broad theoretical framework that motivates the importance of prediction and minimising surprise in terms of the free energy principle and active inference. This section is a bit technical but sets up the formalism for the third section that introduces plausible neuronal architectures that minimise surprise (or more exactly maximise free energy) through predictive coding. In the final section, we consider some canonical examples of song perception using simulations of bird song and the predictive coding scheme of the preceding section. These examples illustrate the basic phenomenology and illustrate some ubiquitous phenomena in the neurosciences, like omission responses and categorisation, as measured both psychophysically and electrophysiologically.

3 Prediction in Music and Cognition

> Predictive information processing is fundamental to music in three ways. (1) Prediction and expectancy incorporate the essence of the dynamics of musical temporality. Further they make the experience of local or large-scale goal-directed processes in music possible (based on, e.g., melodic, harmonic or modal features). (2) Predictive processing constitutes a major process involved in musical interaction and synchronisation. (3) Finally, processes of expectancy and prediction are understood to be linked with specific emotional and aesthetic musical effects. (Rohrmeier and Koelsch 2012)

The Helmholtzian view considers the brain as a learning and inference system—assimilating prior beliefs and violated predictions to predict future events as accurately and as parsimoniously as possible. The evolutionary benefits of such a system are clear: it is through a constant updating of our internal model of the world that our interactions with the world are nuanced and optimised. Prediction consequently plays a deep-seated role in all cognition, including that of music. Predictive processing is fundamental to music in three ways (Rohrmeier and Koelsch 2012): it accommodates musical temporality and underlies musical interactions and synchronisation. Furthermore, it plays a key role in mediating the emotive and aesthetic effects of music.

Meyer (1956) proposed that by confirming or violating the listener's musical expectations—and thereby conveying suspense or resolution—music generates an emotional response. However, to recognise the musical qualities of auditory sensations, we must infer the rules or causal structures that underlie our expectations. This structure is manifest over several levels within music, such as melody and harmony, and entails the recognition of rhythmic or metrical structure. Investigations of Meyer's proposal have focused on individual musical features and—consistent with the prominent contribution of harmony to Western styles of music—the processing of harmonic violations has received much attention. Responses reflecting violated predictions, induced by the preceding harmonic context, are measurable with electroencephalography (EEG). For example, infrequent and unpredictable chords, within chord sequences, elicit an early right-anterior negativity (ERAN) and a late bilateral-frontal negativity (N5) in the event related response (ERP) of the listener (Koelsch et al. 2000). These response components are thought to reflect the violation of harmonic expectations and higher processes of harmonic integration into the on-going musical context respectively. Both of their amplitudes are sensitive to the degree of expectance and the probability of harmonic deviation. Further research showed that they are evoked irrespective of whether the stimulus is attended to (Koelsch et al. 2002; Loui et al. 2005). Furthermore, while the N5 is influenced by emotional expression in the performance of a piece (i.e., deliberate variations in loudness and tempo) the ERAN is not (Koelsch et al. 2008). While these components reflect the brain's capacity to establish expectations given a harmonic context, other studies have demonstrated a late positive component with violation of melodic expectations (Besson and Faita 1995; Verleger 1990)—a prediction—dependent response that is modulated by expertise and familiarity.

While the expression of neural responses to violation of musical expectation is established, the implications of these findings for the emotive effects of music are less well understood; due in part to the difficulty of measuring emotional responses. The possibility that emotional attribution is engaged in violation paradigms has emerged as an intriguing prospect from functional neuroimaging studies: unexpected chords activate both the orbital frontolateral cortex (OFLC)—a paralimbic region associated with evaluating a stimulus' emotional salience—and the anterior insula, associated with autonomic responses to emotionally valent stimuli (Koelsch et al. 2005). Interestingly, similar OFLC activation, in subjects listening to classical music, could be prevented when scrambling music, hence disrupting its structure (Levitin and Menon 2003). The authors attributed this activation to the recognition of fine-structured stimuli that evolve over extended periods of time. The absence of this activation, with scrambled music, suggests that high-level (emotive) attributes of music are associated with longer timescales, as suggested by the studies of the (late) N5 ERP components above.

Some studies have used retrospective subjective accounts of emotional and autonomic responses to music, analysing the musical structure of cited excerpts to identify evocative musical characteristics. For example, violations of harmonic

expectation evoke spine shivers, while a musical phrase occurring earlier than expected reliably increases heart rate (Sloboda 1991). While promising, the link between expectation violation and emotion in such approaches is weakened by the lack of an objectively measurable emotional response. Steinbeis et al. (2006) addressed this by combining the use of subjective scales of tension and emotionality with the recording of heart rate and electrodermal activity (EDA)—physiological measures consistently associated with emotional processing. Harmonic expectation was violated via modification of a single chord in each of six Bach chorales. Tension, emotionality and EDA were found to increase with the degree of harmonic expectation violation, which was also reflected in concurrently recorded ERPs (as discussed above).

While the evidence is intriguing, there are some caveats to consider. For example, violation studies are limited by their design, as the violating stimuli do not occur naturally. Furthermore, harmonic studies in particular typically employ paradigms that assess expectation associated with final chords, so the findings may only apply to tonal closure. The generalisation of violation-dependent responses to all aspects of music perception may be more challenging, because some attributes are more predictable than others. For example, while forthcoming elements of melodic or harmonic expectation may be clearly defined, this is not the case for more complex features, such as key structure, that do not necessarily apply to the next musical event, nor to an unambiguous point in time. Additionally, the interplay of attention and prediction in complex styles of music, with non-aligned (e.g. polyrhythmic) features, remains poorly understood (Rohrmeier and Koelsch 2012). Although (unsupervised) statistical learning models can predict single features such as melody (Pearce et al. 2010), predictive models of complex music are still in developmental stages. Nevertheless, the findings thus far support Meyer's proposal, in which music engages the base mechanisms by which we come to understand our environment. The evidence for violation of expectation in emotion discussed here may underlie the initiation of music's intrigue; as Meyer wrote,

> Such states of doubt and confusion are abhorrent. When confronted with them, the mind attempts to resolve them into clarity and certainty. Meyer (1956)

It may indeed be by a flirtatious generation of disorder and its subsequent resolution that music communicates its emotional effect.

3.1 Summary

In summary, the very fact that neuronal responses to the violation of musical predictions can be elicited speaks to the fact that the brain can construct expectations or predictions about the temporal structure of music. Furthermore, the empirical evidence suggests that these predictions have a hierarchical aspect, in which higher level predictions pertain to a longer timescales. Circumstantial evidence suggests that high-level attributes have an emotive dimension; which,

from the point of view of active inference, mean that these representations provide both exteroceptive (auditory) and interoceptive (autonomic) predictions that may explain the visceral responses that music can elicit—visceral responses associated with the violation and resolution of high-level predictions. These conclusions rest upon a hierarchical model of musical structure that is characterised by a separation of temporal scales. In the next section, we will consider the form of hierarchical models that the brain might use; starting with a purely formal or mathematical description that will be used in the subsequent section to understand the neuronal circuits that might underlie hierarchical inference or predictive coding in the brain.

4 Hierarchical Models and Bayesian Inference

In the introduction, we hinted at the fundamental imperative for all self-organising biological systems to minimise their surprise and actively engage with the environment to selectively sample predicted sensations—this is known as active inference. In the previous section, we saw that the brain can predict the underlying causal structure of (musical) sensory streams over multiple timescales; in other words it can infer the hidden causes of its sensory input. In the remainder of this chapter, we try to put these things together and provide a formal model of perception that can be used to illustrate the nature of perception and prediction. In brief, to maintain a homoeostatic exchange with the environment biological systems must counter the dispersion of their sensory states due to environmental fluctuations. In terms of information theory, this means biological systems must minimise the entropy of their sensory states. This can be achieved by minimising the surprise associated with sensory states at each point in time, because (under ergodic assumptions) the long-term average of surprise is entropy. Crucially, negative surprise is the logarithm of Bayesian evidence. This means that minimising surprise is the same as maximising the evidence for a model of the world entailed by the structure and dynamics of the biological system. In other words, we are all obliged to be Bayes-optimal modellers of our sensorium. In what follows, we consider in more detail the nature of the models that the brain may use to guide active inference and, implicitly, make predictions about hidden causes of sensory input.

4.1 Hierarchical Dynamic Models

Hierarchical dynamic models are probabilistic generative models $p(s, \psi) = p(s|\psi)p(\psi)$ based on state-space models. They entail the likelihood $p(s|\psi)$ of getting some sensory data $s(t)$ given some parameters $\psi = \{x, v, \theta\}$ and priors on those parameters $p(\psi)$. We will see that the parameters subsume different quantities, some of which change with time and some which do not. A dynamic model can be written as

$$\begin{aligned} s &= g(x, v) + \omega_s \\ \dot{x} &= f(x, v) + \omega_x \end{aligned} \tag{1}$$

The continuous nonlinear functions (f, g) of the states are parameterized by θ. The states $v(t)$ can be deterministic, stochastic, or both. They are referred to as sources, or causes. The states $x(t)$ mediate the influence of the input on the output and endow the system with memory. They are often referred to as hidden states because they are seldom observed directly. We assume the random fluctuations (i.e., observation noise) $\omega(t)$ are analytic, such that the covariance of the generalised fluctuations $\tilde{\omega} = (\omega, \omega', \omega'', \ldots)$ is well defined. Generalised states include the state itself and all higher order temporal derivatives.

The first (observer) equation above shows that the hidden states (x, v) are needed to generate an output or sensory data. The second (state) equation enforces a coupling between orders of motion of the hidden states and confers memory on the system. Gaussian assumptions about the fluctuations $p(\tilde{\omega}) = \mathcal{N}(0, \Sigma)$ provide the likelihood of any given sensory data and (empirical) priors over the motion of hidden states. It is these empirical priors that can be exploited by the brain to make predictions about the dynamics or trajectories of hidden states causing sensory input. Hierarchical dynamic models have the following form, which generalizes the model in Eq. 1

$$\begin{aligned} s &= g(x^{(1)}, v^{(1)}) + \omega_v^{(1)} \\ \dot{x}^{(1)} &= f(x^{(1)}, v^{(1)}) + \omega_x^{(1)} \\ &\vdots \\ v^{(i-1)} &= g(x^{(i)}, v^{(i)}) + \omega_v^{(i)} \\ \dot{x}^{(i)} &= f(x^{(i)}, v^{(i)}) + \omega_x^{(i)} \\ &\vdots \end{aligned} \tag{2}$$

Again, $f^{(i)} = f(x^{(i)}, v^{(i)})$ and $g^{(i)} = g(x^{(i)}, v^{(i)})$ are continuous nonlinear functions of the states. The random innovations $\omega^{(i)}$ are conditionally independent fluctuations that enter each level of the hierarchy. These play the role of observation error or noise at the first level and induce random fluctuations in the states at higher levels. The causal states $v = (v^{(1)}, v^{(2)}, \ldots)$ link levels, whereas the hidden states $x = (x^{(1)}, x^{(2)}, \ldots)$ link dynamics over time. In hierarchical form, the output of one level acts as an input to the next. Inputs from higher levels can enter nonlinearly into the state equations and can be regarded as changing its control parameters to produce complicated convolutions with "deep" (i.e., hierarchical) structure.

The conditional independence of the fluctuations means that these models have a Markov property over levels (Efron and Morris 1973), which simplifies the architecture of attending inference schemes. See Kass and Steffey (1989) for a discussion of approximate Bayesian inference models of static data and Friston (2008) for dynamic models. In short, a hierarchical form endows models with the

ability to construct their own priors. For example, the prediction $\tilde{g}^{(i)} = \tilde{g}(\tilde{x}^{(i)}, \tilde{v}^{(i)})$ plays the role of a prior expectation on $\tilde{v}^{(i-1)}$, yet it has to be estimated in terms of $(\tilde{x}^{(i)}, \tilde{v}^{(i)})$. This feature is central to many inference and estimation procedures, ranging from mixed-effects analyses in classical covariance component analysis to automatic relevance determination in machine learning.

4.2 Summary

This section has introduced hierarchical dynamic models (in generalized coordinates of motion). These models are about as complicated as one could imagine; they comprise causal and hidden states, whose dynamics can be coupled with arbitrary (analytic) nonlinear functions. Furthermore, these states can have random fluctuations with unknown amplitude and arbitrary (analytic) autocorrelation functions. A key aspect of these models is their hierarchical structure, which induces empirical priors on the causes. These complement the constraints on hidden states, furnished by empirical priors on their motion or dynamics. Later, we will examine the roles of these structural and dynamical priors in perception. We now consider how these models are inverted to disclose the unknown states generating observed sensory data.

4.3 Model Inversion (inference) and Variational Bayes

The concluding part of this section considers model inversion and provides a heuristic summary of the material in Friston (2008). A generative model maps from hidden causes or states to sensory consequences. Recognition (model inversion) inverts this mapping to infer the hidden causes from sensations. We will focus on variational Bayes, which is a generic approach to model inversion that approximates the conditional density $p(\psi|\tilde{s})$ over the unknown states and parameters, given some data. This approximation is achieved by optimizing the sufficient statistics of a recognition density $q(\tilde{\psi})$ over the hidden generalised states, with respect to a lower bound on the log-evidence $\ln p(\tilde{s}|m)$ of the model m (Feynman 1972; Hinton and von Camp 1993; MacKay 1995; Neal and Hinton 1998; Friston 2005; Friston et al. 2006). The log-evidence or negative surprise can be expressed in terms of a free-energy and divergence term

$$\begin{aligned} \ln p(\tilde{s}|m) &= F + D_{KL}(q(\tilde{\psi})||p(\tilde{\psi}|\tilde{s}, m)) \Rightarrow \\ F &= \left\langle \ln p(\tilde{s}, \tilde{\psi}) \right\rangle_q - \left\langle \ln q(\tilde{\psi}) \right\rangle_q \end{aligned} \tag{3}$$

The free-energy comprises an energy term, corresponding to a Gibb's energy, $G(\tilde{s}, \tilde{\psi}) := \ln p(\tilde{s}, \tilde{\psi})$ expected under the recognition density and its entropy. Equation 3 shows that the free energy is a lower-bound on the log-evidence because the divergence term is, by construction, nonnegative. The objective is to optimize the sufficient statistics of the recognition density by maximising the free-energy and minimizing the divergence. This ensures $q(\tilde{\psi}) \approx p(\tilde{\psi}|\tilde{s}, m)$ becomes an approximate posterior density.

Invoking the recognition density converts a difficult integration problem (inherent in computing the evidence) into an easier optimization problem. We now seek a recognition density that maximizes the free energy at each point in time. In what follows, we will assume the hidden parameters are known and focus on the hidden states $u = (x, y)$. To further simplify things, we will assume the brain uses something called the Laplace approximation. This enables us to focus on a single quantity for each unknown state, the conditional expectation or mean. Under the Laplace approximation, the conditional density assumes a fixed Gaussian form $q(\tilde{u}) = \mathcal{N}(\tilde{\mu}, C)$ with sufficient statistics $(\tilde{\mu}, C)$, corresponding to the conditional expectation and covariance of the hidden states. The advantage of the Laplace approximation is that the conditional covariance is a function of the mean (the inverse curvature of the Gibbs energy at the expectation). This means we can reduce model inversion to optimizing one sufficient statistic; namely, the conditional expectation or mean. This is the solution to

$$\dot{\tilde{\mu}} - D\tilde{\mu} = \partial_u F \tag{4}$$

Here, $\dot{\tilde{\mu}} - D\tilde{\mu}$ can be regarded as motion in a frame of reference that moves with the predicted generalised motion $D\tilde{\mu}$, where D is a matrix derivative operator. Critically, the stationary solution (in this moving frame of reference) maximizes free energy. At this point the mean of the motion becomes the motion of the mean, $\dot{\tilde{\mu}} = D\tilde{\mu}$ and $\partial_u F = 0$. Those people familiar with Kalman filtering will see that Eq. 4, can be regarded as (generalised) Bayesian filtering, where the change in conditional expectations $\dot{\tilde{\mu}} = D\tilde{\mu} + \partial_u F$ comprises a prediction and a correction term that depends upon free energy or—as we will see in the next section—prediction error.

4.4 Summary

In this section, we have seen how the inversion of dynamic models can be formulated as an optimization of free energy. By assuming a fixed-form (Laplace) approximation to the conditional density, one can reduce optimization to finding the conditional means of unknown quantities. For the hidden states, this entails finding a path or trajectory that maximizes free energy. This can found by making the motion of the generalized mean perform a gradient ascent in a frame of

reference that moves with the mean of the generalized motion. The only thing we need to implement this recognition scheme (generalised Bayesian filtering) is the Gibbs energy $G(\tilde{s}, \tilde{\psi}) := \ln p(\tilde{s}, \tilde{\psi})$. This is specified completely by the generative model (Eq. 2). In the next section, we look at what this scheme might look like in the brain—and see that it corresponds to something called predictive coding.

5 Hierarchical Models in the Brain

A key architectural principle of the brain is its hierarchical organization (Felleman and van Essen 1991; Maunsell and van Essen 1983; Mesulam 1998; Zeki and Shipp 1988). This has been established most thoroughly in the visual system, where lower (primary) areas receive sensory input and higher areas adopt a multimodal or associational role. The neurobiological notion of a hierarchy rests upon the distinction between forward and backward connections (Angelucci et al. 2002; Felleman and Van Essen 1991; Murphy and Sillito 1987; Rockland and Pandya 1979; Sherman and Guillery 1998). This distinction is based upon the specificity of cortical layers that are the predominant sources and origins of extrinsic connections. Forward connections arise largely in superficial pyramidal cells, in supra-granular layers, and terminate on spiny stellate cells of layer four in higher cortical areas (DeFelipe et al. 2002; Felleman and Van Essen 1991). Conversely, backward connections arise largely from deep pyramidal cells in infra-granular layers and target cells in the infra and supra-granular layers of lower cortical areas. Intrinsic connections mediate lateral interactions between neurons that are a few millimetres away. There is a key functional asymmetry between forward and backward connections that renders backward connections more modulatory or nonlinear in their effects on neuronal responses (e.g., Sherman and Guillery 1998; see also Hupe et al. 1998). This is consistent with the deployment of voltage-sensitive NMDA receptors in the supra-granular layers that are targeted by backward connections (Rosier et al. 1993). Typically, the synaptic dynamics of backward connections have slower time constants. This has led to the notion that forward connections are driving and elicit obligatory responses in higher levels, whereas backward connections have both driving and modulatory effects and operate over larger spatial and temporal scales.

5.1 Bayesian Filtering and Predictive Coding

This hierarchical structure of the brain speaks to hierarchical models of sensory input. We now consider how the brain‘s functional architecture can be understood as inverting hierarchical models (recovering the hidden causes of sensory input). If we assume that the activity of neurons encodes the conditional mean of states, then Eq. 4 specifies the neuronal dynamics entailed by perception or recognizing states

of the world from sensory data. In Friston (2008), we show how these dynamics can be expressed simply in terms of prediction errors on the causes and motion of the hidden states. Using these errors, we can write Eq. 4 as

$$\begin{aligned}\dot{\tilde{\mu}}_v^{(i)} &= \mathcal{D}\tilde{\mu}_v^{(i)} - \partial_{\tilde{v}}\tilde{\varepsilon}^{(i)} \cdot \xi^{(i)} - \xi_v^{(i+1)} \\ \dot{\tilde{\mu}}_x^{(i)} &= \mathcal{D}\tilde{\mu}_x^{(i)} - \partial_{\tilde{x}}\tilde{\varepsilon}^{(i)} \cdot \xi^{(i)} \\ \\ \xi_v^{(i)} &= \Pi_v^{(i)}(\tilde{\mu}_v^{(i-1)} - g^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)})) \\ \xi_x^{(i)} &= \Pi_x^{(i)}(\mathcal{D}\tilde{\mu}_x^{(i)} - f^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}))\end{aligned} \tag{5}$$

This scheme describes a gradient descent on the (sum of squared) prediction error—or more exactly precision-weighted prediction errors $\xi^{(i)} = \Pi^{(i)}\tilde{\varepsilon}^{(i)}$, where precision $\Pi^{(i)}$ corresponds to the reliability (inverse covariance) of the prediction error $\tilde{\varepsilon}^{(i)}$ at the i-th level of the hierarchy. The first pair of equalities just says that conditional expectations about hidden causes and states $(\tilde{\mu}_v^{(i)}, \tilde{\mu}_x^{(i)})$ are updated based upon the way we would predict them to change—the first term—and subsequent terms that minimise prediction error. The second pair expresses prediction error as the conditional expectations about hidden causes and (the changes in) hidden states minus their predicted values.

It is difficult to overstate the generality and importance of Eq. 5—it grandfathers nearly every known statistical estimation scheme, under parametric assumptions about additive noise. These range from ordinary least squares to advanced Bayesian filtering schemes (see Friston 2008). In this general setting, Eq. 5 corresponds to (generalised) predictive coding. Under linear models, it reduces to linear predictive coding, also known as Kalman-Bucy filtering.

5.2 *Predictive Coding and Message Passing in the Brain*

In neuronal network terms, Eq. 5 says that prediction error units receive messages based on expectations in the same level and the level above. This is because the hierarchical form of the model only requires expectations between neighbouring levels to form prediction errors. Conversely, expectations are driven by prediction error in the same level and the level below—updating expectations about hidden states and causes respectively. This updating corresponds to an accumulation of prediction errors—in that the rate of change of conditional expectations is proportional to prediction error. This means expectations are proportional to the integral or accumulation of prediction errors over time. Crucially, this accumulation requires only the prediction error from the lower level and the level in question. These constitute the bottom-up and lateral messages that drive conditional expectations to provide better predictions—or representations—which suppress the prediction error. Electrophysiologically, this means that one would

expect to see a transient prediction error response to bottom-up afferents (in neuronal populations encoding prediction error) that is suppressed to baseline firing rates by sustained responses (in neuronal populations encoding predictions). This is the essence of recurrent message passing between hierarchical levels to suppress prediction error (see Fig. 1 and Friston 2008 for a more detailed discussion).

We can identify error-units with superficial pyramidal cells, because the only messages that pass up the hierarchy are prediction errors and superficial pyramidal cells originate forward connections in the brain. This is useful because it is these cells that are primarily responsible for electroencephalographic (EEG) signals that can be measured noninvasively. Similarly, the only messages that are passed down the hierarchy are the predictions from state-units that are necessary to form

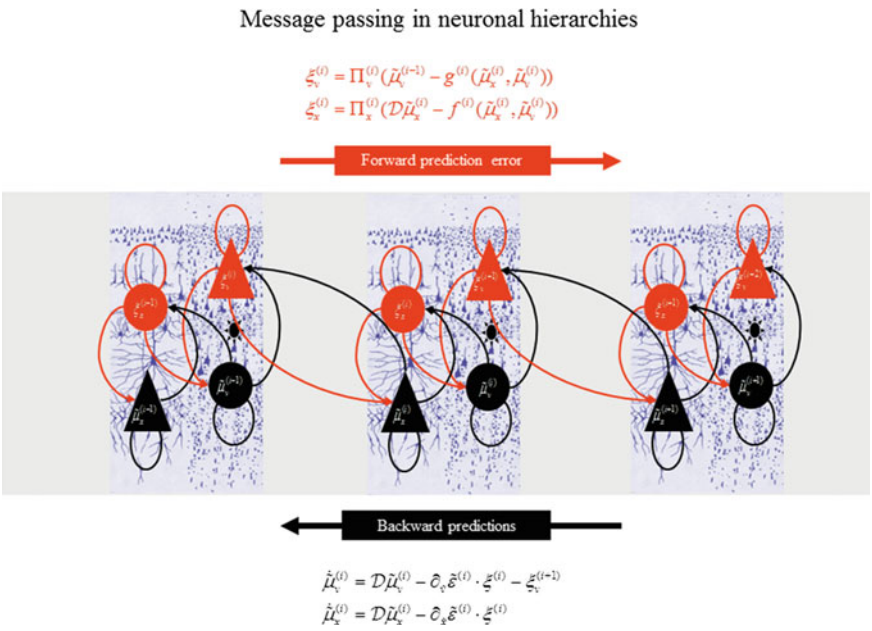


Fig. 1 Schematic, detailing the neuronal architectures that encode a recognition density over the hidden states of a hierarchical model. This schematic shows the speculative cells of origin of forward (*driving*) connections that convey prediction error from a lower area to a higher area and the backward connections that are used to construct predictions. These predictions try to explain away input from lower areas by suppressing prediction error. In this scheme, the sources of forward connections are the superficial pyramidal cell population, and the sources of backward connections are the deep pyramidal cell population. The differential equations relate to the optimization scheme detailed in the main text. The state-units and their efferents are in black and the error-units in red, with causes on the right and hidden states on the left. For simplicity, we have assumed the output of each level is a function of, and only of, the hidden states. This induces a hierarchy over levels and, within each level, a hierarchical relationship between states, where causes predict hidden states. This schematic shows how the neuronal populations may be deployed hierarchically within three cortical areas (or *macro-columns*)

prediction errors in lower levels. The sources of extrinsic backward connections are the deep pyramidal cells, and one might deduce that these encode the expected causes of sensory states (see Mumford 1992 and Fig. 1). Crucially, the motion of each state-unit is a linear mixture of bottom-up prediction error (Eq. 5). This is exactly what is observed physiologically, in that bottom-up driving inputs elicit obligatory responses that do not depend on other bottom-up inputs. The prediction error itself is formed by predictions conveyed by backward and lateral connections. These influences embody the nonlinearities implicit in $\tilde{g}^{(i)}$ and $\tilde{f}^{(i)}$. Again, this is entirely consistent with the nonlinear or modulatory characteristics of backward connections.

5.3 Summary

In summary, we have seen how the inversion of a generic hierarchical and dynamical model of sensory inputs can be transcribed onto neuronal quantities that optimize a variational free energy bound on the evidence for that model. This optimization corresponds, under some simplifying assumptions, to suppression of prediction error at all levels in a cortical hierarchy. This suppression rests upon a balance between bottom-up (prediction error) influences and top-down (empirical prior) influences. In the final section, we use this scheme to simulate neuronal responses. Specifically, we pursue the electrophysiological correlates of prediction error and ask whether we can understand the violation phenomena in event-related potential (ERP), discussed in Sect. 2, in terms of hierarchical inference and message passing in the brain.

6 Birdsong and Attractors

In this section, we examine the emergent properties of a system that uses hierarchical dynamics or attractors as generative models of sensory input. The example we use is birdsong, and the empirical measures we focus on are local field potentials (LFPs) or evoked (ERP) responses that can be recorded noninvasively. Our aim is to show that canonical features of empirical electrophysiological responses can be reproduced easily under attractor models of sensory input. Furthermore, in a hierarchical setting, the use of dynamic models has some interesting implications for perceptual infrastructures (Kiebel et al. 2008). The examples in this section are taken from Friston and Kiebel (2009), to which the reader is referred for more details.

We first describe the model of birdsong and demonstrate the nature and form of this model through simulated lesion experiments. This model is then used to reproduce the violation-dependent responses discussed in Sect. 2 using, perhaps, the most profound form of violation; namely, an omission of an expected event.

We will then use simplified versions of this model to show how attractors can be used to categorize sequences of stimuli quickly and efficiently. Throughout this section, we will exploit the fact that superficial pyramidal cells are the major contributors to observed LFP and ERP signals, which means we can ascribe these signals to prediction error because the superficial pyramidal cells are the source of bottom-up messages in the brain (see Fig. 1).

6.1 Attractors in the Brain

The basic idea in this chapter is that the environment unfolds as an ordered sequence of spatiotemporal dynamics, whose equations of motion entail attractor manifolds that contain sensory trajectories. Critically, the shape of the manifold generating sensory data is itself changed by other dynamical systems that could have their own attractors. If we consider the brain has a generative model of these coupled dynamical systems, then we would expect to see attractors in neuronal dynamics that are trying to predict sensory input. In a hierarchical setting, the states of a high-level attractor enter the equations of motion of a low-level attractor in a nonlinear way, to change the shape of its manifold. This form of generative model has a number of sensible and appealing characteristics.

First, at any level the model can generate and therefore encode structured sequences of events, as the states flow over different parts of the manifold. These sequences can be simple, such as the quasi-periodic attractors of central pattern generators (McCrea and Rybak 2008) or can exhibit complicated sequences of the sort associated with chaotic and itinerant dynamics (e.g., Breakspear and Stam 2005; Canolty et al. 2006; Friston 1997; Haken Kelso et al. 1990; Jirsa et al. 1998; Kopell et al. 2000; Rabinovich et al. 2008).

Second, hierarchically deployed attractors enable the brain to generate and therefore predict or represent different categories of sequences. This is because any low-level attractor embodies a family of trajectories that correspond to a structured sequence. The neuronal activity encoding the particular state at any one time determines where the current dynamics are within the sequence, while the shape of the attractor manifold determines which sequence is currently being expressed. In other words, the attractor manifold encodes what is being perceived and the neuronal activity encodes where the current percept is located on the manifold or within the sequence.

Third, if the state of a higher attractor changes the manifold of a subordinate attractor, then the states of the higher attractor come to encode the category of the sequence or dynamics represented by the lower attractor. This means it is possible to generate and represent sequences of sequences and, by induction, sequences of sequences of sequences, and so on. This rests upon the states of neuronal attractors at any cortical level providing control parameters for attractor dynamics at the level below. This necessarily entails a nonlinear interaction between the top-down

effects of the higher attractor and the states of the recipient attractor. Again, this is entirely consistent with the known functional asymmetries between forward and backward connections and speaks to the nonlinear effects of top-down connections in the real brain.

Finally, this particular model has implications for the temporal structure of perception and, in particular, music. Put simply, the dynamics of high-level representations unfold more slowly than the dynamics of lower level representations. This is because the state of a higher attractor prescribes a manifold that guides the flow of lower states. In the limiting case of the higher level having a fixed-point attractor, its fixed states will encode lower level dynamics, which could change quite rapidly. We will see an example of this later, when considering the perceptual categorization of different sequences of chirps subtending birdsongs. This attribute of hierarchically coupled attractors enables the representation of arbitrarily long sequences of sequences and suggests that neuronal representations in the brain will change more slowly at higher levels (Kiebel et al. 2008; see also Botvinick et al. 2007; Hasson et al. 2008). One can turn this argument on its head and use the fact that we are able to recognize sequences of sequences (e.g., Chait et al. 2007) as an existence proof for this sort of generative model. In the examples that follow, we will try to show how autonomous dynamics furnish generative models of sensory input, which behave much like real brains, when measured electrophysiologically.

6.2 A Synthetic Avian Brain

The toy example used here deals with the generation and recognition of birdsongs (Laje and Mindlin 2002). We imagine that birdsongs are produced by two time-varying control parameters that control the frequency and amplitude of vibrations emanating from the syrinx of a songbird (see Fig. 2). There has been an extensive modelling effort using attractor models at the biomechanical level to understand the generation of birdsong (e.g., Laje et al. 2002). Here, we use the attractors at a higher level to provide time-varying control over the resulting sonograms. We drive the syrinx with two states of a Lorenz attractor, one controlling the frequency (between 2 and 5 kHz) and the other (after rectification) controlling the amplitude or volume. The parameters of the Lorenz attractor were chosen to generate a short sequence of chirps every second or so. To endow the generative model with a hierarchical structure, we placed a second Lorenz attractor, whose dynamics were an order of magnitude slower, over the first. The states of the slower attractor entered as control parameters (known as the Rayleigh and Prandtl number) to control the dynamics exhibited by the first. These dynamics could range from a fixed-point attractor, where the states of the first are all zero, through to quasi-periodic and chaotic behaviour, when the value of the Prandtl number exceeds an appropriate threshold (about 24) and induces a bifurcation. Because higher states

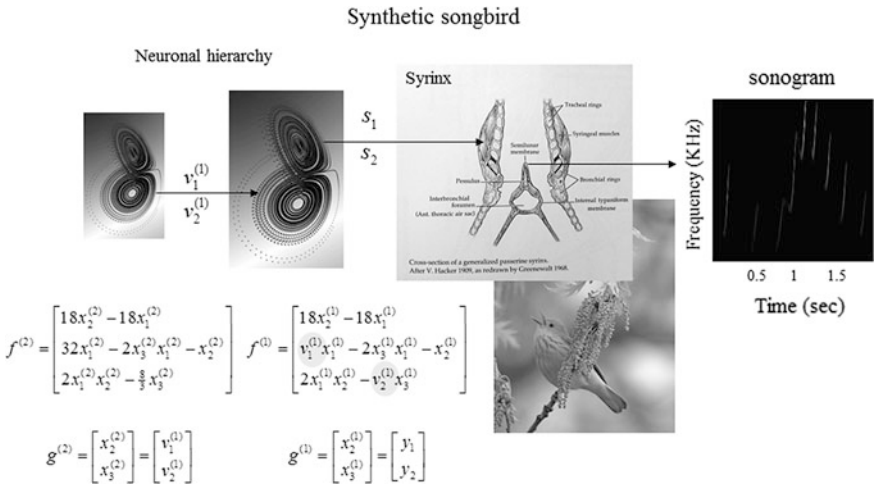


Fig. 2 Schematic, showing the construction of a generative model for birdsongs. This model comprises two Lorenz attractors, where the higher attractor delivers two control parameters (*grey circles*) to a lower level attractor, which, in turn, delivers two control parameters to a synthetic syrinx to produce amplitude and frequency modulated stimuli. This stimulus is represented as a sonogram in the right panel. The equations represent the hierarchical dynamic model in the form of Eq. 2

evolve more slowly, they switch the lower attractor on and off, generating distinct songs, where each song comprises a series of distinct chirps (see Fig. 3).

6.3 Song Recognition

This model generates spontaneous sequences of songs using autonomous dynamics. We generated a single song, corresponding roughly to a cycle of the higher attractor and then inverted the ensuing sonogram (summarized as peak amplitude and volume) using the message-passing scheme described in previous sections. The results are shown in Fig. 3 and demonstrate that, after several hundred milliseconds, the veridical hidden states and superordinate causes can be recovered. Interestingly, the third chirp is not perceived, in that the first-level prediction error was not sufficient to overcome the dynamical and structural priors entailed by the model. However, once the subsequent chirp had been predicted correctly the following sequence of chirps was recognized with a high degree of conditional confidence. Note that when the second and third chirps in the sequence are not recognized, first-level prediction error is high and the conditional confidence about the causes at the second level is low (reflected in the wide 90 % confidence intervals). Heuristically, this means that the synthetic bird listening to

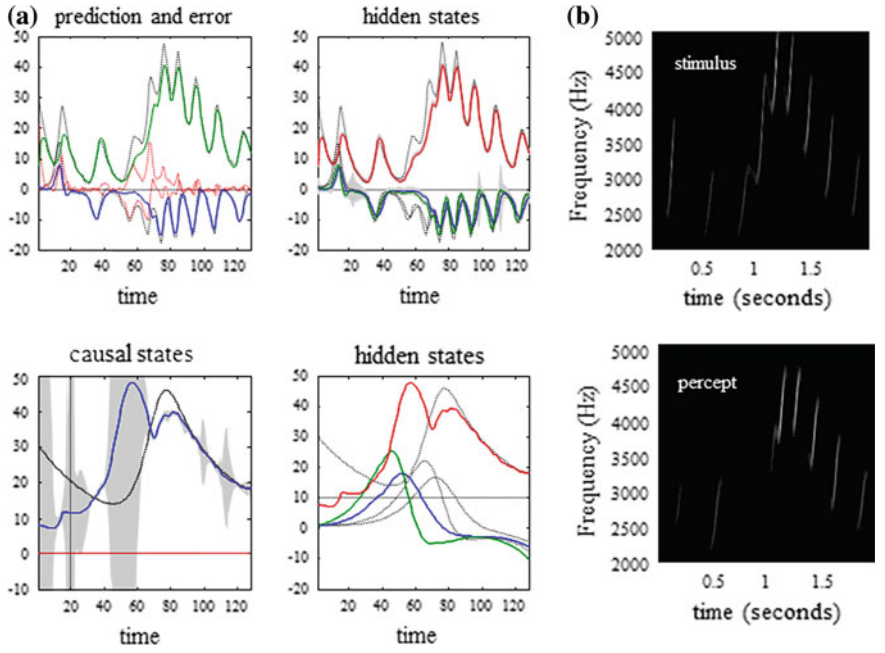


Fig. 3 Results of a Bayesian inversion or deconvolution of the sonogram shown in Fig. 2. **a** Upper panels show the time courses of hidden and causal states. (*Upper left*) These are the true and predicted states driving the syrinx and are simple mappings from two of the three hidden states of the first-level attractor. The solid lines respond to the conditional mode and the dotted lines to the true values. The discrepancy is the prediction error and is shown as a broken red line. (*Upper right*) The true and estimated hidden states of the first-level attractor. Note that the third hidden state has to be inferred from the sensory data. Confidence intervals on the conditional expectations are shown in grey and demonstrate a high degree of confidence, because a low level of sensory noise was used in these simulations. The panels below show the corresponding causes and hidden states at the second level. Again the conditional expectations are shown as solid lines and the true values as broken lines. Note the inflated conditional confidence interval halfway through the song when the third and fourth chirps are misperceived. **b** The stimulus and percept in sonogram format, detailing the expression of different frequencies generated over peristimulus time

the song did not know which song was being emitted and was unable to predict subsequent chirps.

6.4 Structural and Dynamic Priors

This example provides a nice opportunity to illustrate the relative roles of structural and dynamic priors. Structural priors are provided by the top-down inputs that dynamically reshape the manifold of the low-level attractor. However, this

attractor itself contains an abundance of dynamical priors that unfold in generalized coordinates. Both provide important constraints on the evolution of sensory states, which facilitate recognition. We can selectively destroy these priors by lesioning the top-down connections to remove structural priors (over hidden causes) or by cutting the intrinsic connections that mediate dynamic priors (over hidden states). The latter involves cutting the self-connections in Fig. 1 among the causal and state units. The results of these two simulated lesion experiments are shown in Fig. 4. The top panel shows the percept as in the previous panel, in terms of the predicted sonogram and prediction error at the first and second level. The subsequent two panels show exactly the same information but without structural (middle) and dynamic (lower) priors. In both cases, the bird fails to recognize the sequence with a corresponding inflation of prediction error, particularly at the last level. Interestingly, the removal of structural priors has a less marked effect on recognition than removing the dynamical priors. Without dynamical priors there is a failure to segment the sensory stream and although there is a preservation of frequency tracking, the dynamics *per se* have completely lost their sequential structure. Although it is interesting to compare structural and dynamics priors, the important message here is that both are necessary for veridical perception and that removal of either leads to suboptimal inference. Both of these empirical priors prescribe dynamics that enable the synthetic bird to predict what will be heard next. This leads to the question 'What would happen if the song terminated prematurely?'

6.5 Omission and Violation of Predictions

We repeated the above simulation but terminated the song after the fifth chirp. The corresponding sonograms and percepts are shown with their prediction errors in Fig. 5. The left panels show the stimulus and percept as in Fig. 4, while the right panels show the stimulus and responses to omission of the last syllables. These results illustrate two important phenomena. First, there is a vigorous expression of prediction error after the song terminates abruptly. This reflects the dynamical nature of the recognition process because, at this point, there is no sensory input to predict. In other words, the prediction error is generated entirely by the predictions afforded by the dynamic model of sensory input. It can be seen that this prediction error (with a percept but no stimulus) is almost as large as the prediction error associated with the third and fourth stimuli that are not perceived (stimulus but no percept). Second, it can be seen that there is a transient percept, when the omitted chirp should have occurred. Its frequency is slightly too low, but its timing is preserved in relation to the expected stimulus train. This is an interesting stimulation from the point of view of ERP studies of omission-related responses. These simulations and related empirical studies (e.g., Nordby et al. 1994; Yabe et al. 1997) provide clear evidence for the predictive capacity of the brain. In the context

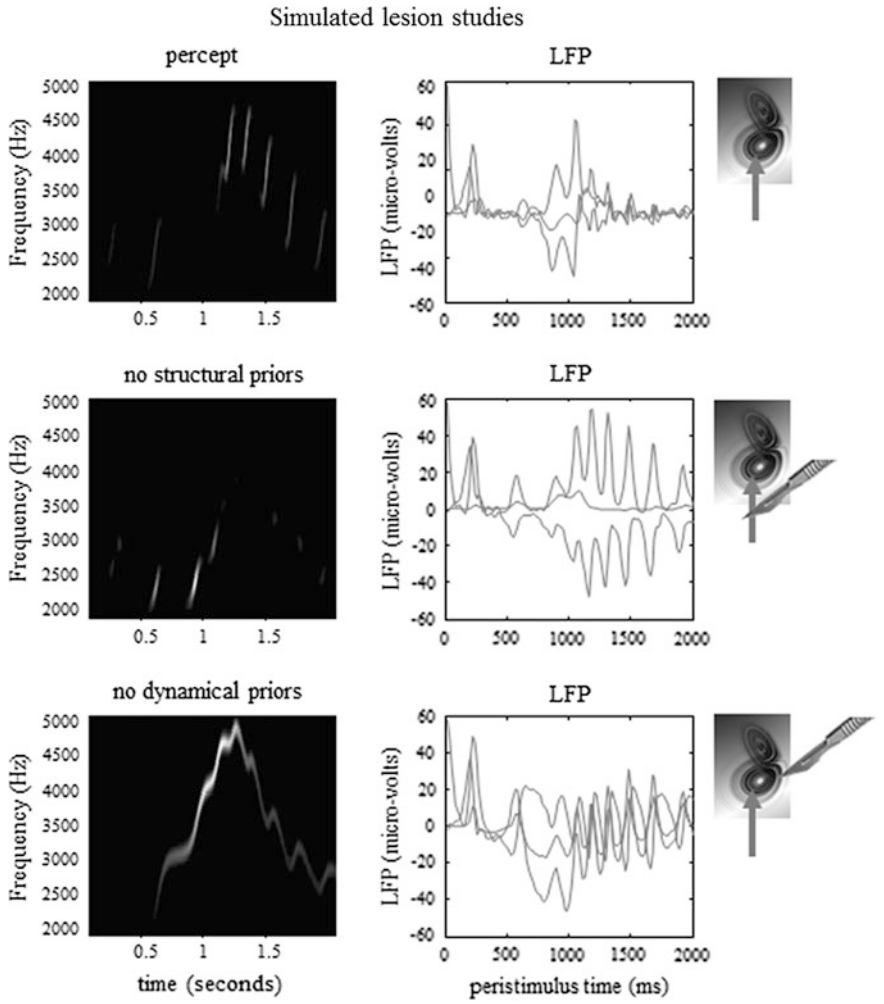
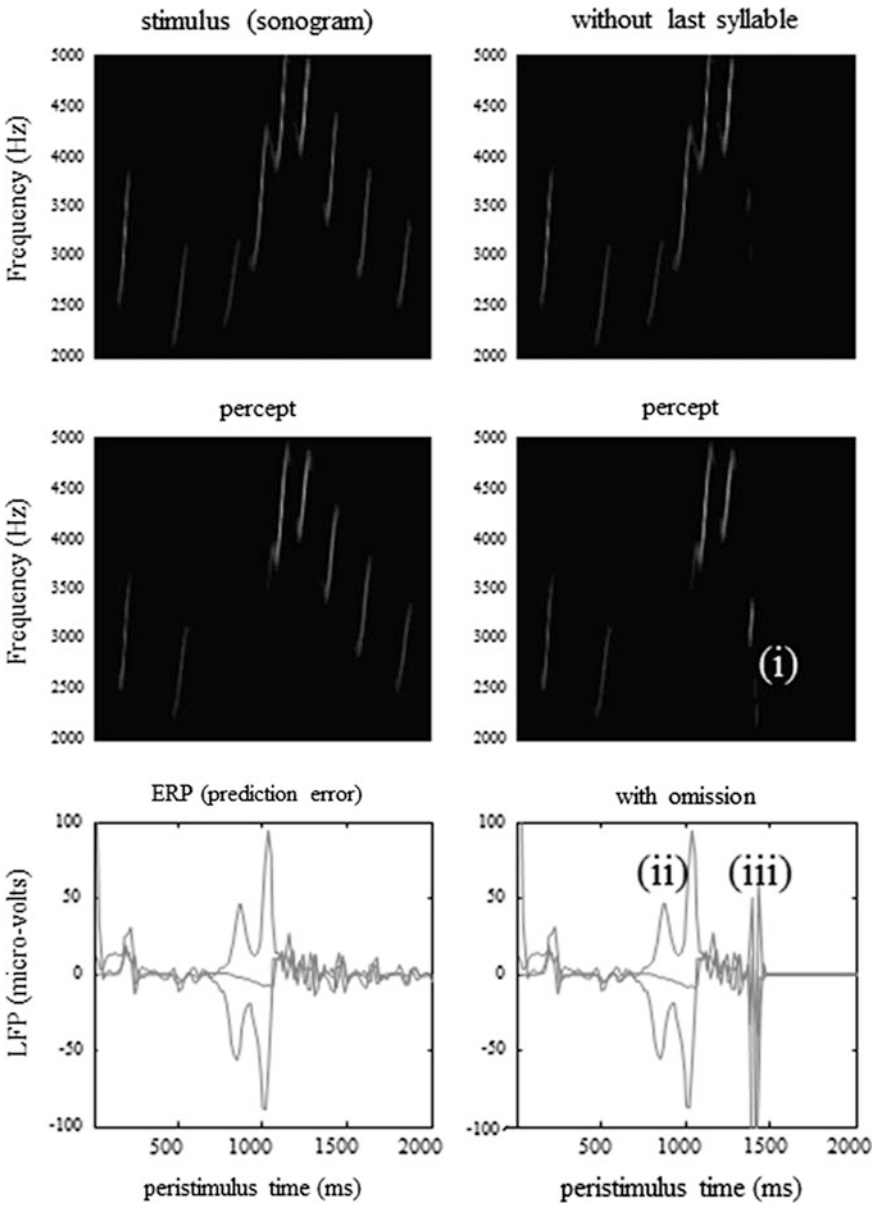


Fig. 4 Results of simulated lesion studies using the birdsong model of the previous figures. The left panels show the percept in terms of the predicted sonograms, and the right panels show the corresponding prediction error (at both levels); these are the differences between the incoming sensory information and the prediction and the discrepancy between the conditional expectation of the second-level cause and that predicted by the second-level hidden states. *Top panels* The recognition dynamics in the intact bird. *Middle panels* The percept and corresponding prediction errors when the connections between the hidden states at the second level and their corresponding causes are removed. This effectively removes structural priors on the evolution of the attractor manifold prescribing the sensory dynamics at the first level. *Lower panels* The effects of retaining the structural priors but removing the dynamical priors by cutting the connections that mediate inversion in generalized coordinates. These results suggest that both structural and dynamical priors are necessary for veridical perception

of music studies, the results in Fig. 5 can be seen as a rough model of the violation (harmonic) responses described in Sect. 2 (Koelsch et al. 2000, 2002, 2008, Loui et al. 2005; Besson and Faita 1995; Verleger 1990). In this example, prediction rests upon the internal construction of an attractor manifold that defines a family of trajectories, each corresponding to the realization of a particular song. In the last simulation we look more closely at perceptual categorization of these songs.

6.6 Perceptual Categorization

In the previous simulations, we saw that a song corresponds to a sequence of chirps that is preordained by the shape of an attractor manifold controlled by top-down inputs. This means that for every point in the state-space of the higher attractor there is a corresponding manifold or category of song. In other words, recognizing or categorizing a particular song corresponds to finding a fixed location in the higher state-space. This provides a nice metaphor for perceptual categorization; because the neuronal states of the higher attractor represent, implicitly, a category of song. Inverting the generative model means that, probabilistically, we can map from a sequence of sensory events to a point in some perceptual space, where this mapping corresponds to perceptual recognition or categorization. This can be demonstrated in our synthetic songbird by ignoring the dynamics of the second-level attractor and exposing the bird to a song and letting the states at the second level optimize their location in perceptual space, to best predict the sensory input. To illustrate this, we generated three songs by fixing the Rayleigh and Prandtl variables to three distinct values. We then placed uninformative priors on the second-level causes (that were previously driven by the hidden states of the second-level attractor) and inverted the model in the usual way. Figure 6a shows the results of this simulation for a single song. This song comprises a series of relatively low-frequency chirps emitted every 250 ms or so. The causes of this song (song C in panel b) are recovered after the second chirp, with relatively tight confidence intervals (the blue and green lines in the lower left panel). We then repeated this exercise for three songs. The results are shown in Fig. 6b. The songs are portrayed in sonogram format in the top panels and the inferred perceptual causes in the bottom panels. The left panel shows the evolution of these causes for all three songs as a function of peristimulus time and the right shows the corresponding conditional density in the causal or perceptual space of these two states after convergence. It can be seen that for all three songs the 90 % confidence interval encompasses the true values (red dots). Furthermore, there is very little overlap between the conditional densities (grey regions), which means that the precision of the perceptual categorization is almost 100 %. This is a simple but nice example of perceptual categorization, where sequences of sensory events


stimulus (sonogram)
without last syllable
Frequency (Hz)
5000
4500
4000
3500
3000
2500
2000
percept
percept
(i)
ERP (prediction error)
with omission
LFP (micro-volts)
100
50
0
-50
-100
(ii)
(iii)
500
1000
1500
2000
peristimulus time (ms)
peristimulus time (ms)

◀ **Fig. 5** Omission-related responses. Here, we have omitted the last few chirps from the stimulus. The left-hand panels show the original sequence and responses evoked. The right-hand panels show the equivalent dynamics on omission of the last chirps. The top panels show the stimulus and the middle panels the corresponding percept in sonogram format. The interesting thing to note here is the occurrence of an anomalous percept after termination of the song on the lower right (*i*). This corresponds roughly to the chirp that would have been perceived in the absence of omission. The lower panels show the corresponding (precision-weighted) prediction error under the two stimuli at both levels. A comparison of the two reveals a burst of prediction error when a stimulus is missed (*ii*) and at the point that the stimulus terminates (*iii*) despite the fact that there is no stimulus present at this time. The darker lines correspond to prediction error at the first level, and the lighter lines correspond to prediction error at the second level

with extended temporal support can be mapped to locations in perceptual space, through Bayesian filtering (predictive coding) of the sort entailed by the free-energy principle.

7 Conclusion

This chapter has suggested that the architecture of cortical systems speaks to hierarchical generative models in the brain. The estimation or inversion of these models corresponds to a generalized Bayesian filtering (predictive coding) of sensory inputs to disclose their causes. This predictive coding can be implemented in a neuronally plausible fashion, where neuronal dynamics self-organize when exposed to inputs to suppress prediction errors. The focus of this chapter has been on the nature of the hierarchical models and, in particular, models that show autonomous dynamics. These models may be relevant for music perception because they enable sequences of sequences to be inferred or recognized. We have tried to demonstrate their plausibility, in relation to empirical observations, by interpreting the prediction error, associated with model inversion, with observed electrophysiological responses. These models provide a graceful way to map from complicated sensory trajectories to points in abstract perceptual spaces. Furthermore, in a hierarchical setting, this mapping may involve trajectories in perceptual spaces of increasingly higher order. The mathematical formalism (and simulations) of hierarchical Bayesian inference in the brain provides a nice link between the generic principles of perceptual inference (and self-organisation) and the perception of music—in particular, it enabled us to simulate one of the most prominent electrophysiological phenomena in music research; namely violation—dependent responses.

The ideas presented in this chapter have a long history, starting with the notion of neuronal energy (Helmholtz 1860), covering ideas like efficient coding and analysis by synthesis (Barlow 1961; Neisser 1967) to more recent formulations in terms of Bayesian inversion and predictive coding (e.g., Ballard et al.1983; Dayan et al. 1995; Kawato et al. 1993; Mumford 1992; Rao and Ballard 1998). This work

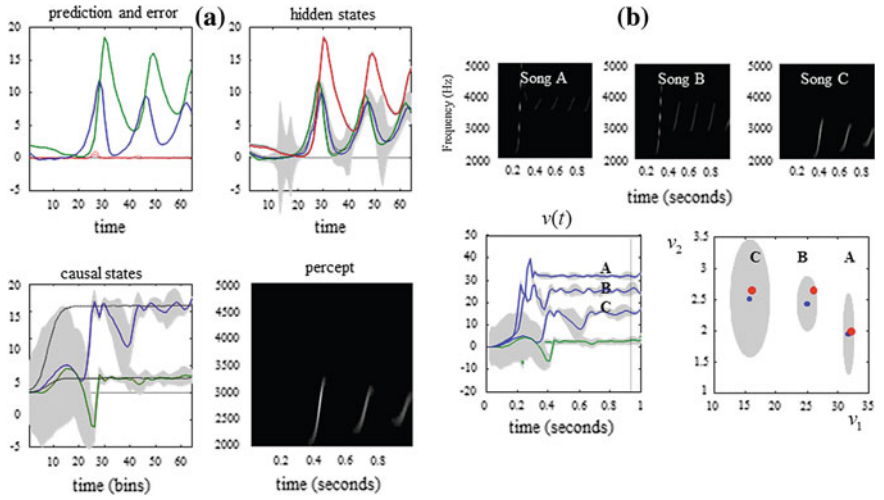


Fig. 6 **a** Schematic demonstration of perceptual categorization. This figure follows the same format as Fig. 3. However, here there are no hidden states at the second level, and the causes were subject to stationary and uninformative priors. This song was generated by a first-level attractor with fixed control parameters of 16 and 8/3, respectively. It can be seen that, on inversion of this model, these two control variables, corresponding to causes or states at the second level, are recovered with relatively high conditional precision. However, it takes about 50 iterations (about 600 ms) before they stabilize. In other words, the sensory sequence has been mapped correctly to a point in perceptual space after the occurrence of the second chirp. This song corresponds to song C on the right. **b** The results of inversion for three songs each produced with three distinct pairs of values for the second-level causes (the Rayleigh and Prandtl variables of the first-level attractor). *Upper panel* The three songs shown in sonogram format corresponding to a series of relatively high-frequency chirps that fall progressively in both frequency and number as the Rayleigh number is decreased. *Lower left* These are the second-level causes shown as a function of peristimulus time for the three songs. It can be seen that the causes are identified after about 600 ms with high conditional precision. *Lower right* This shows the conditional density on the causes shortly before the end of peristimulus time (*dotted line on the left*). The blue dots correspond to conditional means or expectations, and the grey areas correspond to the conditional confidence regions. Note that these encompass the true values (*red dots*) used to generate the songs. These results indicate that there has been a successful categorization, in the sense that there is no ambiguity (from the point of view of the synthetic bird) about which song was heard

has tried to provide support for the notion that the brain uses attractors to represent and predict causes in the sensorium (Byrne et al. 2007; Deco and Rolls 2003; Freeman 1987; Tsodyks 1999). More generally, one might conclude that we need large brains, with deep hierarchical structure, to perceive and appreciate the world we inhabit. This resonates with Einstein's conclusion:

> He who joyfully marches to music in rank and file has already earned my contempt. He has been given a large brain by mistake, since for him the spinal cord would suffice. Albert Einstein

although motivated from a slightly different perspective.

Acknowledgments The Wellcome Trust funded this work. We would also like to thank Larry Goodyer for invaluable discussions.

References

Angelucci, A., Levitt, J. B., Walton, E. J., Hupe, J. M., Bullier, J., & Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience, 22*, 8633–8646.

Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature, 306*, 21–26.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Besson, M., & Faita, F. (1995). Event-related potential (ERP) study of musical expectancy—comparison of musicians with nonmusicians. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1278–1296.

Botvinick, M. M. (2007). Multilevel structure in behaviour and in the brain: A model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1485), 1615–1626.

Breakspear, M., & Stam, C. J. (2005). Dynamics of a neural system with a multiscale architecture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*, 1051–1107.

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychology Review, 114*(2), 340–375.

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., et al. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science, 313*, 1626–1628.

Chait, M., Poeppel, D., de Cheveigné, A., & Simon, J. Z. (2007). Processing asymmetry of transitions between order and disorder in human auditory cortex. *Journal of Neuroscience, 27*(19), 5207–5514.

Dayan, P., Hinton, G. E., & Neal, R. M. (1995). The Helmholtz machine. *Neural Computation, 7*, 889–904.

Deco, G., & Rolls, E. T. (2003). Attention and working memory: A dynamical model of neuronal activity in the prefrontal cortex. *European Journal of Neuroscience, 18*(8), 2374–2390.

DeFelipe, J., Alonso-Nanclares, L., & Arellano, J. I. (2002). Microstructure of the neocortex: Comparative aspects. *Journal of Neurocytology, 31*, 299–316.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association, 68*, 117–130.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*, 1–47.

Feynman, R. P. (1972). *Statistical mechanics*. Reading, MA: Benjamin.

Freeman, W. J. (1987). Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biological Cybernetics, 56*(2–3), 139–150.

Friston, K. J . (1997). Transients, metastability, and neuronal dynamics. *NeuroImage, 5*(2), 164–171.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*, 815–836.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris, 100*(1–3), 70–87.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), e1000211.

Friston, K., & Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks, 22*(8), 1093–1104.

Haken, H., Kelso, J. A. S., Fuchs, A., & Pandya, A. S. (1990). Dynamic pattern-recognition of coordinated biological motion. *Neural Networks, 3*, 395–401.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience, 28*, 2539–2550.

Helmholtz, H. (1860/1962). Handbuch der physiologischen Optik. In J. P. C. Southall (Ed.) (Vol. 3). New York: Dover, (Translation).

Helmholtz, H. (1866/1962). Concerning the perceptions in general. In J. Southall (Ed.) *Treatise on physiological optics* (3rd ed., Vol. III). New York: Dover, (Translation).

Helmholtz, H. (1877). On the sensations of tone as a physiological basis for the theory of music. In A. J. Ellis (Ed.), *Fourth German edition, translated, revised, corrected with notes and additional appendix*. New York: Dover Publications Inc., 1954 (Reprint).

Hinton, G. E., & von Camp, D. (1993). Keeping neural networks simple by minimising the description length of weights. *In Proceedings of COLT-93*, 5–13.

Hupe, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature, 394*, 784–787.

Jirsa, V. K., Fuchs, A., & Kelso, J. A. (1998). Connecting cortical and behavioral dynamics: bimanual coordination. *Neural Computation, 10*, 2019–2045.

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association, 407*, 717–726.

Kawato, M., Hayakawa, H., & Inui, T. (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network, 4*, 415–422.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4*(11), e1000209.

Koelsch, S., Gunter, T., Friederici, A. D., & Schröger, E. (2000). Brain indices of music processing: "nonmusicians" are musical. *Journal of Cognitive Neuroscience, 12*(3), 520–541.

Koelsch, S., Schroger, E., & Gunter, T. C. (2002). Music matters: Preattentive musicality of the human brain. *Psychophysiology, 39*(1), 38–48.

Koelsch, S., Fritz, T., Schulze, K., Alsop, D., & Schlaug, G. (2005). Adults and children processing music: An fMRI study. *Neuroimage, 25*(4), 1068–1076.

Koelsch, S., Kilches, S., Steinbeis, N., & Schelinski, S. (2008). Effects of unexpected chords and of performer's expression on brain responses and electrodermal activity. *PloS ONE, 3*, e2631.

Kopell, N., Ermentrout, G. B., Whittington, M. A., & Traub, R. D. (2000). Gamma rhythms and beta rhythms have different synchronization properties. *Proceedings of the National Academy of Sciences USA, 97*, 1867–1872.

Laje, R., Gardner, T. J., & Mindlin, G. B. (2002). Neuromuscular control of vocalizations in birdsong: a model. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 65, 051921.1–8.

Laje, R., & Mindlin, G. B. (2002). Diversity within a birdsong. *Physical Review Letters, 89*, 288102.

Levitin, D. J., & Menon, V. (2003). Musical structure is processed in "language" areas of the brain: a possible role for Brodmann Area 47 in temporal coherence. *Neuroimage, 20*(4), 2142–2152.

Loui, P., Grent-'t-Jong, T., Torpey, D., & Woldorff, M. (2005). Effects of attention on the neural processing of harmonic syntax in Western music. *Brain Research. Cognitive Brain Research, 25*(3), 678–687.

MacKay, D. J. C. (1995). Free-energy minimisation algorithm for decoding and cryptoanalysis. *Electronics Letters, 31*, 445–447.

Maunsell, J. H., & van Essen, D. C. (1983). The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience, 3*, 2563–2586.

McCrea, D. A., & Rybak, I. A. (2008). Organization of mammalian locomotor rhythm and pattern generation. *Brain Research Reviews, 57*(1), 134–146.
Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121*, 1013–1052.
Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.
Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.
Murphy, P. C., & Sillito, A. M. (1987). Corticofugal feedback influences the generation of length tuning in the visual pathway. *Nature, 329*, 727–729.
Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental sparse and other variants. In M. I. Jordan (Ed.), *Learning in graphical models, 355–368*. Dordrecht: Dordrecht Kulver Academic Press.
Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
Nordby, H., Hammerborg, D., Roth, W. T., & Hugdahl, K. (1994). ERPs for infrequent omissions and inclusions of stimulus elements. *Psychophysiology, 31*(6), 544–552.
Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *Neuroimage, 50*(1), 302–313.
Rabinovich, M., Huerta, R., & Laurent, G. (2008). Neuroscience: Transient dynamics for neural processing. *Science, 321*(5885), 48–50.
Rao, R. P., & Ballard, D. H. (1998). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience, 2*, 79–87.
Rockland, K. S., & Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research, 179*, 3–20.
Rohrmeier, M. A., & Koelsch, S. (2012). Predictive information processing in music cognition. *A critical review. Int J Psychophysiol., 83*(2), 164–175.
Rosier, A. M., Arckens, L., Orban, G. A., & Vandesande, F. (1993). Laminar distribution of NMDA receptors in cat and monkey visual cortex visualized by [3H]-MK-801 binding. *Journal of Comparative Neurology, 335*, 369–380.
Sherman, S. M., & Guillery, R. W. (1998). On the actions that one nerve cell can have on another: Distinguishing "drivers" from "modulators". *Proceedings of the National Academy of Sciences USA, 95*, 7121–7126.
Sloboda, J. A. (1991). Music structure and emotional response: Some empirical fmdings. *Psychology of Music, 19*, 110–120.
Steinbeis, N., Koelsch, S., & Sloboda, J. A. (2006). The role of harmonic expectancy violations in musical emotions: evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience, 18*(8), 1380–1393.
Tsodyks, M. (1999). Attractor neural network models of spatial maps in hippocampus. *Hippocampus, 9*(4), 481–489.
Verleger, R. (1990). P3-evoking wrong notes: unexpected, awaited, or arousing? *International Journal of Neuroscience, 55*, 171–179.
Yabe, H., Tervaniemi, M., Reinikainen, K., & Näätänen, R. (1997). Temporal window of integration revealed by MMN to sound omission. *NeuroReport, 8*(8), 1971–1974.
Zeki, S., & Shipp, S. (1988). The functional logic of cortical connections. *Nature, 335*, 311–331.

Change and Continuity in Sound Analysis: A Review of Concepts in Regard to Musical Acoustics, Music Perception, and Transcription

Albrecht Schneider

1 On measuring Basic Properties of Sound: A Brief Retrospect

Over the past decades, a broad range of software and hardware tools has become available suited to perform sound analysis both in the time domain and in the frequency domain. Though musical acoustics as a field of research based on both calculation and experiment started to evolve around 1600 (cf. Cohen 1984), and had gained impetus by 1700 with experiments on, for example, vibration of strings as conducted by J. Sauveur and many other scholars since (see Cannon and Dostrovsky 1982), investigation of actual sound as produced by instruments and voices was limited since appropriate tools for measurement and analysis were scarce. Experiments at the time of Chladni (1805) were mostly done with sets of tuning forks or organ pipes (see Beyer 1999). After Charles Cagniard de la Tour had invented the siren (1819), such instruments came into use soon (as in the pioneering experiments of August Seebeck that led to the first formulation of 'periodicity pitch' published in 1841; see Hesse 1972, 58ff.; de Boer 1976; Schneider 1997b, 134f.). Tuning forks, pipes, sirens as well as resonance boxes and resonance bottles such as described by Helmholtz (1863/1870/1896) were the basic toolkit of the acoustician then. Even Stumpf conducted most of his experiments on perception of consonant and dissonant sounds (1890, 1898) with the aid of sets of tuning forks (mounted on resonance boxes), reed pipes and organ mixture stops. One proven method used to study the structure of harmonic partials in complex tones was resonance, another was additive synthesis of sounds such as

This article is dedicated to the memory of Walter Graf (1903–1982), one of the pioneers of sound research in Systematic and Comparative Musicology.

A. Schneider (✉)
Institut für Musikwissenschaft, Universität Hamburg,
Neue Rabenstr. 13, D-20354 Hamburg, Germany
e-mail: aschneid@uni-hamburg.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_3,

vowels by means of a set of tuning forks that could be excited in an electromagnetic circuit, and thereby would produce a continuous sound. Mechanical devices such as the *Tonmesser* built by Appun (of Hanau, Germany) and the *Tonvariator* developed by Stern (1902) even allowed measurement and production of tones varying continuously in frequency as well as synthesis of sonorities ranging from perfect harmonic to inharmonic. Besides resonance, interference was a principle used for analysis. By the end of the 19th century, an apparatus for dampening or cancelling partials out of complex sounds (such as vowels) had been developed (cf. Graf 1980, 212). A range of mechano-acoustical devices available at the time were used by Helmholtz and other scholars (e.g., Carl Stumpf and co-workers, Wilhelm Wundt and co-workers) to explore sound as well as auditory phenomena such as the sensation of consonance and dissonance, roughness, beats, and combination tones.

For the study of sound wave characteristics, Helmholtz (1870, 33f.) also saw the need to record sound as radiated from musical instruments (including the human voice) on some suited graph paper or other material. He rightly emphasized that the shape of a sound wave per period determines the corresponding timbre (German: *Klangfarbe*; see below, Sect. 3). Helmholtz pointed to the *Phonautograph* of Scott de Martinville who, around 1857–1860, had actually recorded sound waves (including a few seconds of a French folk song, *Au clair de la lune*) on a rotating cylinder.[1] In 1877, Edison presented his *Phonograph* (the early tinfoil version), which was used in the 1880s and 1890s for the study of sounds from musical instrument and the human voice by the German physiologist Georg Meissner (see below). With Edison's improved 1888 model of the *Phonograph* (Edison 1888), recording and reproduction of sound had become a standard method. In particular the physiologist Ludimar Hermann published numerous articles on speech sounds (vowels, consonants; see below, Sect. 3) recorded and analyzed by way of *Phonophotographie* (e.g., Hermann 1889, 1893, 1894, 1895, 1911). As an alternative to the *Phonograph*, the physiologist Victor Hensen had developed a machine called *Sprachzeichner* (1879, 1888) that offered a very subtle registration of sound waves (plus the recording of a tuning fork as referent). The *Sprachzeichner* was one of the fundamental tools in phonetics that has been employed, for instance, by the Finish-Swedish linguist, Hugo Pipping, for the study of spoken and sung vowels (e.g., Pipping 1894). The study of sung vowels of course played a central role in the theory of formants (see below, Sect. 3.3). Meissner, Hermann, Pipping and others took great pains in analyzing complex sounds for finding the period length and the fundamental frequency as well as to determine the spectral content for each single period. Fourier analysis was done by

[1] It is reported (cf. Beyer 1999, 140f., Figures 6,7,8,9,10,11) that Karl Rudolph Koenig (who invented a range of instruments for acoustics) had improved the *Phonautograph* by adding a conical horn to collect the sound (as was similarly done later by Edison with the *Phonograph*) as well as the rotating cylinder on which the sound was actually recorded as an oscillogram. As to the early history of sound analysis, see also Graf (1980, 211ff).

hand with the aid of rulers and curve templates in a time-consuming geometrical analysis procedure.[2]

2 Frequency Measurement, Periodicity Estimation, Melographic 'Pitch' Notation

Besides Meissner, Hermann, and Pipping, another scholar with a medical training, Edward W. Scripture, became a specialist in the study of speech curves for which he developed a methodology to determine actual (fundamental) frequencies of speech or sung melodies (Scripture 1906, 1927). This approach still was based on a manual reading of wave-lengths (Scripture 1906, 60ff.) or periods that had to be transferred to their respective frequency values. A similar yet improved technique again called *Phonophotography* was employed by Metfessel (1928) who had the sound wave (plus a referent vibration of 100 Hz obtained from a tuning fork usable as a time-line) recorded on film. It was applied to the study of vibrato and intonation in "western" music as well as to the microstructure of African-American singing styles.

The basic idea behind the analyses performed by Hermann, Scripture, Metfessel as well as by other researchers is to use the information provided by the temporal structure of a sound wave in order to determine the frequency (Hz) corresponding to a given period length (ms). The fundamental frequency can be calculated by making use of the inverse relation between period T(ms) and frequency f(Hz), which is given by $T = 1/f$, and $f = 1/T$. In case the signal is a complex tone comprising a series of harmonic partials, the period typically is determined by the first partial acting as the fundamental frequency f_1, which is acoustically real if it can be measured as a spectral component (for example, in the sound of a flute in normal blowing condition). In the case of a harmonic complex tone, partial frequencies can be determined according to $f_n = n \times f_1$. In a strict sense, the frequency corresponding to any particular period of a sine tone or to f_1 of a complex tone (with all partials locked in zero-phase) can be taken as its instantaneous frequency (German: *Augenblicksfrequenz*).[3] An additional aspect is that, in a signal where f_1 is weak or even missing, the period length T(ms) determined by a sufficient number of harmonic partials superimposed on each other and locked in phase is identical with the period length τ(ms) defined by the f_1 partial (for

[2] Some of the early research is reported (with details in regard to methods of measurement and the technology then available) in Meissner (1907), Krueger (1907), Hermann (1893), Herrmann (1908). See also Panconcelli-Calzia (1941/1994).

[3] This is a practical definition applicable to empirical observation and measurement. In signal processing, the notion 'instantaneous frequency' has a more technical definition based on calculating the instantaneous (or local) phase $\varphi(t)$ (a real-valued function) for a (complex-valued) function $x(t)$ representing the signal; the instantaneous angular frequency can be determined as the derivative of the phase, that is $\omega(t) = \varphi'(t)$; see Cohen (1995, 39ff).

examples, see Schneider 1997a, b and 2000). What is measured, in this case, is the repetition frequency of the envelope of a complex waveshape. This repetition frequency can be labeled f0 of F0, and in contemporary psychoacoustics is generally addressed as the "fundamental frequency" of a complex sound (what is rather misleading since it implies the concept of a complete harmonic spectrum where f_1 is present). The envelope repetition frequency f0 of course permits to calculate also f_1 (whether this is present in the spectrum or not; if it is, f0[Hz] = f_1[Hz]). The sensation corresponding to f0 is a 'periodicity pitch' that proved of importance to explain pitch perception in the case of a 'missing fundamental', that is, where the sound carries little or no energy at f_1 (cf. de Boer 1976). The sensation of a 'periodicity pitch' evoked from detecting f0 at times has also been labeled 'low pitch' since it is typically found below the spectral components actually present. For example, in major and minor chords in "root" position comprising a triad of three complex tones each synthesized from a number of partials in just intonation, the "fundamental" f0 that can be calculated by either autocorrelation or subharmonic matching appears as the common denominator of the total harmonic series (cf. Schneider and Frieler 2009; Schneider 2011).

In practice, finding the frequencies for musical tones from reading periods manually (as was done by Scripture et al.) is an arduous task since the sound segments under investigation must be small to allow manual search for appropriate zero crossings defining periods. By about 1930, measurements had been improved in that electronic equipment including amplifiers and oscilloscopes (based on either galvanometer or cathode ray technology) had become available that allowed for continuous recording of oscillograms up to ca. 15 s on film or special paper strips. Such a setup has been used, for example, to study intonation patterns of cellists playing scales and melodic phrases in a microtonal context (Kreichgauer 1932). Still the length of periods for each tone played by skilled musicians had to be determined manually (i.e., by counting the number of periods per time unit); the amount of labour Kreichgauer (1932) has put into his study of intonation patterns seems incredible since a corpus of several hundred meters of oscillographic registrations resulting from a series of experiments had to be analyzed.

The advantage of taking the sound wave as a prime source for analysis is that, if we assume linear behaviour of all tools used in the recording, it can be assumed that no alterations in the signal have yet occurred before analysis. Hence, looking for the period lengths of a time signal and transferring the temporal information to corresponding frequency values can be regarded as an objective way of measurement. Such an analysis allows, most of all, to determine the fundamental frequency per period, and to see if, and possibly to which extent, there is a fluctuation in period length (meaning that there is some frequency modulation in the signal as well). To illustrate the case, we might look into a very short segment of the sound wave recorded as part of a song, *Abu Zeluf*, as sung by a woman in Lebanon (Fig. 1).[4]

[4] *Abu Zeluf*, sung by Dunya Yunis, recorded by Poul Rovsing Olsen on 4th of February, Beirut 1972. Published on the LP *Music in the World of Islam, Vol. I: The Human Voice*, Tangent

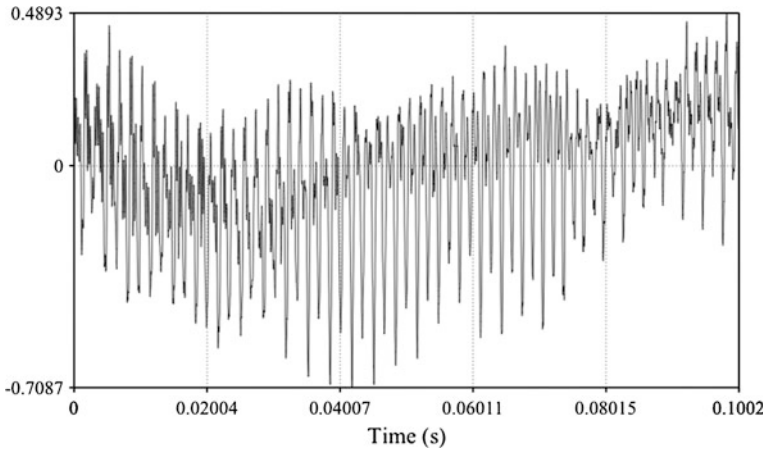


Fig. 1 Short segment (100 ms) of sound recorded from a female singer in Lebanon

The period length at the beginning of this segment is 3.426 ms corresponding to 291.91 Hz while at the end a period length of 3.025 ms is found corresponding to 330.54 Hz. Hence, there is a shift in fundamental frequency that can be viewed also as a shift in 'pitch' (in regard to hearing and psychoacoustics; see below) by more than one semitone. With modern digital equipment, measuring the period lengths and even calculating the 'pitch shift' can be done with great precision. In the days of analogue measurement, tracking of zero crossings that define the periods in complex wave shapes mostly was done with an oscilloscope, and often the signal was put through a low pass or band pass filter in order to simplify the wave shape, and possibly to extract the fundamental from each complex tone (as this was believed to represent the 'pitch') in a melody. Already in the 1930s, techniques for finding the fundamental frequency of a sequence of tones forming a (quasi-continuous) melodic contour and plotting such a curve over time by means of a so-called *Tonhöhenschreiber* or *pitch recorder* suited for monophonic signals had been developed (Grützmacher and Lottermoser 1937; Obata and Kobayashi 1937). The range of the German *Tonhöhenschreiber* has been given as 2.5 octaves, and the precision of 'pitch' as ca. 25 cent (Lottermoser 1976/1977, 139). By the end of the 1940s, ideas of Charles Seeger, one of the pioneers in systematic and comparative musicology, for constructing a 'melograph' capable to plot the ups and downs of a melodic contour as well as the dynamic changes of a signal over time had reached the state of a beta version (cf. Seeger 1951). The aim was to develop an electronic device that could perform *electronic sound-writing in the laboratory* (Seeger 1951), that is, automated transcription of (monophonic) music recorded in the field or available on phonograph records. Melographic analysis was

(Footnote 4 continued)
Records TGS 131 (London 1976), B2. As to the history of the recording and its later (ab)use, see Feld and Kirgegaard (2010).

considered not only to make transcription easier and subtle changes in pitch detectable but also to provide a tool for objective analysis and notation independent of the ear of the individual listener (see also Schneider 1986, 1987).

During the 1950s, more attempts at constructing 'melographs' were made (of which one realized in Oslo was notable; see Dahlback 1958). Seeger's idea for a melograph finally materialized in his 'Model C', which was used as an aid in transcription of, for example, saxophone parts Charlie Parker had played in his rendition of *Parker's Mood* (rec. New York 1948; cf. Owen 1974). Model C besides tracking of 'pitch' offered recording of dynamics as well as a (rather elementary) representation of spectral energy over time. A more advanced concept of a melograph was developed by Miroslav Filip who, instead of low-pass filtering a given signal for extraction of the fundamental (f_1 of a complex tone) took a nonlinear approach to envelope periodicity detection that could handle also signals with a "missing fundamental" and proved to be robust in actual measurement (Filip 1969, 1970). The melograph developed by Filip has been used as an aid for transcription and analysis of orally transmitted music (cf. Elschek 1979, 2006). The main purpose of melographic analysis is to determine the exact fundamental frequency contour or (for harmonic/periodic signals) its equivalent, the 'pitch contour' calculated from f0 values.[5] Also, the onset and duration of tones as they occur in a certain melodic context are of interest. In addition to finding pitch contours at large, one may look into details of intonation and fluctuations of pitch due to glissandi, vibrati, etc. An example to illustrate the case is given in Fig. 2, which shows the 'pitch' trajectory of a segment of ca. 20 s of a female singing a song, *Abu Zeluf*, recorded in the Lebanon.

The segment in question includes three short introductory notes (the musical notes being, roughly, c4 at about 264 Hz [repeated once], and d4 taken a bit sharp at ca. 301 Hz) followed by a very long note e4 (at ca. 333 Hz) which is held almost constant in pitch for more than 9 s. Then, after a short e4 and a sudden jump upwards to a short peak at ca. 380 Hz, a strong melisma follows with a modulation between the notes e4 and d4. Of course, all notes or, rather, note names like c4, d4, e4, etc. must be taken as relative pitches only that have to be interpreted in regard to the Arab modal scale used in this song (apparently, a scale of the *maqam rast* group of modes). The modulation frequency found in the melisma is about 5.5 Hz (Fig. 3):

After the melisma, we find a sequence of short notes beginning on b3 (at ca. 245 Hz), with a transition from c#4 and d4 to e4, followed by a d4, another e4 and then d4 (on which note the phrase ends). What the melogram (Fig. 2) clearly reveals is the small fluctuation of 'pitch' on the long note e4 before the melisma, and the fairly regular modulation applied when rendering the melisma d4 ↔ e4.

[5] The measurements for pitch curves shown in Figs. 2 and 3 were performed with the Praat software (Boersma and Weenink 2011). A special autocorrelation method (Boersma 1993) was chosen for calculation of pitch with a time resolution of 1 ms.

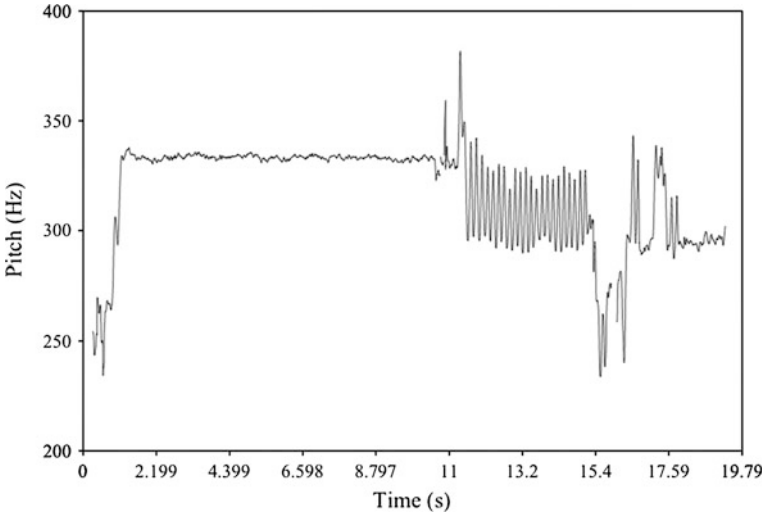


Fig. 2 Melogram of a segment (ca. 20″) of a female singing *Abu Zeluf*, Lebanon

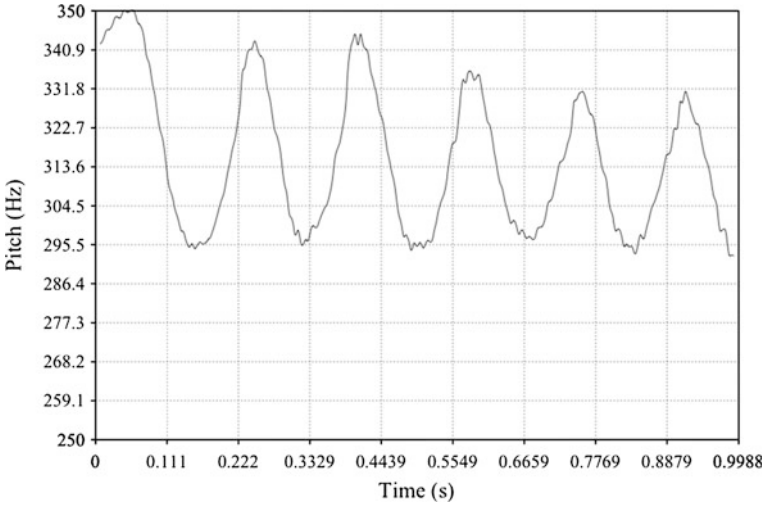


Fig. 3 Dunya Yunis singing *Abu Zeluf*, melisma, frequency modulation

In addition to continuous registration of the fundamental frequency as a function of time, another approach was taken by establishing histograms of fundamental frequency distributions (Tjernlund et al. 1972; Filip 1978). This approach includes statistical considerations and needs computers for practical realization. Recently, such a histogram analysis was chosen (in modern software implementations) for the study of Turkish *makam* music (Bozkurt 2008; Bozkurt et al. 2009;

f0 calculation is done with the well-known *Yin* algorithm of de Cheveigné and Kawahara 2002) and of Cambodian melismatic chanting (of the genre *smot*; Bader 2011). In these recent studies, melographic analysis is combined with an automated, algorithmic analysis of fundamental frequency distributions. To be sure, while melographic analysis can show deviations of 'pitch' from expected values (i.e., frequency values defined by a certain tone system or scale) that occur at a certain time of a performance, and in a specific melodic context, frequency histograms eliminate the time dimension as well as the melodic context in which certain pitch deviations or melodic embellishments (such as vibrato or melisma) take place. The advantage of the histogram technique, however is that the number of occurrences of particular fundamental frequencies leads to a pattern of pitches prevalent in a certain piece of music. In most cases (depending of course on the shape of the frequency distribution and the statistics for frequencies obtained in a particular study) a melodic scale or mode can be inferred from the histogram data.

The melographic approach that, in its original concepts, was directed to finding trajectories of fundamental frequency which then could be plotted as pitch contour curves against time on some graph paper (possibly with a linear or logarithmic frequency [y] and a linear time [x] scale), has been expanded in several directions over the past decades. First, besides normal algorithms suited to achieve either f_1 (= fundamental frequency of complex tones/sounds with n harmonics; f_1 is typically extracted by low pass or band pass filtering) or f0 (= frequency with which the envelope of a complex sound repeats per second), a number of 'tracking algorithms' have become available that allow to calculate trajectories for partials according to some model. One of the best-known such models is that issued by McAulay and Quatieri (1986) that utilizes frame-to-frame peak matching, thereby establishing a number of frequency tracks (depending on the spectral energy distribution and the number and strength of peaks per frame). Instead of covering a full spectrum, tracking algorithms may be directed to the fundamental frequency of complex sounds determined by means of, for example, a constant Q transform (Brown and Puckette 1993; see also Brown 2007). This approach allows precise tracking even of signals changing rapidly in frequency over time (like the example shown in Fig. 1). Other methods applied to speech and music for f0 estimates include autocorrelation plus additional processing steps (de Cheveigné and Kawahara 2002) and Principal Component (PC) autoregressive frequency estimation based on the ModCov (modified covariance, see Marple 1987) model (Hekland 2001). Also, software especially suited for the study of micromelodic ornamentation (such as found in Hindustani music of India) has been developed that uses autocorrelation for an initial pitch estimate but also performs calculation of the spectral centroid; smoothed pitch contours are found by fitting Bézier spline curves to the data (Battey 2004).

Second, while algorithms suited to extract either f_1 or f0 from complex sound for a long time were restricted to deal with monophonic signals only (i.e., signals containing one tone and, typically, one pitch per time), there have been a number of attempts in recent years to cover polyphonic signals as well (for an overview, see Klapuri 2004; Klapuri and Davy 2006, part III, chs. 7 and 8). The goal of

polyphonic analysis often is to provide (if possible) a MIDI file of musically defined (pitch, duration) notes from audio sample data (see, e.g., Paiva et al. 2008). Some of the concepts chosen are related to those of Auditory Scene Analysis (ASA, Bregman 1990) since a number of source signals have to be separated into (spectrally correlated) "streams" or, in this case, "voices". However, though a range of commercial and free software for PCM-to-MIDI-conversion is at hand, one has to bear in mind that the rate of accuracy achieved for polyphonic pitch tracking analysis even with advanced models by now comes close to 60 % (cf. Cañadas Quesada et al. 2010). As it seems, there is still a lot of work ahead for automated transcription of polyphonic music.

3 'Sound Colour' (*Klangfarbe*) and Spectrum: Acoustical and Psychological Aspects

3.1 On the Archaeology of 'Sound Colour' or 'Timbre'

Man (like other mammalian species) has experienced various types of sounds as occurring in the natural environment for thousands of years. Also, communication between humans as well as with animals needed articulated sound. It is very likely that individuals discovered phenomenal differences as well as similarities between various sounds early on, and that categorization of sounds according to certain attributes was undertaken. Though direct evidence is difficult to adduce, there are "ethnographic parallels" that at least can illustrate the case (see, e.g. Feld 1990 with a survey of sound phenomena that are part of the natural environment as well as of the sociocultural life of the Kaluli tribe of Papua New Guinea). Regarding 'Old World' cultures, there is ample evidence for sound phenomena resulting from singing or from various instruments in written sources of Greek and Roman antiquity, and also in medieval writings. For example, the mathematician and music theorist Nikomachos, dealing with sounds produced by various instruments (harm. en. IV), rightly states that sound is an impact on air, and that sounds can be distinguished as to large and small, dull and sharp (low pitches corresponding to dull sound because of a loss of tension in stringed instruments). By the end of the Middle Ages, and more so during 'Renaissance', musical instruments in use had increased by far, and so had the number of different stops or ranks of organ pipes producing different sounds (the variety of instruments is reflected in Praetorius' *De Organographia* of 1619). These facts indicate an awareness of people listening to music for characteristic tonal registers and sound qualities.

As is well-known, theory of vibration and other fundamental issues in acoustics were developed steadily between, roughly, 1500 and 1800 (cf. Cannon and Dostrovsky 1982; Cohen 1984). Resonance in strings and stringed instruments was a major topic of research already when Sauveur (1701) succeeded in the determination of harmonic partials in vibrating strings. However, his lectures, demonstrating that

even a single musical 'tone', when played on, for example, a harpsichord, contained a series of harmonics, were widely recognized, and had great impact also on music theory (as is evident in particular in Rameau's writings, see Schneider 2011). Since vibration theory was based on observations of the swinging pendulum, the notion of a period of vibration (expressed as the duration the pendulum needed to complete a full circle of motion) and its relation to 'pitch' was fairly well understood. The faster the motion, the higher the frequency of vibration and the number of pulses transmitted through the air to the ear would be. And the stronger the excitation force applied to a string, the louder the resulting sound. 'Pitch', then, was dependent on the number of pulses per time unit, and 'loudness' on the amplitude of vibration. Also, one did distinguish 'tone' (the result of a regular vibration) and 'noise' (irregular or even arbitrarily changing motion).

When Chladni entered the stage with his comprehensive work *Die Akustik* (1805/1830), he rightly argued that the source of audible sound (German: *Schall*) is an elastic body set to regular vibration (*Klang*) or irregular motion (*Geräusch*). Criticising Rameau and others who had claimed one could hear a number of harmonic partials in addition to the fundamental (Chladni 1805, 3, in this respect, speaks of *Hauptschwingung*), Chladni argued that a given sound is not an agglomerate or complex but something that is single (*etwas ganz Einfaches*). Notwithstanding his many insights into types of vibration (transversal, longitudinal, torsional) and the modal structure of vibrations in elastic bodies (many of them demonstrable by means of *Klangfiguren*), Chladni did not quite come to grips with 'sound colour' or *timbre* though he mentions this French term in a chapter treating modifications and articulations of sound (Chladni 1805/1830, § 248). In line with his fundamental theory of sound as emanating from vibration of elastic bodies, he saw such modifications and articulations of sound depending on various materials (such as steel, brass, or gut used for strings) as well as in small differences in the motion within vibrating bodies resulting probably from strain and stress. This was a modern view as far as vibration is concerned yet not enough in regard to the perception of 'sound colour'. In a later writing, Chladni (1817, 58) again refers to the French word timbre,[6] saying that it denotes *die qualitative Verschiedenheit des Klanges auf die Würkung, wofür man im Deutschen keinen bestimmten Ausdruck hat*. This indicates that timbre, or *Klangfarbe* by this time was not yet a central issue for research. Moreover, Chladni's resistance against Rameau's perception of string partials as audible tones led even major music theorists in Germany to subscribe to his view. For instance, Weber (1817/1824/1830, §§ I–V, 180) relates to Chladni when stating that a *Klang ist ein einfacher und ungemischter Laut*, and that the very nature of musical instruments is to produce sounds or tones (a *Klang* of a definite pitch is called *Ton*) as pure as possible. Only instruments of a musically "lower" grade (like a snare drum and, more so, cymbals and other Turkish instruments) apparently have a number of

[6] The meaning of *timbre* in the French language comprises sound, sound colour, and also bike bell (*timbre d'une bicyclette*) as well as brand, mark and (postal) stamp.

Nebenschwingungen that result in audible *Beitöne*. These, however, Weber claims, are almost not audible in "higher" instruments such as strings and wind instruments. Partials, he (1830, 12) concludes, even though these might be found in vibrating strings, have nothing to do with the nature or the beauty of a sound and can be considered an imperfection that, however, is harmless since higher partials or *Beitöne* in reality would not be audible. Such statement left Weber in a position where he, on the one hand, could not but admit that there is something one perceives as the quality or character of a sound (*sein eigenthümliches Gepräge* in the sense of the French *timbre*), and on the other he refuses to take notice of the acoustical basis underlying any *Tonfarbe* or *Klangfarbe*.

Due to ongoing research in acoustics and also in optics (where the wavelength and spectrum of light had become an issue since Newton, and, by analogy, sound thus could hardly be conceived without the concept of a spectrum), the situation had changed by about 1850. In Opelt's *Theorie der Musik* (1852, § 7) it is hypothesized that "the so-called *Klangfarbe* depends on different kinds and shapes of pulses" (by pulses periodic wavetrains are meant), even though this cannot be proved with certainty by ear alone. Opelt assumed that sound as transmitted from an instrument indeed is a complex whole to which several vibrating structures of each instrument contribute.

Since 'pitch' was associated with wavelength and the number of vibrations per time unit, and the sensation of intensity (or, rather, loudness) with the amplitude of vibration, 'sound colour' could not be attributed to anything else but the shape of a wave. This is what Helmholtz (1863/1870/1896) elaborated in great detail consequent to the basic statement that 'sound colour' (*Klangfarbe*) depends on the microstructure of vibration of a sounding body. This microstructure of course is reflected in the shape of each period of a sound wave recorded from an instrument. In regard to perceiving 'sound colour', Helmholtz (1863) argued that the ear is capable to perform decomposition of any complex periodic sound wave into its constituents according to Ohm's law, that is, into sinusoidals (each of a given frequency and a given amplitude). Of course, Helmholtz knew the theorem of Joseph Fourier, according to which (in a brief interpretation Helmholtz gave as part of a popular lecture in Bonn 1857) any arbitrary [periodic] waveshape can be synthesized from a number of simple waves of different wavelength.[7] Accordingly, Helmholtz held that our ear does the same what the mathematician does by applying Fourier's theorem, namely "it dissolves [periodic] complex waves into a sum of simple (or elementary) waves". Helmholtz (1870/1896) also explored in an approach one may call 'analysis-by-synthesis' production of complex periodic sounds (and even inharmonic sounds) from superposition of sinusoidals. The idea was to compare natural with synthesized sounds; if a synthesized sound came close enough to the original, its spectral composition could be stated in terms of the formula used in the additive synthesis.

[7] Translations are from the edition of Helmholtz' collected lectures and speeches (Helmholtz 1896).

3.2 *A Matter of* Ausdehnung*: Stumpf on 'Sound Colour' (*Klangfarbe*)*

The study of 'sound colour' was begun, as far as instrumentation and method is concerned, on a mechanical basis, namely by use of tuning forks and other devices as resonators, and of interference tubes for cancellation of partials. Along with analysis, empirical work on 'analysis-by-synthesis' was undertaken in that complex sounds were synthesized from a set of (almost pure) tones. Though scholars recognized that sounds can undergo marked changes not only in regard to pitch and dynamic level over time but also with respect to timbral characteristics, 'sound colour' was first viewed in a more static sense, namely as an attribute of sound that is present and audible for a certain time. As Stumpf (1890, 520ff.) has discussed in a phenomenological approach to perception of specifics of 'sound colour' (*Klangfarbe im engeren Sinne*), we can assign three basic attributes to sounds: height (*Höhe*), intensity (*Stärke*), and extension *(Größe*). While height and intensity can be directly related to physical parameters (frequency, amplitude), this is not the case for the third attribute that, in certain ways, for Stumpf has to do with perception of space and of spatial properties (cf. Stumpf 1890, 50ff.). The key word he uses in this context is *Ausdehnung* (that can be translated, in this context, as extension or volume), however, in his discourse there are several 'spatial' aspects to which he relates. These range from localizing the source of a sound in three-dimensional space, that is, a task performed in binaural hearing (a process that can be described in terms of acoustics and psychoacoustics), to a more subjective assessment of sound qualities such as 'volume', 'density', 'brightness', and 'sharpness'. For example, musical tones low in fundamental frequency and 'pitch' seem to be more extended (to fill a larger 'volume', or to have a larger 'body') than high-pitched tones that, typically, appear as small and lacking 'volume'. Further, low tones often appear as dull as well as soft while high-pitched tones appear as bright and also as sharp, etc. There are a number of such phenomenal attributes that, according to Stumpf (1890; also Stumpf 1926, Kap. 15) we use to characterize the quality of certain sounds we perceive. The attributes, however, are assigned to sounds in a quantitative way. If sounds are classified, for instance in terms of 'thickness', certain sounds may be rated as being 'thicker' than others.

Stumpf (1890, 535ff.) argued that 'spatial' attributes (such as massiveness or sharpness) apply even to pure (sine) tones, that those attributes are immanent to a tone of given height and intensity, and that these attributes vary in parallel with the 'height' and the brightness of tones.[8] Also, certain combinations of the primary (objectively measurable) attributes tone height and intensity will result in certain

[8] For an in-depth discussion of the tonal and sound attributes that were used by Stumpf, Hornbostel, Wellek, etc., see Albersheim (1939) and Schneider (1997b), Kap. III.1, 404–430.

sensations that are the basis of the 'tone colour' we assign even to pure tones.[9] For example, a sine tone of high frequency and high intensity (SPL) made audible appears as sharp, or even as "piercing". In general, sensation of pitch and brightness can be said to vary with the physical property of frequency (intensity is held constant), sensation of loudness varies with physical intensity (frequency held constant); density and sharpness vary (in upward direction) with frequency and intensity, volume varies against frequency, brightness and density (with maximal volume for tones low in frequency, 'tone height', and brightness). In regard to the issue: what is the "dimensionality" of sounds, it seems obvious that even pure tones have a number of physical properties and sensational as well as perceptional attributes that can vary in a quantitative way. For the physical properties (frequency, intensity [dB]), which can be varied along a continuum, and independent of each other, the notion of a 'dimension' applies. With respect to sensation and perception, the matter is much more difficult since, as is well-known from sensation of 'pitch' and brightness of pure tones, these are 'integral variables' whereas volume and brightness are (in principle) separable (cf. Schneider 1997b, 429f.). If the notion of 'dimension' is taken in a strict sense (requiring quasi-continua for measurement on at least interval scale level and correlational independence of variables so that such 'dimensions' representing variables can be unfolded as an orthogonal structure in a *k*-dimensional vector space), it will be hard to separate phenomenal attributes according to well-defined 'dimensions'. Stumpf saw the interrelation of tonal attributes already in simple tones, for which 'tone color', he said (1890, 540) is not an attribute besides 'height' and 'intensity' but comprises "partly intensity, partly height, and partly extension" (or volume).

In regard to 'sound colour' (of complex sounds), Stumpf distinguishes characteristics he calls 'inner' moments from such he refers to as 'outer'. The 'inner' moments relate to the structure of partials in a given sound, that is, the number of partials present and the relative amplitude of each partial. Hence, the 'inner' structure of sounds concerns the energy distribution in a (typically harmonic or, in certain types of sounds, also an inharmonic) spectrum. Inner moments of sound colour can be represented by a line spectrum as well as by a spectral envelope. The 'outer' moments relate to, first of all, temporal and dynamic aspects such as the onset and the decay of sound, the occurrence of transients and noisy components at the onset (as well as noisy components even in the steady-state portion), modulation and other fluctuations. Most of these phenomena can be described in terms of the temporal envelope. Taken together, 'inner' and 'outer' moments of 'sound colour' thereby form a three-dimensional structure (Fig. 4), in which *x* represents the partials (ordered by their harmonic number and/or their frequency [Hz or kHz]),

[9] The term 'tone colour' (*Tonfarb*e, Stumpf 1890, 1926) applies to pure tones, while 'sound colour' (*Klangfarbe*) is reserved to complex tones. More empirical evidence for the fact that, by varying two physical properties (frequency, amplitude) of pure tones, one can vary more than two sensational attributes (namely, pitch, loudness, and volume), was later given by Stevens (1934).

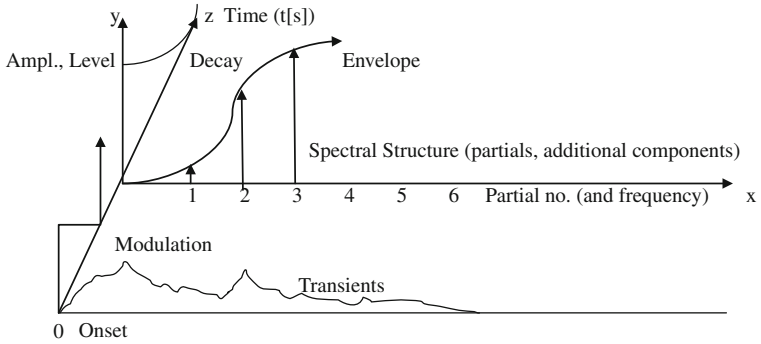


Fig. 4 'Inner' and 'outer' moments of *Klangfarbe* (Stumpf 1926, Kap. 15)

y represents the (linear) amplitude of sound pressure or the intensity of the sound,[10] and *z* represents time [s or ms].

Stumpf rightly emphasized that sounds produced from instruments quite often vary more in regard to the register in which they are located (from low or very low to medium to high or even very high depending on the ambitus or range of octaves an instrument can cover) than from one instrument to the other (the phenomenal difference for certain sounds produced by either a French Horn or a Cello in a low register thereby appears smaller than for two sounds produced by each instrument in either a very low or a very high register). This fact has led to considerations according to which one would need to keep the number and relative intensity of partials constant in order to maintain a certain 'sound colour' over several registers (cf. Slawson 1985). This idea (which owes to Slawson's concept of composing sequences of 'sound colour' equivalent to sequences of tones in a melody) would imply shifting an identical spectral envelope along the frequency axis; such an operation would, however, not preserve zones of spectral energy concentrations that are often viewed as 'formants'. Consequently, such a linear shift can have unwanted effects because sounds appear unnatural in timbre when envelopes are transposed by an octave up or down (cf. Rodet and Schwarz 2007, 177).

Another finding Stumpf reported was that even experienced musicians and instrument makers failed significantly to identify sounds produced from various instruments (presented at random) correctly if the onset including transients and the final decay were cut off from the sound, and only the quasi-stationary portion was audible for 2 s for the subjects. To be sure, such experiments again were carried out basically on a mechanical level, in this case, with sounds presented through tubes in a wall that could be rapidly opened and closed to cut out portions of sound (cf. Stumpf 1926, 374f.).

[10] Where, approximately, $I = p_{eff}\, v_{eff} = p^2/Z$ that can be transferred to SPL [dB].

3.3 The Quest for 'Formants' in Musical Instruments

Helmholtz (1863/1870) devoted one chapter of his comprehensive work to differences various musical instruments show in regard to their respective *Klangfarbe*. He included a section on vowels as occurring in speech and in singing and reported many of his own observations plus some of the research done by other scholars. After going into details of resonance phenomena in the vocal tract and in particular in the mouth cavity, Helmholtz (1870, 179) came to the conclusion that vocal sounds differ significantly from most other musical instruments in that the relative power of partials is not dependent on their harmonic number but on their absolute pitch (or frequency) position.[11] As an example, he said that the vowel /A/, when sung on the musical note Es (E^b_2), would have a resonance peak at b'' (B_5), which is the twelfth partial of Es (E^b_2). If, however, the same vowel /A/ is sung on the note b' (B_4), there is still a peak at b'' (B_5) though in this case it is the second partial. In this respect, vowels apparently did differ from a range of musical sounds where the strength of partials typically decreases with harmonic number (such as A = 1/n or in any other suitable ratio of amplitude to harmonic number).

The finding of Helmholtz, in a generalized form, implies that vowels retain spectral energy concentrations at certain frequencies or small frequency bands irrespective of the absolute fundamental frequency where phonation takes place. Though direct observation of phonation had become possible when Johann Nepomuk Czermak (1828–1873, a Czech-Austrian professor of physiology) had introduced the larynx mirror into the study of the vocal folds when in action, Helmholtz still saw the vocal tract (German: *Ansatzrohr*) consisting of the pharynx and the mouth cavity as the main part relevant for spectral shaping of vowels. Explaining vowel production basically by resonances in the vocal tract, Helmholtz (1870, 178) referred to the work of Robert Willis (1830), a professor of mechanics at Cambridge who had conducted relevant experiments with reed pipes of variable length and had claimed that each vowel can be related to a single resonance. Helmholtz's concept of vowels characterized by distinct resonances of the vocal tract and mouth cavity corresponding to certain musical pitches (e.g., /U/ to f [F_3], /A/ to b'' [B_5], /E/ to b'''[B_6]) was confirmed, in the main, by observations and measurements made by several scholars independently. It is not possible here to review the many notable contributions to the theory of vowels put forward, for the most part, by professors of anatomy and physiology like the Czech-Austrian Johann N. Czermak (1828–1873), the Dutch Franciscus Cornelis Donders (1818–1889), and the Germans Georg Meissner (1848–1905) and Ludimar Hermann (1838–1914).[12] Also of importance was the Finish-Swedish linguist Hugo Pipping (1864–1944) who published several relevant articles as well as a

[11] Helmholtz (1870, 179): *…dass die Stärke ihrer Obertöne nicht von der Ordnungszahl derselben, sondern von deren absoluter Tonhöhe abhängt.*

[12] Meissner was a professor in Göttingen when conducting his research on vowels and sounds from musical instruments in the 1880s and 1890s. Hermann was a professor in the University of

monograph (Pipping 1894) on this matter and explicitly dealt with the sound colour of vowels as sung.

It should be noted that there were some controversies concerning the nature of the *formant*, a term coined by Hermann (1894, 267) who defined it as the "characteristic tone" within a spectrum. For many sounds, Hermann had calculated the formant as a 'center of gravity' (he called it *Schwerpunktmethode*) by taking three adjacent partials with their amplitudes. Since determination of frequency and amplitudes for partials could only be done by approximation (given the mechanical tools available for analysis, see also Meissner 1907; Herrmann 1908), Hermann came to the conclusion that the formant must not always stand in a harmonic frequency ratio to the fundamental but could be inharmonic as well. This provoked severe criticism because the formant was primarily understood as a resonance phenomenon in a tube (the vocal tract) filled with air that undergoes longitudinal vibration only.

From his own measurements, in line with those of Pipping (1890), Hermann concluded *dass die Höhe der hervorragenden Partialtöne der Vocale sich mit der Notenhöhe nicht wesentlich ändert* (Hermann 1891, 181). Since the formant should be kept fixed in a certain frequency band while the musical notes of a scale rise in fundamental frequency, Hermann (1894, 268) stated as a *Grundgesetz* (fundamental law) that *der Formant mit steigender Stimmnote in der Ordnungszahl immer weiter herabgeht, seine absolute Lage dagegen behält*.

Following to the investigations of Helmholtz in regard to the sound colour of musical instruments and the human voice, it was proposed that the basic sound quality of certain instruments could be compared to vowels, whereby the bassoon should be similar to /U/, the French horn to /O/, the trombone to /A/, the oboe to the German /Ä/, and so on (cf. von Qvanten 1875). In much of the early research, the formant was viewed either as a single harmonic partial of strong intensity (the most prominent resonance in the vocal tract or in a tube filled with air such as a flute or flue pipe) or as kind of a *Mundton* (as Hermann claimed to exist) resulting from a separate regime of vibration. However, it had become clear quite soon that the specific quality of vowels as well as of sounds characteristic for a certain musical instrument resulted from groups of partials rather than from single spectral components. Meissner (1907, 595), summing up his findings, stated that (1) it is groups of higher partials with significant amplitudes that characterize both the sound of wind instruments (aerophones, *Blasinstrumente*) as well as vowel sounds of the human voice. He had found (2) such concentrations of spectral energy not to shift with different musical notes played or sung (that is, such concentrations are relatively independent of the 'pitch' played). Meissner concluded that the groups of relevant partials (3) are forming regions of spectral energy concentration which are fixed in frequency (*Eine den Klang eines Blasinstruments wesentlich*

(Footnote 12 continued)
Königsberg and internationally acknowledged as author/editor of textbooks on physiology. Some of the historical background to this era of research is briefly summed up in Graf (1980, 211ff).

charakterisierende Gruppe höherer Obertöne …ist ein festes Gebiet oder eine feste Region bestimmter absoluter Tonhöhen…). In research done in the 20th century, the term for the regions Meissner had found usually was *Formantregionen* or *Formantstrecken* (cf., e.g. Stumpf 1926; Vierling 1936; Winckel 1960).

One interesting point reported in several of the early publications (e.g., Meissner 1907; Herrmann 1908) is that, in vocal sounds such as vowels as well as in sounds recorded from various pipes, the fundamental f_1 was found quite weak or almost missing. This led to questioning the Ohm/Helmholtz-theory of pitch as applicable to all kinds of sounds and to suggesting that 'periodicity pitch' as proposed by Seebeck might be a valid alternative in certain cases. Further, it was found that air in a cylindrical tube excited by a reed (as a valve periodically opened and shut) did not yield a sound spectrum confined to odd harmonics since even harmonics with significant amplitudes were measured as well.

When Stumpf finally wrote his book on speech sounds (*Sprachlaute* such as vowels and consonants, Stumpf 1926), he could drew on a broad range of previous research plus a wealth of observations and data he and his co-workers had collected. Stumpf offered a descriptive and analytical treatment of many phenomena relevant for phonetics though he (1926, VI) considered this book mainly as a continuation of the chapters on *Klangfarbe* he had offered in vol. II of the *Tonpsychologie* (1890, § 28). In his *Sprachlaute* monograph, Stumpf also included chapters on the psychology of listening and on topics relevant for psychoacoustics. Concluding his book, he added a chapter on instrumental sounds (Kap. 15: *Über Instrumentalklänge*) in which he treated the problems of 'sound colour' once more (on the basis of experiments he had conducted for many years after the second volume of the *Tonpsychologie* had been published).

Naturally, one of the chapters on the analysis of sung vowels relates to 'formants' (Kap. 2, 62ff.). Formants according to Stumpf are not necessarily confined to single partials (as most of previous research from Helmholtz to Hermann had suggested) but rather may include several partials so that energy is concentrated in a frequency band (*eine Strecke des Tongebietes*). Stumpf distinguished between a *Hauptformant* (the frequency band in which the most prominent spectral peak is found) and one or several *Nebenformanten* that appear as additional relative maxima of spectral energy distribution and can be found above or below the *Hauptformant*. As to the debate whether formants are rising in parallel with f_1 of the notes sung or played on instruments (usually within one octave) or are almost fixed in their frequency position, Stumpf (1926, 62ff., 191f., 377f.) did not accept the view of Hermann (1891) concerning "frequency-fixed formants" but took kind of an intermediate approach in that he argued that formants are only relatively stable in their frequency position, and will shift in the direction of the notes played or sung in a scale: *ihre Bewegung erfolgt in gleicher Richtung, aber weit langsamer, sie umfaßt (abgesehen vom Grundton c2) nur wenige Töne*. Thus formants should shift only slightly with rising 'pitch', and only within certain boundaries. Probably due to his focus on vowels, Stumpf reported spectral energy maxima for a number of instruments (in particular woodwinds and brass) which he, in analogy to speech, addressed as formants. He saw a main formant (*Hauptformant*), typically

comprising harmonic partials no. 4–6, shifting relative to the fundamental frequency of different musical tones that were played, while several *Nebenformanten* (adjunct formants) that might also occur according to his observations should remain within a certain frequency range.

It should be noted that, after Hermann's *Schwerpunktmethode* for finding the center of formant regions, a new method was proposed by Vierling (1936) on the basis of the phase relationship between several partials making up a formant group for which a spectral envelope with a single peak can be found in an amplitude spectrum. Taking the frequency corresponding to an envelope maximum as the resonance frequency ω_{res} (for a resonance filter in a generator/resonator model), its phase would be zero; it follows from an equivalent resonance filter curve that for all partials below ω_{res} (as the point of phase inversion), the phase is negative, and for all partials higher in frequency, the phase is positive. Vierling suggested that, by superimposing the partials with correct amplitudes and phases, a new wave results, whose envelope periodicity frequency would represent the center of the formant region.

Not long after *Die Sprachlaute* had been published, Stumpf's former student Karl Erich Schumann submitted his Habilitationsschrift on *Die Physik der Klangfarben* (typewritten, 1929) to the Berlin university.[13] Schumann, who—from what is known today as factual evidence—took his Ph.D. with Stumpf (in 1922),[14] had specialized in musical acoustics and had published a small book on this matter in a popular scientific series. His treatment of formants in this publication (Schumann 1925, 76–79) sums up some of the earlier findings and discussions (Helmholtz, Hermann) and refers briefly to experiments Stumpf had conducted on the synthesis of vowels. It is not possible, at this place, to discuss Schumann's work in any detail though it became influential in some circles in the 1970s, and has been referred to more frequently since (see Mertens 1975; Fricke 1993; Reuter 1995; Reuter 1996). Ironically, Schumann's *Physik der Klangfarben* never got published due to, it seems, effects of World War II (while Schumann played an important role in the organization of the Nazi war effort and in the development of weapons, in particular high-grade explosives; cf. Nagel 2007).

Schumann had postulated four *Klangfarbengesetze* (laws reigning 'sound colour'; see Mertens 1975; Reuter 1996, 110ff.) which all relate to formants in the spectra of musical instrument sounds. Though there are certain patterns or even

[13] A copy of this Habilitationssschrift is kept in the library of the Staatliche Institut für Musikforschung of Berlin. Another copy apparently is in the Institute of Musicology at Cologne.

[14] As to Schumann's biography, see his Wikipedia entry and Nagel (2007, 232ff). Stumpf became an emeritus in 1921 but continued to teach (including supervision of candidates) until late in the 1920s. According to Reinecke (2003, 185) who knew Schumann personally, he was *Schüler und langjähriger Assistent Carl Stumpfs*. However, Stumpf (1926, 7), in the *Einleitung* to his *Sprachlaute*, mentions eleven co-workers and colleagues who had helped him in experiments on the analysis of vowels (notably Dr. von Allesch), but not Schumann who (cf. Nagel 2007, 232ff.) apparently was a paid assistant in the Institute of Physics where he worked with Arthur Wehnelt, well known for his discoveries in electronics.

regularities that can be found empirically in the structure of spectra of certain types of instruments (e.g., reed-driven aerophones, see Voigt 1975), the notion of a 'law' would require generalization that, from the evidence at hand, seems hard to justify yet. The concept of 'formant' itself has been given somewhat different interpretations in regard to vowels and musical sounds (see above). In a general view, *the term is synonymous with resonances in the vocal tract* (Slawson 1985, 38). According to acoustic fundaments of speech production in a system that can be described as a source/filter model comprising a generator (the vocal folds) and a resonator (the vocal tract up to and including the lips), formants are regarded as resonances in a cylindrical tube of length L(cm) closed at one end (a $\lambda/4$ resonator). Hence, the resonance frequencies should be found at $F1 = c/4\ l$, $F2 = 3c/4\ l$, and $F3 = 5c/4\ l$ of a resonator of length L with c = speed of sound in air (see Neppert and Pétursson 1986) where $F1$, $F2$, $F3$ represent the three (most prominent) formants. For the sound pattern corresponding to the neutral vowel /ə/, and the typical male resonator tube of $L = 17.5$ cm length, the formant frequencies should be close to 500 Hz, 1500 Hz, and 2500 Hz, respectively. This is an idealization, though, since both the generator and the resonator can be varied dynamically by speakers or singers during phonation. As has been demonstrated by Fant (1960) by means of measurement and modeling, the cross section profile and even the length of the vocal tract (German: *Ansatzrohr*) as well as other parameters relevant for 'pitch' and spectral energy distribution can be modified in the course of phonation processes. There is a range of articulation effects resulting from small changes in the position of the larynx, the lower jaw, the tongue (body and tip as two separate subsystems), and the lips (for details, see Sundberg 1997) which lead to modifications also of formants in regard to their bandwidth and frequency position relative to a fundamental as well as to changes in spectral energy distribution and spectral envelope of sound radiated from the mouth.

The distribution of spectral energy for an individual singer and a particular vowel sung with the 'colour' of a certain language can be determined by analysis of steady-state portions of sounds. For example, taking the long note sung by a woman on the vowel /ɒ/ at the beginning of her rendering *Abu Zeluf* (see the left half of the melogram in Fig. 2), one gets a fairly stable spectrum with a pattern of peaks shown in Fig. 5.

The sound actually contains spectral energy up to at least 12 kHz. The fundamental is found near 333 Hz, with the second harmonic (667 Hz) strongest in amplitude. Taking the spectral envelope in Fig. 5 calculated from a formant filter analysis, one finds peaks at harmonics nos. 2, 4, 9 and a zone of condensed spectral energy carried with the harmonics 11, 12, 13. Taking a template for vowel formants (cf. Födermayr 1971, Abb. 21), one sees that for the vowel in question energy maxima should occur in four frequency bands: the first is around 700 Hz, the second around 1 kHz, the third covers a band from, roughly, 1.8 to 2.7 kHz, and the fourth a band from about 3.3 to 4 kHz.

As had been stressed already in earlier research, certain sounds produced from various musical instruments appear as similar to vowels when such are sung. Phenomenal similarity of course can be rated by subjects, and especially by those

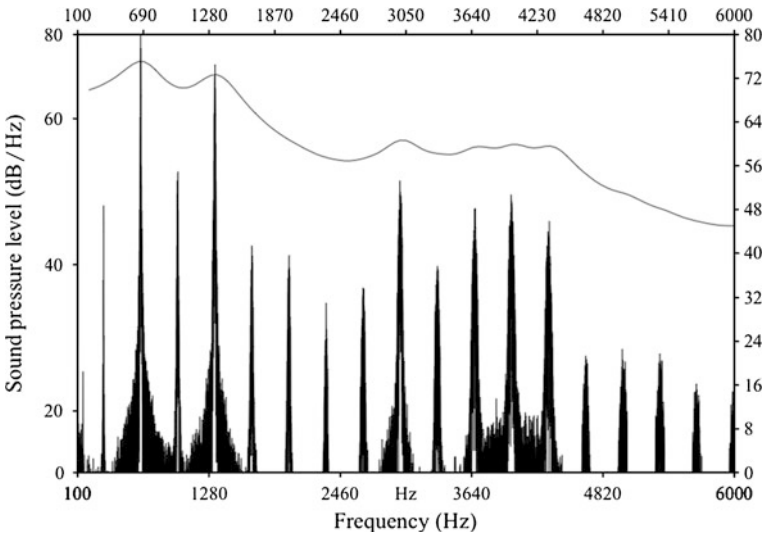


Fig. 5 Spectrum and formant filter envelope, vowel / ɒ / as sung by Dunya Yunis

with a musical training. However, 'vowel quality' in violin sounds can be detected also with signal processing tools where the output is referenced to a Jones-type diagram of vowel categories (see Mores 2010). Since subjective ratings of the 'vowel quality' in violin sounds and automated extraction of vowels show a high degree of convergence, there must be distinctive features which permit to classify sounds as to their 'vowel quality'. In retrospect, this has led to the issue of 'formants' as well as to several explanations why formants are genuine to musical sounds. Given that the steady-state of most vowels produced in singing shows a pattern of peaks and/or concentrations of energy (the latter may be connected with a single partial or with a group of partials) in the envelope, one can address formants accordingly as (1) single prominent spectral peaks, as (2) groups of partials with amplitudes or intensities above the average level of neighbouring partials, or (3) simply as frequency bands in which more spectral energy is found than in adjacent regions of the spectrum. The "ideal" formant spectrum, then, would be characterized by a regular pattern of spectral maxima and minima so that the envelope would show a cyclic structure defined by zeros. Such spectra can be obtained from trains of rectangular pulses with a duty cycle of τ/T (where τ is pulse width [ms], and T is the pulse period[ms]). Amplitudes of partials in the spectrum of a train of rectangular pulses conform to a sinc function (sin [x]/x) where the zeros are found at $n\ \tau$/T = 1, 2, 3, …(cf. Meyer and Guicking 1974, 40f.). The function useful for demonstrating a cyclic spectrum would be of the form Sin[x]^2/x so that the envelope is all positive (Fig. 6); the roots of the function of course are found at $n\ \pi$, $n = 1, 2, 3, \ldots$ The peaks and troughs of the envelope can be conceived as covered by another envelope that represents the overall exponential decay of amplitudes (dashed line):

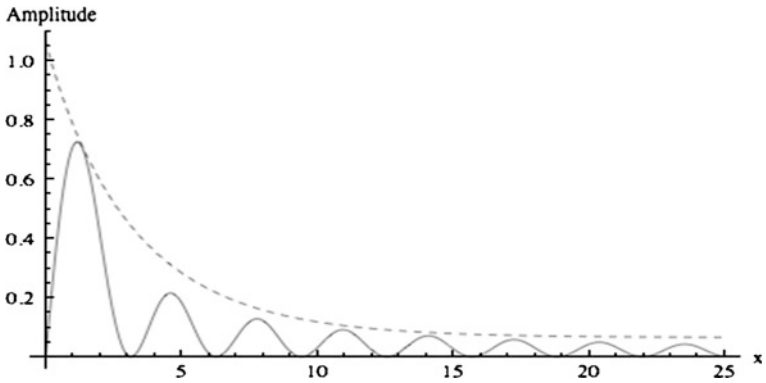


Fig. 6 Model of a cyclic spectrum (envelope of the form $\mathrm{Sin}[x]^2/x$)

Since spectra obtained for sounds of certain musical instruments such as reed-driven aerophones basically exhibit a more or less cyclic structure [as is the case for a number of double reeds (cf. Voigt 1975) and also reed pipe stops in organs, see Beurmann et al. 1998], the explanation at hand is taking the reed as a pulse generator that should produce, in theory, a sequence of rectangular pulses where the pulse of height A and duration t corresponds to the time the valve is open so that a flow U of air at a certain pressure p is released into the resonator while the time between pulses marks the duration for which the valve is shut. Though such a model serves to explain the basic principle, things are quite complex in reality since even for instruments with a beating reed such as the clarinet the vibration of the reed and the pressure measured in the mouthpiece (cf. Backus 1963) do not yield a rectangular shape equivalent to a pulse (which is defined as a [discontinuous] jump-function). Rather, reed vibration and pressure seem to be almost sinusoidal when excitation of the reed is very soft, and approximate a triangular shape with medium blowing pressure. Only when driven hard, motion of the clarinet reed becomes more of a rectangular wave. It was observed that for soft blowing the valve never completely shuts while it shuts for about a half-cycle when high blowing pressure meaning a very strong force acting on the valve is applied.

If we take a comparatively large double reed mounted on a small brass tube as it is used for the French-Breton *bombarde,*[15] the reed generator produces a periodic change in pressure at the output (see Schneider 1998, Figs. 1, 2) that becomes audible as the source signal since the double reed is fixed to a small cylindrical

[15] The triangular double reed used for measurements is 10.5 mm wide at the opening. It is 20 mm long and is mounted on a cylindrical tube of 25 mm length and 4 mm diameter. Thereby, the total length of reed and tube is 45 mm. Recordings were made with the reed and tube put close in front to a condenser microphone (Neumann U 67, AKG C 414 B-TL II) set to cardioid and fed into a preamp and A/D converter system (Telefunken V 76, Panasonic DAT SV 3800 at 16 bit/48 kHz and RME Fireface 800 at 32 bit float/96 kHz on hard disc.

tube of 25 mm length that already may act as a (first) resonator. The *bombarde* can be modeled as a coupled system comprising the valve as such, this small brass tube, the conical bombarde pipe of 27.5 cm length (with 8 mm bore at the upper, 15 mm at the lower end) plus the bell which has an effective length of 3.4 cm and opens from 16 mm to 48 mm. For the double reed on the small brass tube, at soft to medium pressure force, the wave shape of the source signal is approximately triangular (comparable to the beating reed generator of the clarinet) while the signal becomes more complex with respect to its fine structure per period with increasing blowing pressure. The reed generator (double reed bound to its proper tube) for medium excitation yields a spectrum comprising a series of quasi-harmonics starting at ca. 698 Hz as listed in Table 1.

Table 1 *bombarde*, reed generator output spectrum 0.5–14 kHz (FFT: 16384 pts, Hanning)

Partial no.	Frequency	dB[a]	Partial no.	Frequency	dB
1	698	−59.4	11	7,658	−73.3
2	1,395	−13.4	12	8,372	−43.7
3	2,094	−54.8	13	9,063	−73.6
4	2,790	−27.4	14	9,767	−42.7
5	3,510	−77.9	15	10,446	−73.7
6	4,186	−17.8	16	11,162	−42.2
7	4,837	−66.0	17	11,853	−72.5
8	5,581	−41.3	18	12,556	−52.0
9	6,248	−74.2	19	13,247	−83.3
10	6,976	−34.4	20	13,957	−62.8

[a] SPL relative to 0 dBFS; therefore, all dB readings for undistorted signals are negative (from 0 to −100 dB)

There are some more partials above 14 kHz up to 22 kHz so that excitation of modes in the resonator is secured even at modest blowing pressure. It is obvious that partials no. 2, 4, 6 are exceedingly stronger than their neighbours, partials 1, 3, 5 and 7, so that they will make up good candidates for ‘formants’. In fact, if the reed generator sound is put through a formant analysis using the routine implemented in the Praat software,[16] it turns out three ‘formants’ (Fig. 7).

One can see that the three ‘formants’ in fact correspond to partials no. 2, 4 and 6 of the reed generator spectrum (Table 1), and that there is considerable energy fluctuation in partials 2 and 6 over time while partial 4 remains stable.[17] Whether one may call the three strongest partials of the generator spectrum ‘formants’ because they ‘stand out’ against their neighbours seems a matter of terminology

[16] The analysis is based on STFT with an approximation of spectral peaks in the envelope per analysis frame by the Burg method (cf. Marple 1987). The three ‘formants’ in fact represent tracks of spectral envelope peaks plotted against time.

[17] The fluctuation is evident from so-called speckles (black dots) which indicate spectral energy peaks per analysis frame.

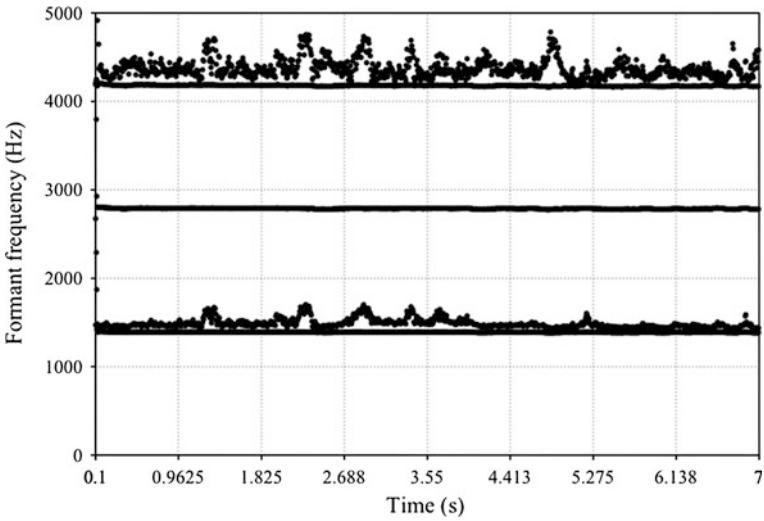


Fig. 7 *bombarde* reed generator output, formant analysis (Burg algorithm)

rather than of principle. However, in this case each of the three 'formants' of the reed generator would consist of but one partial.

In case the generator is excited with maximum admissible blowing pressure (there is of course a limit beyond which the two reeds are pressed against each other and the valve simply remains shut), crests of the wave become quite steep and the wave shape even more complex per period due to additional modes of vibration.[18] Consequent to overblowing, the bandwidth of the source spectrum in this condition starts with partial no. 2 indicated in Table 1, and increases with blowing force and pressure applied to the valve so that the source spectrum of the reed generator carries energy up to about 33 kHz. The spectrum shows distinct and strong peaks up to 10 kHz from where on peaks are broader; also, there is considerable energy distributed in between peaks due to noise from the air flow through the valve. The reed generator spectrum however does not reveal a cyclic structure with clear dips or gaps at certain partials as one would expect from a system that produces rectangular pulse trains as output.

If driven with a medium blowing force applied to the reed generator, *bombarde* sounds in the normal playing range (i.e., not overblown into the higher octave) have spectra like that displayed in Fig. 8:

[18] In reed-driven instruments such as the clarinet, a range of nonlinearities is observed with respect to the vibration of the (single or double) reed, the flow of air through the slit as well as in other parameters. For details in regard to modeling and calculation, see Dalmont et al. (2000), Kergomard et al. (2000).

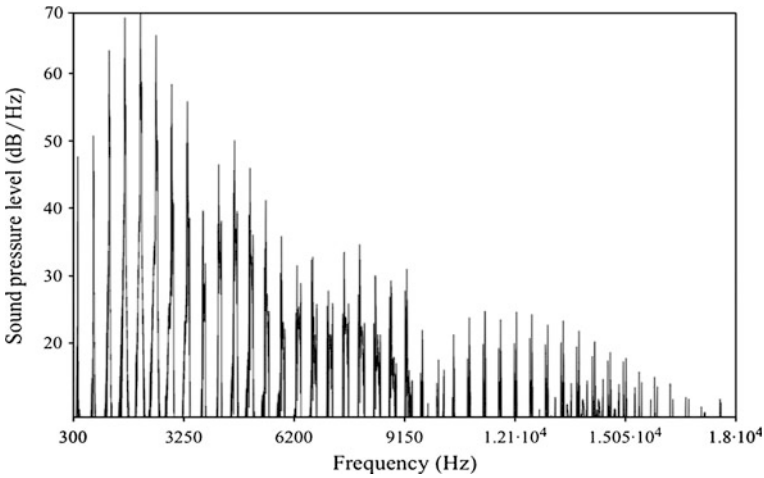


Fig. 8 Spectrum, *bombarde* note 1, $f_1 \approx 419$ Hz (0.3–18 kHz, 42 partials displayed)

The spectrum can be structured into four main groups of partials, which cover (I) partials 1–9, (II) 9–15, (III) 15–24, and (IV) 24–41 so that the partials two groups have in common are spectral dips that separate two neighbouring groups from each other. If an envelope is put on the partial amplitudes for each group, it will approximate somehow a shape like ∩ (an inverse U), however, the spectrum hardly can be called cyclic in any strict form (according to the model in Fig. 6). Though sound spectra for most notes of the *bombarde* played within one octave are similar in structure, there is still some variation which is also evident from spectral statistics given in Table 2 (all sounds normalized to -3dB before analysis):

Table 2 *bombarde*: notes 1–8, f_1, spectral center of gravity and SD[a]

Note/tone	f_1(Hz)	Center of gravity(Hz)	SD(Hz)
1	419	2082.77	642.83
2	452	2156.88	528.73
3	502	2098.26	804.65
4	566	2451.21	847.87
5	597	2561.56	1101.84
6	673	2720.42	1051.01
7	752	2824.61	1317.79
8	839	2928.04	1225.11

[a] The center of gravity is the average of f over the frequency band covered by the analysis, weighted (in this case) by the power spectrum (for details, see documentation in the Praat manual)

The data with the center of gravity rising from one note and sound to the next (except for note/sound 3) indicate spectral energy basically seems to shift with the rise of f_1 rather than to remain fixed in a certain frequency position. If segments of

one second each of the sounds 1–8 are subjected to a formant analysis (Burg method), there will be up to five, and typically four, tracks that are created in the frequency range relevant for formants with respect to a male (4.5 kHz) or a female voice (5.5 kHz) (Fig. 9).

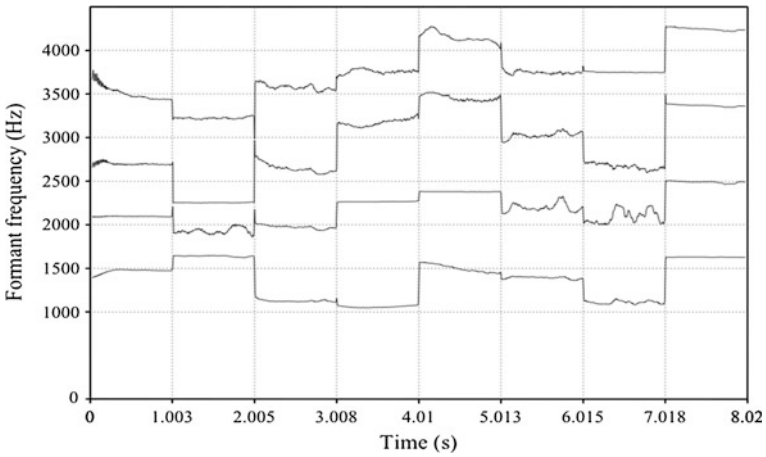


Fig. 9 Formant analysis (Burg), segments of *bombarde* sounds 1–8

One can mostly relate the four formants per sound to particular partials or to pairs of such partials. For instance, in the first sound segment the lowest formant at 1.5 kHz is very close to the strong partial no. 4 of the spectrum (Fig. 8) at 1674 Hz/68.2 dB, formant 2 represents the most prominent partial of the spectrum (no. 5 at ca. 2093 Hz/68.7 dB), formant no. 3 is in between partial no. 6 (ca. 2511 Hz/65.6 dB) and no. 7 (2929 Hz/58.6 dB), etc. Whether there are any "formant laws" inherent in the sound especially of reed instruments (cf. Mertens 1975; Voigt 1975) that could be derived from spectral analysis at present seems hard to tell; in any event, one would have to analyze a sample of many sounds to justify such a conclusion in regard to principles of inductive generalization.

Though reasonable approximations to cyclic spectra can be found in the sound of reed-driven instruments (and in particular in organ reed pipes, cf. Beurmann et al. 1998; Schneider et al. 2001), there are probably much closer approximations to cyclic spectra at hand for plucked strings of harpsichords where the harmonics that are cancelled out are determined by the ratio of the string length and the plucking point, *L*/*l* (for examples, see Beurmann and Schneider 2008, 2009). An approximate cyclic spectral envelope of course indicates that there are bands where energy is concentrated relative to the dips or gaps in between. Following common terminology, one may address such concentrations of energy in groups of partials as 'formants' even for harpsichord sounds. It would need empirical evaluation involving experienced listeners to find out to which extent such sounds in perception might appear as similar to sung vowels. After all, it should be remembered that the concept of 'formant' was developed for *vowels* as observed in *singing*, in the first place.

3.4 *Stumpf Reconfirmed: Periodicity, Harmonicity,* Verschmelzung, *Consonance*

'Inner' and 'outer' features of *Klangfarbe* condensed into Fig. 4 are suited to characterize a large number of sounds in an objective way (namely, by signal analysis as well as through psychoacoustic experiments). In regard to objective criteria, Stumpf (1926, 390) identified *Klangfarbe im engeren Sinne* with spectral structure, and subjectively with the sum of so-called *Komplexeigenschaften* resulting from listening to sounds composed of partials. To Stumpf (1926, 278), sounds comprising a number of partials (and in particular vowels) in normal perception (i.e., holistic perception not directed to structural analysis) will appear as complexes that are not segregated or ordered (*ungegliederte Komplexe*). To Stumpf, sound colours were classic examples of *Komplexeigenschaften* since they are perceived as wholes that must not necessarily be analyzed into their constituents. A different matter though is a complex of partials that is experienced as giving a special perceptual quality, that of *Verschmelzung* into a highly consonant formation (cf. Stumpf 1890, 1898, 1926). Stumpf himself had experienced highly consonant complexes of partials by listening to organ mixture stops as well as by synthesizing vowels; applying attentive listening to such sounds, he regarded *Verschmelzung* an experience that reflects not just "fusion" of partials but even more so their interpenetration (*Durchdringung*; cf. Stumpf 1890, 128ff.). Stumpf's view was that the perception of spectral *Verschmelzung* and *Durchdringung* must have a neural basis, and that apperception of consonance would most likely be effected on the cortical level.

It is characteristic of such complexes (cf. Schneider 1997a, b, 2000) that perception can switch between two different modes, one being holistic and one analytic.[19] While the holistic experience of such a complex of harmonic partials as produced by a chord of three or more complex tones in just intonation (see examples in Schneider 1997a; Schneider and Frieler 2009) rests on facts that can be described in terms of acoustics and psychoacoustics (namely, strict periodicity of the time signal, clear microstructure of the waveshape with steep crests at the beginning of each period, absence of beats and roughness, strict harmonicity of the spectrum, strong difference and combination tones present at appropriate SPL), the analytic approach can be directed to segregating the complex sound into several constituents whose relational structure is perceived and cognitively evaluated (this is one of the reasons why Stumpf and other psychologists with a philosophical background maintained to use the notion of apperception for such a process; cf. Stumpf 1926, 279, 372; Schneider 1997a, b). To illustrate the case, we synthesized a sound Stumpf himself (1926, 394) has proposed for demonstrating the effect of a special concord (*Zusammenklang*) that is perceived as *sehr einheitlich, aber noch reicher und von wunderbarer Schönheit* (even more coherent and richer in

[19] The difference is not identical with, yet bears some parallels to, that between bottom-up and top-down analysis as outlined in Bregman's *Auditory Scene Analysis* (Bregman 1990).

harmony than another concord that had been proposed by Dayton Miller). Stumpf's concord comprises five perfect triads played simultaneously that are build on the root frequencies of 100, 200, 400, 800 and 1600 Hz, respectively. Hence, we need 15 partials, which are arranged so as to match Stumpf's desired spectral structure (that included an envelope where the intensity decreases regularly from one partial to the next; in our sound, damping per partial is set to -3dB). Also, when synthesizing the sound (done with Mathematica) we put a temporal envelope on the partials for a smooth decay just as Stumpf had suggested (Fig. 10).

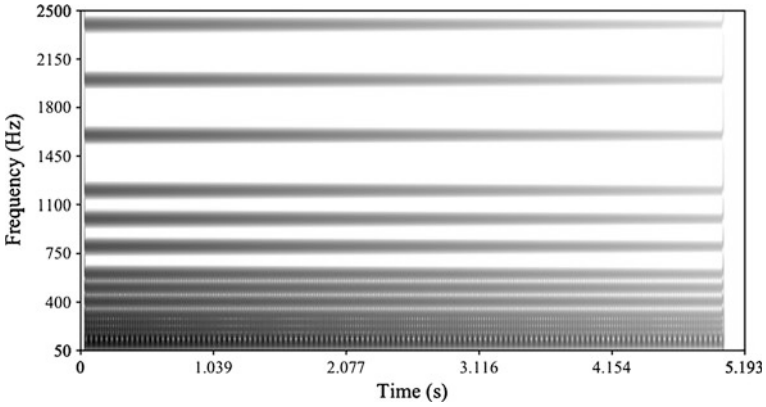


Fig. 10 Stumpf's perfect *Zusammenklang*, 5 × 3 partials = 5 just triads

Hearing such a complex, the attentive listener indeed can switch between a holistic mode where he or she will perceive a chord-like sonority of fifteen tones, and a more analytic mode directed to evaluating the spectral structure of the complex as well as the tonal relations between the five triads. Though there are certain limits for such an analysis as far as frequency resolution of the ear and, consequently, identification of components by means of perception is concerned, the pitch structure and also the interplay of some of the partials in this complex can be recognized. The basis for a cognitive evaluation is an analysis that takes place in the auditory pathway consequent to two interacting mechanisms, peripheral spectral analysis and the detection of periodicities arising from spectral components as well as from the temporal envelope. The autocorrelation process (ACF) will also reveal periodicities corresponding to virtual pitches including a series of subharmonics.[20] Hence when fed into a model of the auditory periphery working in the time domain that processes complex sounds from basilar membrane filtering and hair cell transduction to neural nerve spike representation in the auditory nerve

[20] Not to be confused with the subharmonic matching process that has been proposed by Terhardt (1998) for 'pitch' (f0) estimation of complex sounds; of course, there are some correspondences in both concepts.

(see Meddis and O'Mard 1997), the output aggregates periodicities found in the neural activity pattern (NAP) into a sum ACF (SACF). As can be expected, the neural output contains all periodicities inherent in the sound input plus virtual pitches and subharmonics (see Schneider and Frieler 2009). Exactly this will happen also with Stumpf's 'paradigm sound' (1926, 394: *idealer Klang*) since an analysis performed with various algorithms including standard and advanced autocorrelation as well as cepstrum analysis (see Mertins 1996, 1999; Arfib et al. 2002) not only turns out lags (ms) corresponding to actual partial frequencies (such as 10 ms = 100 Hz) but also gives the two main periods (10, 25 ms) governing this complex sound. In addition, the analysis yields a series of subharmonics below the fundamental ranging from 50 Hz down to 10 Hz. Hence we confirm once more (cf. Schneider 1997a; Schneider and Frieler 2009) what Stumpf had experienced and what could be expected taking fundamentals of acoustics and psychoacoustics into account: sounds composed of harmonic partials organized into several complex tones representing a concord in just intonation will result in the perception of a highly consonant formation having a distinct 'Gestalt' and sensory quality. The explanation on the level of the sound signal of course is the causal relation between strict periodicity of the time signal and perfect harmonicity of the spectrum as defined by the Wiener-Khintchine theorem (cf. Meyer and Guicking 1974, 110ff.; Hartmann 1998, Chap. 14).

The following example should demonstrate the robustness of the principle: consider three cosine pulse trains which have fundamental frequencies at 300, 400, 500 Hz plus a number of harmonics so that equal energy is contained in the spectrum at {300, 400, 500, 600, 800, 900, 1,000, 1,200, 1,600 Hz}. The pulse trains (sampled at 16 bit/44.1 kHz) are in a harmonic ratio and might be regarded as representing both a sound signal and the corresponding neural output. The time function $s(t)$ arising from the superposition of the pulse trains for 100 ms is shown in Fig. 11 (for an almost identical SACF output derived from processing in an auditory model, see Schneider and Frieler 2009, Fig. 3).

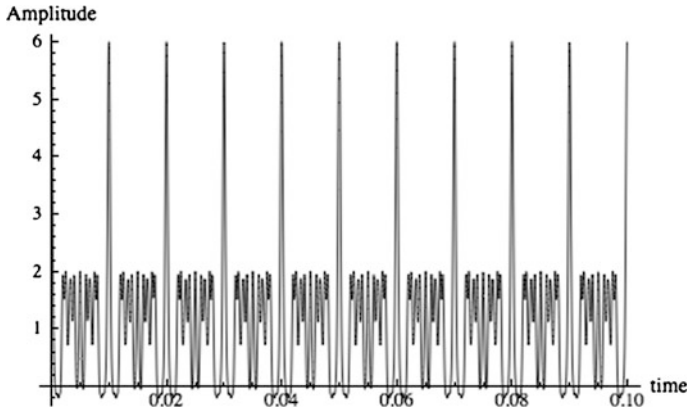


Fig. 11 Periodicities inherent in pulse trains based on harmonic ratios 3:4:5

From Fig. 11 it is easy to see that the signal is strictly periodic with a period $T = 10$ ms implying a repetition frequency f0 of the compound pulse of 100 Hz that constitutes the 'root' of the harmonic series and gives rise to the sensation of a salient low virtual pitch. The complex sound thus composed of course is highly consonant in regard to Stumpf's concept of *Verschmelzung*.

If the pulse trains are subjected to frequency modulation (FM) with individual modulation frequencies per Fourier component, the resulting sound is audibly shifting up and down in pitch and spectrum so that the degree of *Verschmelzung* one may assign to the percept is much less than that of the unmodulated sound (Fig. 11), and the time function $s(t)_{\mathrm{fm}}$ of the compound pulse at a first glance appears quite irregular (Fig. 12; 100 ms displayed).

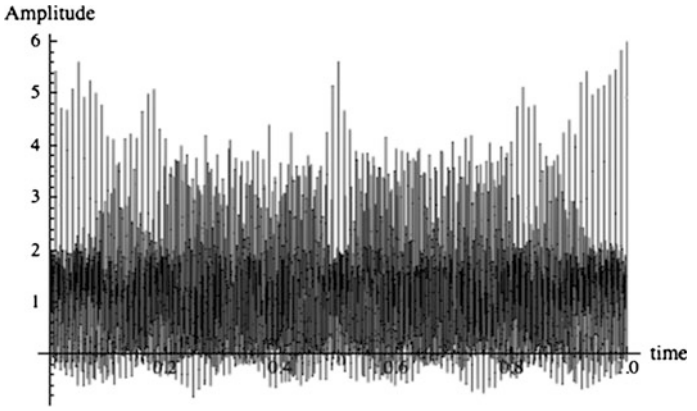


Fig. 12 Pulse trains with fundamental frequencies in the ratio of 3:4:5 subjected to FM; all Fourier components have individual modulation frequencies $A_n \sin(2\pi\, kt)$, $k = 1, 2, 3, \ldots$

Though it is not easy to detect the periodicities still inherent in the modulated time function $s(t)_{\mathrm{fm}}$ from looking into the graph, autocorrelation analysis still finds some of the Fourier components making up the compound as well as f0 clearly marked at 100 Hz. The reason is that FM in this sound also is periodic (though with different modulation frequencies applied to different components), and that the autocorrelation function (ACF) by Norbert Wiener (1961) explicitly had been defined as the temporal mean of the product $s(t)\ s(t + \tau)$ in order to detect periodicities even in complex and/or noisy signals (in particular, EEG recordings). Hence, ACF and, similarly, cross-correlation (CCF; see Hartmann 1998, 346ff.; Ingle and Proakis 2000) such as used to compare two sequences (e.g., two audio signals or one audio signal recorded from both ears) is a tool suited to detect periodicities in various signals; ACF and CCF is robust in regard to angular modulation (frequency, phase) as long as the modulation itself is periodic (or nearly so). Therefore, pitch extraction based on ACF or CCF works also for slightly detuned intervals (as is the case in equal temperament) or even for inharmonic signals up to a certain degree of inharmonicity (since increasing inharmonicity of the spectrum implies decreasing periodicity of the time signal).

As far as auditory perception is concerned, Licklider (1951, 1956) it seems was the first to draw on ACF for a model of pitch perception where *the autocorrelation mechanism consists of many delay-line autocorrelators in parallel*. In addition, Licklider proposed a cross-correlational operation conceived as a *time-coincidence arrangement with two inputs and two delay lines running in opposite direction*. Though the basic idea of time-domain pitch analysis (that had been proposed by scholars such as Seebeck and Schouten) has been widely accepted and has been realized in many models of "neurally inspired" auditory perception (for an overview, see de Cheveigné 2005), the ACF-based model suffered from a lack of convincing neuroanatomical and neurophysiological evidence since an array of neural delay-line autocorrelators was not yet discovered. Also, processing speed in the auditory pathway is at odds with the huge delay needed to perform ACF for low pitches. On the other hand, in neurophysiological research undertaken from the 1960s onwards (summarized in Hesse 1972; Keidel 1975; Ehret 1997; Schneider 1997a, b; Nelken 2002), there had been measurements of periodicities in the auditory nerve as well as on the level of higher relays of the auditory pathway (notably, the ICC and the CGM) corresponding to periodicities in input signals. More recently, the neural basis for differences between consonant and dissonant musical intervals has been demonstrated by all-order interspike interval histograms recorded from auditory nerve fibers of cats in animal experiments (Tramo et al. 2001; see also Cariani 2004). Also, an improved auditory processing model that includes the auditory nerve, the cochlear nucleus (CN), and the inferior collicus (ICC) has replaced the ACF process by operations on a huge array of components which are *physiologically plausible* (Meddis and O'Mard 2006).

3.5 Transients and Dynamic Evolution of Sound Spectra

In addition to finding pitch curves based on either f_1 or f0 measurements (see above), spectral structure and spectral energy distribution have always been of interest. After decades of research where analysis and synthesis of sounds had been confined to mechanical instrumentation, the 1920s saw the breakthrough of electro-acoustics. Stumpf (1926, 408f.) points already to experiments in radio stations on stereophonic recordings done with several microphones and finds such trends fitting his own concepts, namely, spatialization of sound.

In the 1920s, sound analysis gained new impetus when filtering based on the so-called search tone method (Grützmacher 1927; see also Küpfmüller 1968, 122ff.) allowed spectral decomposition of a time signal. This led first to finding spectra for steady-state sounds. At the same time, transients preceding almost stable sounds such as vowels in speech or tones played on woodwind and bowed string instruments were investigated. The problem in such an approach is that transients lack the clear periodic time structure which governs the steady-state portion of a given sound. Hence, a Fourier analysis that (in principle) assumes strict periodicity of the signal is difficult to conduct if possible at all. Backhaus (1932) who offered a

thorough study of transients (labeled *Ausgleichsvorgänge* then) had the idea to take the period found for the steady-state portion of the signal following the transient portion and see if spectral decomposition could be carried out by tentatively applying the known period length of the stationary signal to the transient part. In many instances, a fair approximation was possible (including manual interpolation of step functions to obtain smoothed curves for the dynamic evolution of partials over time; see Backhaus 1932, 32–34). The results were presented as a family of curves for the partials studied in a 2D amplitude/time diagram. The concept resembles a modern approach, namely the phase vocoder and the 3D-Plots one can obtain for the evolution of partials with appropriate software (such as *sndan*, see Beauchamp 2007). Since the phase vocoder analysis can be understood as a bank of band pass filters tuned to a fundamental frequency (whereby the center frequency of the lowest band pass should closely match f_1 of the sound to be analyzed), a decision as to f_1 or f0 of the signal has to be made not unlike the considerations Backhaus had outlined. Phase vocoder analysis effects spectral decomposition of a complex harmonic sound into its partials. One option of *sndan* (see Beauchamp 2007) is to plot the amplitude for each partial against time to facilitate the study of onsets and the evolution of spectral energy with time for various partials. For example, looking into the sound produced on a bassoon for the note B_2 ($f_1 \approx 120$ Hz), the 3D plot shows different trajectories for the first six partials within the first 300 ms of sound. The slow rise of most of the partials indicates a rather soft attack for this sound (Fig. 13).

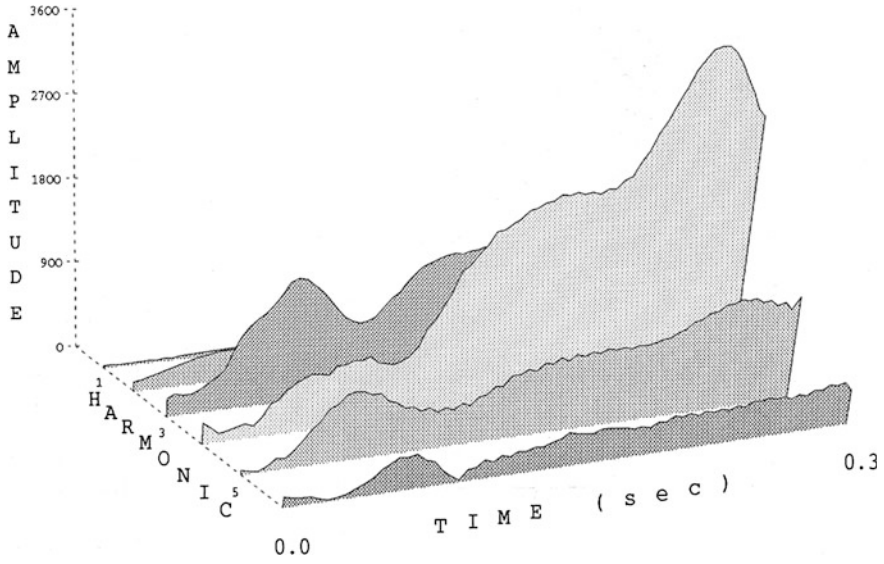


Fig. 13 Phase vocoder analysis, harmonics 1–6, bassoon note B_2, linear amplitude

Of course, analysis techniques such as the phase vocoder are based on digital signal processing and were not at hand in the 1930s. Acousticians found, however, a clever technique for the study of transients by applying octave-band filters where, because of the wide bandwidth, transient response time of the filter was short enough not to affect the filter output in a significant way.[21] The filter output per octave and the original signal were recorded on sound film and then plotted as oscillograms whereby different onset times for partials in consecutive octave bands became apparent. Such an analysis was carried out, among others (see Graf 1972; Reuter 1995), on organ flue and reed pipes (Trendelenburg et al. 1936). Studying pipe ranks from the famous organ Arp Schnitger had built for the chapel in the palace of Berlin, in 1706 (cf. Edskes and Vogel 2009, 138–143; the organ was destroyed in WW II), it was found that the onset of partials differs significantly from reed pipes to flue pipes (what can be explained by taking into account characteristics of different generators and regimes of vibration) and also with respect to various pipe geometries and pipe sizes per octave.

Basically the same approach as *Suchtonanalyse* is known also as heterodyne filter analysis (see Fant 1952; Roads and Strawn 1996, 548ff.), which has been adopted in the construction of the analogue sonagraph (designed by Bell Labs) that came into research institutions as a stand-alone hardware unit (built by Kay Elemetrics Co.) in the 1950s and was primarily used for visualization of speech patterns ("visible speech"; cf. Potter et al. 1947; Neppert and Pétursson 1986). The sonagraph offered analysis of a segment of sound either directly recorded into the machine (by microphone) or fed from tape or record player into a line input (mono). The Kay sonagraph became a standard tool in phonetics and, beginning in the 1960s, also for systematic and comparative musicology (see Graf 1972, 1976, 1980; Födermayr 1971; Rösing 1972). The following sonogram displays the same melisma of a female singer from Lebanon that has been under study above (Figs. 2, 3); the sonagram was produced with the (advanced) Kay model 7030 A that offered two exchangeable filter units with four different band filter settings (wide: 300 or 150 Hz; narrow: 45 or 10 Hz).[22]

The sonagram clearly shows the modulation and a concentration of spectral energy in bands corresponding to vocal formants (these bands cover ca. 600–800 Hz, 1200–1400 Hz, and from 3.4 to above 4 kHz). The results of the analogue sonagraphic analysis are well in line with the spectral and formant analysis obtained by digital signal processing (see Figs 2, 3 and 5).

[21] For the calculation of analogue low pass and band pass filter parameters, see Fant (1952), Küpfmüller (1968).

[22] The sonagram was produced, in spring 1978, in the lab of the Kommission für Schallforschung of the Austrian Academy of Sciences at Vienna where the present author was working for a period on invitation of Walter Graf, then head of the Kommission für Schallforschung and its lab. Fant (1952) developed a heterodyne filter that, different from the constant bandwidth filter employed in the sonagraph, offers continuously variable bandwidth along with the continuously variable center frequency.

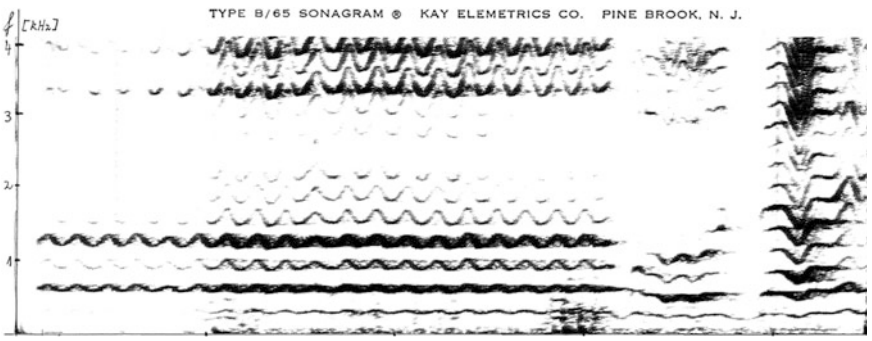


Fig. 14 *Abu Zeluf*, melisma, Sonagram 0–4 kHz, 4.8 s of the time signal

Since the analogue sonagraph employs a linear frequency scale, one could measure (manually peak-to-peak) the modulation (as in Fig. 14) at the *n*th partial and then calculate the frequency modulation deviation for the fundamental simply by dividing the shift found at the *n*th partial by *n*. Also, one could make use of both wide band and narrow band filters in the same probe and thereby overcome—at least to a certain degree—the limitations known from linear systems (cf. Küpfmüller 1968, Kap. IV), where the relation df dt ≥ 1 applies. In regard to band pass filters this means that the filter response time is $\tau = 2\pi/\Delta\omega = 1/\Delta f$, that is, the response takes the longer the narrower the pass band of the filter is chosen. Having the same probe processed by filters of different bandwidth, one could have the advantage of better temporal resolution with the 300/150 Hz filter, and improved frequency resolution by using the 45/10 Hz narrow band (cf. Schneider 1986).

In the 1970s, digital signal analysis based on the Discrete Fourier Transforms (DFT) making use of highly efficient algorithms such as FFT (see Randall 1987; DeFatta et al. 1988) was developing fast. Spectrum analyzers offering DFT/FFT became available though sampling rates, length of the transform for Fourier analysis and, hence, temporal and frequency resolution were still modest (cf. Randall 1987) mostly due to limits in memory needed for storage and processing of signals. In addition to narrow-band spectrum analyzers (such as the B and K 2031 and 2033 models that, by about 1980/81, were found in many labs), also a digital sonagraph (Sona-Graph DSP 5500, Kay Elemetrics) was constructed which, like the analogue model, allowed to measure the temporal evolution of spectra and to plot time/frequency representations as spectrograms but also had an option for analysis of two independent signal channels. This machine has been used in the study of transients in organ flue pipes as well as in recorders (Castellengo 1999).

By the end of the 1970s, the Synclavier (and by about 1980 the Synclavier II) became available that, besides its capabilities as a digital synthesizer and high resolution sampler (with sampling up to 100 kHz at 16 bit), offered also state-of-the-art FFT-based spectral analysis including 3D spectral plots and harmonic grid display to check the harmonicity of spectral components (cf. Beurmann and Schneider 1995; Schneider 1997b). Moreover, the Synclavier II offered automated

transcription, whereby sound was directly transformed into western staff notation. We have tried this option along with the spectral analysis and component frequency and level (dB) calculation on, for example, two-part pan pipe music of the 'Are'are (Salomon Islands, recordings made by Hugo Zemp; cf. Schneider 1997b, 393ff.).

By about 1990, powerful workstations (like SGI, SUN, NeXT) had become available, some with A/D and D/A conversion as well as DSP hardware onboard. Also, software for sound analysis and synthesis such as *sndan* (cf. Beauchamp 2007), *Spectro* (Gary Scavone and P. Cook, CCRMA/Stanford), *Sonogram* (Hiroshi Momose, UC Davis), *SuperVP* (IRCAM) as well as versatile platforms for programming (like CSound, CMusic, ESPS +) along with multi-functional math packages (Mathematica, MatLab) opened a completely new age especially for systematic musicologists as well as for composers and musicians interested in acoustics and psychoacoustics, sound synthesis, computer music, etc. [for an overview of tools and fields of application, see comprehensive books edited by Roads and Strawn 1996; Roads et al. 1997; Zölzer 2003; Beauchamp 2007 and the monograph Smith (2007) has supplied]. In regard to sound analysis, careful application of short-time Fourier transforms (STFT) allowed to study transients as well as inharmonic signals. By choosing a very low hop ratio (down to a few samples or even one sample) and, thus, a high percentage of overlap of frames that are processed (and in addition zero padding as needed, depending on the signal structure and transform length; see DeFatta et al. 1988), a quasi-continuous spectral documentation of transients in onsets as well as of modulation processes as observed in various musical instruments became accessible (for examples, see Schneider 1997b, 1998, 2000; Beurmann et al. 1998). Also, spectral envelope estimates by means of LPC or AR algorithms (cf. Marple 1987; Rodet and Schwarz 2007, Schneider and Mores, this volume) allowed close inspection of transient parts of sound. In addition to STFT, wavelet analysis has been employed as a high-resolution time/frequency representation of signals (cf. Mertins 1996, 1999). Since the transient portion of a signal contains most of the information, calculation of fractal dimensions and other methods known from dynamic system analysis (e.g., phase space and limit cycle analysis, Hopf bifurcation) help to identify processes characteristic of the transient part as distinct from the steady-state sound (cf. Bader 2002). One of the motives central to conduct such studies is to understand complex acoustic systems such as musical instruments from their sound as it is radiated from vibrating parts and surfaces. The sound patterns thus are likely to indicate patterns of vibration. Therefore, results obtained from the analysis of sound of real instruments must converge to a high degree with data from vibration analysis and modeling of instruments such as is done with, for example, the finite element and finite difference methodologies (FEM, FDM; see Bader 2005 for his model of the classical guitar and Lau et al. 2010 for a model of a swinging bell as compared to a real bell). In this respect, advanced sound analysis has gained an important role in musical acoustics as well as in areas of psychoacoustics where properties of sound in regard to perception are of interest. To be sure, most of the developments addressed in the present survey (covering, in the main, the era of Helmholtz to

the time when computers and digital signal processing had become widespread) happened within a span of but a hundred years.

4 Summing Up: Continuity and Change

Modern times up to the present seem to be governed by unprecedented speed and ever increasing rates of acceleration. Travel within the 20th century has become faster and faster (from simple automobiles to jet planes), and technological developments have followed one another in ever shorter cycles. A field where the rate of change perhaps is most obvious is information and communication systems including the hardware involved. The computational power of a mainframe of the 1960s or even the 1970s appears small when compared to an average PC in use now. Musical acoustics and sound research have clearly benefited from rapid developments in electronics and in computer technology. Because of the progress in technology and in actual research, the number of publications relating to sound analysis and synthesis is vast. In such a situation, there is a risk that significant portions of previous knowledge will fall into oblivion notwithstanding massive data storage and archiving of relevant publications in digital formats.

Of course, there is always the possibility to look back on earlier achievements in research and technology as has been done for acoustics. In regard to science, historical accounts, rather than being conceived as a plain "narrative" of past endeavours and achievements, must try to reveal topics central to research in a certain field along with shedding light on issues in approach and methodology (including basic mathematical and physical background; see, e.g., Cannon and Dostrovsky 1982; Beyer 1999).

While history conceived as "narrative" can hardly be written without some concept of continuity, Canguilhem (1966/1994) has stressed the importance of discontinuity in science as a driving force for progress. One could also point to the role of changing 'paradigms' that supplant or replace one another to underpin the dynamics of science. Though discontinuity is a fact that can be observed in many instances in regard to theories, methodology, and also practical matters, there is also persistence of previous knowledge as well as of problems that are unsolved yet and hence need to be investigated. In order to find improved experimental designs and possibly more appropriate solutions, one has to know about previous research and its outcome (including valid results as well as failures).

The present article and some other publications I have contributed making use of sound analysis and synthesis in one way or another (e.g., Schneider 1997a, b, 2001, 2011) were conceived with the intent to connect the present to the past, and to outline certain developments in regard to both change and continuity in research. While change is often induced by new tools (technical as well as conceptual and methodological) that become available for research, leading to fresh perspectives, in quite many areas a certain continuity can also be observed either because problems resist to be solved completely, or because previous research is

acknowledged as still worth noticing and perhaps calls for a continuation of efforts.

References

Albersheim, G. (1939). *Zur Psychologie der Ton- und Klangeigenschaften unter Berücksichtigung der ‚Zweikomponententheorie' und der Vokalsystematik*. Straßburg: Heitz.

Arfib, D., Keiler, F., & Zölzer, U. (2002). Source-Filter Processing. In U. Zölzer (Ed.), *DAFX. Digital Audio Effects* (pp. 299–372). Chichester: Wiley.

Backhaus, H. (1932). Über die Bedeutung der Ausgleichsvorgänge in der Musik. *Zeitschrift für technische Physik, 13*, 31–46.

Backus, J. (1963). Acoustical investigations of the clarinet. *Sound, 2*, 22–25.

Bader, R. (2002). *Fraktale Dimensionen, Informationsstrukturen und Mikrorhythmik der Einschwingvorgänge von Musikinstrumenten*. Ph.D. dissertation, University of Hamburg.

Bader, R. (2005). *Computational Mechanics of the Classical Guitar*. Berlin etc.: Springer.

Bader, R. (2011). Buddhism, animism, and entertainment in Cambodian melismatic chanting *smot*—history and tonal system. In A. Schneider & A. von Ruschkowski (Eds.), *Systematic Musicology: Empirical and theoretical studies* (pp. 283–305). P. Lang: Frankfurt/M.

Battey, B. (2004). Bézier spline modeling of pitch-continuous melodic expression and ornamentation. *Computer Music Journal, 28*, 25–39.

Beauchamp, J. (2007). Analysis and Synthesis of musical instrument sounds. In J. Beauchamp (Ed.), *Analysis, Synthesis, and Perception of musical sounds* (pp. 1–89). New York: Springer.

Beurmann, A., & Schneider, A. (1995). Zur akustischen Untersuchung von Volksmusikinstrumenten. *Studia instrumentorum musicae popularis* (Vol. X, pp. 113–121). Stockholm: Musikmuseet.

Beurmann, A., & Schneider, A. (2008). Acoustics of the harpsichord: a case study. In A. Schneider (Ed.), *Systematic and Comparative Musicology: Concepts, methods, findings* (pp. 241–263). P. Lang: Frankfurt/M.

Beurmann, A., & Schneider, A. (2009). Acoustics and sound of the harpsichord: another study. In R. Bader (Ed.), *Musical Acoustics, Neurocognition and Psychology of Music* (pp. 57–72). Frankfurt/M: P. Lang.

Beurmann, A., Schneider, A., & Lauer, E. (1998). Klanguntersuchungen an der Arp-Schnitger-Orgel zu St. Jacobi, Hamburg. *Systematische Musikwissenschaft, 6*, 151–187.

Beyer, R. (1999). *Sounds of our time. Two hundred years of acoustics*. New York: Springer.

Boersma, P. (1993). Accurate short-term Analysis of the fundamental frequency and the harmonic-to-noise ratio of a sampled sound. *IFA Proceedings, 17*, 97–110.

Boersma, P., Weenink D. (2011). *Praat. Doing Phonetics by Computer*. Amsterdam: University of Amsterdam, Institute of Phonetic Sciences. (Praat version 5323).

Bozkurt, B. (2008). An automatic Pitch analysis method for Turkish Maquam music. *Journal of New Music Research, 37*, 1–13.

Bozkurt, B., Yarman, O., Karaosmanoğlu, K., & Akkoş, C. (2009). Weighing diverse theoretical Models on Turkish *Maquam* music against pitch measurements: a comparison of peaks automatically derived from frequency histograms with proposed scale tones. *Journal of New Music Research, 38*, 45–70.

Bregman, A. (1990). *Auditory Scene Analysis*. Cambridge: MIT Press.

Brown, J. (2007). Fundamental frequency tracking and applications to musical signal analysis. In J. Beauchamp (Ed.), *Analysis, Synthesis, and Perception of musical sounds* (pp. 90–121). New York: Springer.

Brown, J., & Puckette, M. (1993). A high resolution fundamental frequency determination based on phase changes of the Fourier transform. *Journal of Acoustical Society of America, 94*, 662–667.

Cañadas Quesada, F. J., Ruiz Reyes, N., Vera Candeas, P., Carabias, J. J., & Maldonado, S. (2010). A multiple-F0 estimation approach based on Gaussian spectral modeling for polyphonic music transcription. *Journal of New Music Research, 39*, 93–107.

Canguilhem, G. (1966/1994). L'Objet de l'histoire des sciences (= lecture, Montréal 1966). In G. Canguilhem (Ed.), *Études d'histoire et de philosophie des sciences* (pp. 9–23). 7e ed. Paris: Vrin 1994.

Cannon, J., & Dostrovsky, S. (1982). *The Evolution of Dynamics: Vibration theory from 1687 to 1742*. New York: Springer.

Cariani, P. (2004). A temporal Model for Pitch multiplicity and tonal Consonance. *Proceedings of International Conference on Music Perception and Cognition* (IMPC), Evanston, 310–314.

Castellengo, M. (1999). Analysis of initial transients in flute like instruments. *Acustica, 85*, 387–400.

Chladni, E. F. F. (1805/1830). *Die Akustik*. Leipzig: Breitkopf & Haertel (2nd ed. 1830).

Chladni, E. F. F. (1817). *Neue Beyträge zur Akustik*. Leipzig: Breitkopf & Haertel.

Cohen, H. F. (1984). *Quantifying Music. The Science of music at the first stage of the scientific revolution, 1580–1650*. Dordrecht, Boston: D. Reidel.

Cohen, L. (1995). *Time-Frequency analysis*. Upper Saddle River: Prentice Hall.

Dahlback, K. (1958). *New Methods in vocal folk music research*. Oslo: University Press.

Dalmont, J. P., Gilbert, J., & Kergomard, J. (2000). Reed Instruments, from small to large amplitude periodic oscillations and the Helmholtz motion analogy. *Acustica, 86*, 671–684.

de Boer, E. (1976). On the "Residue" and Auditory Pitch Perception. In W. D. Keidel & W. D. Neff (Eds.), *Handbook of Sensory Physiology* (pp. 479–583). Berlin: Springer.

de Cheiveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111*, 1917–1930.

de Cheveigné, A. (2005). Pitch perception models. In Chr. Plack, A. Oxenham, R. Fay, A. Popper (Eds.). *Pitch. Neural Coding and Perception* (pp. 169–233). New York: Springer.

DeFatta, D., Lucas, J., & Hodgkiss, W. (1988). *Digital Signal Processing. A system design approach*. New York: Wiley.

Edison, T. (1888). The perfected Phonograph. *The North American Review, 146*(379), 641–650.

Edskes, C., & Vogel, H. (2009). *Arp Schnitger und sein Werk*. Bremen: Hauschild.

Ehret, G. (1997). The auditory midbrain, a "shunting yard" of acoustical information processing. In G. Ehret & R. Romand (Eds.), *The Central Auditory System* (pp. 259–316). Oxford: Oxford University Press.

Elschek, O. (1979). *Melographische Interpretationscharakteristika von Flötenmusik*. *Studia instr. mus. pop*. VI (pp. 43–58), Stockholm: Musikhist. Museet.

Elschek, O. (2006). *Fujara. The Slovak Queen of European Flutes*. Bratislava: Hudobné Centrum.

Fant, G. (1952). *The Heterodyne Filter*. Göteborg/Stockholm: Elander/Lindstahl.

Fant, G. (1960). *Acoustic theory of speech production. With calculations based on x-ray studies of Russian articulations*. The Hague: Mouton.

Feld, S. (1990). *Sound and Sentiment: Birds, weeping, poetics, and song in Kaluli expression* (2nd ed.). Philadelpia: University of Pennsylvania Press.

Feld, S., Kirgegaard, A. (2010). Entangled Complicities in the prehistory of 'World Music': Poul Rovsing Olsen and Jean Jenkins encounter Brian Eno and David Byrne in the Bush of Ghosts. *Popular Musicology Online*. www.popular-musicology-online.com/issues/04/feld.html. (retrieved 30th of August, 2012).

Filip, M. (1969). Envelope periodicity detection. *Journal of the Acoustical Society of America, 45*, 719–732.

Filip, M. (1970). Frekvenčné merania a tónová sústava [Frequency measurement and tonal system] (pp. 50–85), *Nové Cesty Hudbe, 2,* Praha.

Filip, M. (1978). Acoustic measurements as auxiliary methods in ethnomusicology. *Musicologica Slovaca, 7*, 77–87.

Födermayr, F. (1971). *Zur gesanglichen Stimmgebung in der außereuropäischen Musik* (Vol. 2). Wien: Stiglmayr.

Fricke, J. (1993). Die Wechselwirkung von Mensch und Musik im Zusammenspiel von Physik und Physiologie. In B. Enders & St. Hanheide (Eds.), *Neue Musiktechnologie* (pp. 169–196). London: Schott.

Graf, W. (1972). Musikalische Klangforschung. *Acta Musicologica, 44*, 31–78.

Graf, W. (1976). Zum Klang der Stainer-Geigen. In *Festschrift Walter Senn zum 70. Geburtstag* (pp. 98–117), hrsg. von E. Egg u.a., München: Katzbichler.

Graf, W. (1980). *Vergleichende Musikwissenschaft. Ausgewählte Aufsätze*. Hrsg. von F. Födermayr. Stiglmayr: Wien-Föhrenau.

Grützmacher, M. (1927). Eine neue Methode der Klanganalyse. *Elektrische Nachrichtentechnik (ENT), 4*(12), 533–545.

Grützmacher, M., & Lottermoser, W. (1937). Über ein Verfahren zur trägheitsfreien Aufzeichnung von Melodiekurven. *Akustische Zeitschrift, 2*, 242–248.

Hartmann, W. (1998). *Signals, Sound, and Sensation*. New York: AIP/Springer.

Hekland, F. (2001). *Automatic Music transcription using autoregressive frequency estimation.* Project paper. Toulouse: ENSEEIHT.

Helmholtz, H. (1863/1870/1896). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Braunschweig: Vieweg (3rd ed. 1870, 5th ed. 1896).

Helmholtz, H. von (1857/1896). Über die physiologischen Ursachen der musikalischen Harmonien (Vortrag, Bonn 1857). In H. von Helmholtz (Ed.), *Vorträge und Reden* (pp. 119–155). 4. Aufl. Braunschweig: Vieweg 1896, Bd 1.

Hermann, L. (1889). Phonophotographische Untersuchungen. [Pflügers]. *Archiv für die gesamte Physiologie, 45*, 582–592.

Hermann, L. (1891). Bemerkungen zur Vocalfrage. *Archiv für die gesamte Physiologie, 48*, 181–194.

Hermann, L. (1893). Phonophotographische Untersuchungen. IV: Untersuchungen mittels des neuen Edison'schen Phonographen. *Archiv für die gesamte Physiologie, 53*, 1–51.

Hermann, L. (1894). Phonophotographische Untersuchungen. VI. Nachtrag zur Untersuchung der Vocalcurven. (Nach Versuchen in Gemeinschaft mit Dr. F. Matthias und stud. med. Alfred Ehrhardt.). *Archiv für die gesamte Physiologie, 58*, 264–279.

Hermann, L. (1911). Neue Beiträge zur Lehre von den Vokalen und ihrer Entstehung. *Archiv für die gesamte Physiologie, 141*, 1–62.

Hermann, L. (und H. Hirschfeld) (1895). Weitere Untersuchungen über das Wesen der Vocale. *Archiv für die gesamte Physiologie, 61*, 169–209.

Herrmann, E. (1908). Über die Klangfarbe einiger Orchesterinstrumente und ihre Analyse. Diss. Königsberg 1908. In *Beiträge zur Physiologie und Pathologie. Festschrift zum 70. Geburtstag L. Hermann …gewidmet* (pp. 59–105), hrsg. von O. Weiss, Stuttgart: Union Deutsche Verlagsges.

Hesse, H. P. (1972). *Die Wahrnehmung von Tonhöhe und Klangfarbe als Problem der Hörtheorie*. Köln: Volk/Gerig.

Ingle, V., & Proakis, J. (2000). *Digital Signal Processing using MATLAB*. Pacific Grove ets.: Brooks/Cole.

Keidel, W. D. (Ed.). (1975). *Physiologie des Gehörs. Akustische Informationsverarbeitung*. Stuttgart: Thieme.

Kergomard, J., Ollivier, S., & Gilbert, J. (2000). Calculation of the Spectrum of self-sustained oscillators using a variable truncation method: application to cylindrical reed instruments. *Acustica, 86*, 685–703.

Klapuri, A. (2004). Automatic music transcription as we know it today. *Journal of New Music Research, 33*, 269–282.

Klapuri, A., & Davy, M. (Eds.). (2006). *Signal Processing Methods for music transcription*. New York: Springer.

Kreichgauer, A. (1932). *Über Maßbestimmungen freier Intonationen*. Berlin: Phil. Diss.
Krueger, F. (1907). Beziehungen der experimentellen Phonetik zur Psychologie. *Bericht über den II. Kongress für exp. Psychol. Würzburg 1906* (pp. 58–122). Leipzig: Barth.
Küpfmüller, K. (1968). *Die Systemtheorie der elektrischen Nachrichtenübertragung*. 3. Aufl. Stuttgart: Hirzel.
Lau, B., Bader, R., Schneider, A., & Wriggers, P. (2010). Finite-Element transient calculation of a bell struck by its clapper. In R. Bader, Chr. Neuhaus, & U. Morgenstern (Eds.), *Concepts, Experiments, and Fieldwork: Studies in Systematic Musicology and Ethnomusicology* (pp. 137–156). Frankfurt/M.: P. Lang.
Licklider, J. C. (1951). A duplex theory of pitch perception. *Experientia, 7*, 128–134.
Licklider, J. C. (1956). Auditory Frequency Analysis. In C. Cherry (Ed.), *Information theory* (pp. 253–268). London: Butterworth.
Lottermoser, W. (1976/1977). Frequenzschwankungen bei musikalischen Klängen. *Acustica,* 36, 138–146.
Marple, S. L. (1987). *Digital Spectral Analysis*. Englewood Cliffs, NJ: Prentice Hall.
McAulay, R., Quatieri, T. (1986). Speech Analysis/Synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing* Vol. ASSP-34, *34*(4), 744–754.
Meddis, R., & O'Mard, L. (1997). A unitary Model of pitch perception. *Journal of the Acoustical Society of America, 102*, 1811–1820.
Meddis, R., & O'Mard, L. (2006). Virtual pitch in a computational physiological model. *Journal of the Acoustical Society of America, 120*, 3861–3869.
Meissner, G. (1907). Klangaufnahmen an Blasinstrumenten, eine Grundlage für das Verständnis der menschlichen Stimme. Nachgelassenes Manuscript …hrsg. durch Richard Wachsmuth. [Pflügers]. *Archiv für die gesamte Physiologie, 116*, 543–599.
Mertens, P. H. (1975). *Die Schumannschen Klangfarbengesetze und ihre Bedeutung für die Übertragung von Sprache und Musik*. Frankfurt/M: Bochinsky.
Mertins, A. (1996). *Signaltheorie*. Stuttgart: Teubner.
Mertins, A. (1999). *Signal Analysis*. Chichester: Wiley.
Metfessel, M. (1928). *Phonophotography in Folk Music. American Negro songs in new notation*. Chapel Hill: University of North Carolina Press.
Meyer, E., & Guicking, D. (1974). *Schwingungslehre*. Braunschweig: Vieweg.
Mores, R. (2010). Vowel quality in violin sounds. In R. Bader, Chr. Neuhaus, & U. Morgenstern (Eds.), *Concepts, Experiments, and Fieldwork: Studies in Systematic Musicology and Ethnomusicology* (pp. 113–135). Frankfurt/M.: P. Lang.
Nagel, G. (2007). Sprengstoff- und Fusionsforschung an der Berliner Universität: Erich Schumann und das II. Physikalische Institut. In R. Karlsch & H. Petermann (Eds.), *Für und Wider 'Hitlers Bombe'. Studien zur Atomforschung in Deutschland* (pp. 229–260). Münster: Waxmann.
Nelken, I. (2002). Feature detection by the auditory cortex. In D. Oertel, R. Fay, & A. Popper (Eds.), *Integrative Functions in the Mammalian auditory Pathway* (pp. 358–416). New York: Springer.
Neppert, J., & Pétursson, M. (1986). *Elemente einer akustischen Phonetik* (2nd ed.). Hamburg: Buske.
Obata, J., & Kobayashi, R. (1937). A direct-reading pitch recorder and its application to music and speech. *Journal of the Acoustical Society of America, 9*, 156–161.
Opelt, F. W. (1852). *Allgemeine Theorie der Musik auf den Rhythmus der Klangwellenpulse gegründet*. Leipzig: Barth.
Owen, T. (1974). Applying the Melograph to "Parker's Mood". *Selected Reports in Ethnomusicology* (pp. 167–175), Vol. II, 1. Los Angeles: UCLA.
Paiva, R. P., Mendes, T., & Cardoso, A. (2008). From Pitches to notes: creation and segmentation of pitch tracks for melody detection in polyphonic audio. *Journal of New Music Research, 37*, 185–205.

Panconcelli-Calzia, G. (1941/1994). *Geschichtszahlen der Phonetik. 3000 Jahre Phonetik*. Hamburg: Hansischer Gildenverlag 1941; Repr. Amsterdam: Benjamins 1994.

Pipping, H. (1890). Zur Klangfarbe der gesungenen Vocale. *Zeitschrift für Biologie* 27, 1ff., 433ff.

Pipping, H. (1894). *Über die Theorie der Vokale*. Helsingfors: Soc. Litterariae Fennicae (=Acta Soc. Scient. Fenn. XX, no. 11).

Potter, R., Kopp, G., & Green, H. (1947). *Visible Speech*. New York: van Nostrand.

Randall, R. B. (1987). *Frequency Analysis* (3rd ed.). Naerum: Bruel & Kjaer.

Reinecke, H. P. (2003). Hermann von Helmholtz, Carl Stumpf und die Folgen. Von der Akustik zur Tonpsychologie. Über ein Kapitel Wissenschaftsgeschichte in Berlin. In M. Kaiser-El-Safti & M. Ballod (Eds.), *Musik und Sprache. Zur Phänomenologie von Carl Stumpf* (pp. 185–197). Würzburg: Königshausen & Neumann.

Reuter, C. (1995). *Der Einschwingvorgang nichtperkussiver Musikinstrumente*. Frankfurt/M: P. Lang.

Reuter, C. (1996). *Die auditive Diskrimination von Orchesterinstrumenten. Verschmelzung und und Heraushörbarkeit von Instrumentalklangfarben im Ensemblespiel*. Frankfurt/M: P. Lang.

Roads, C., Strawn, J., et al. (1996). *The Computer Music Tutorial*. Cambridge: MIT Press.

Roads, C., Pope, S. T., Piccialli, A., & de Poli, G. (Eds.). (1997). *Musical Signal Processing*. Lisse etc.: Swets & Zeitlinger.

Rodet, X., & Schwarz, D. (2007). Spectral Envelopes and additive + residual analysis/synthesis. In J. Beauchamp (Ed.), *Analysis, Synthesis, and Perception of musical sounds* (pp. 174–227). New York: Springer.

Rösing, H. (1972). *Die Bedeutung der Klangfarbe in traditioneller und elektronischer Musik. Eine sonagraphische Untersuchung*. München: Katzbichler.

Sauveur, J. (1701). Système général des intervalles des sons, et son application à tous les systèmes et à tous les instrumens de musique. *Histoire de L'Académie Royale des Sciences. Année 1701, Mémoires de Mathematique & de Physique*, 297–364.

Schneider, A. (1986). Tonsystem und Intonation. *Hamburger Jahrbuch für Musikwissenschaft, 9*, 153–199.

Schneider, A. (1987). Musik, Sound, Sprache, Schrift: Transkription und Notation in der Vergleichenden Musikwissenschaft und Musikethnologie. *Zeitschrift für Semiotik, 9*, 317–343.

Schneider, A. (1997a). "Verschmelzung" Tonal Fusion, and Consonance: Carl Stumpf revisited. In M. Leman (Ed.), *Music, Gestalt, and Computing. Studies in Cognitive and Systematic Musicology* (pp. 117–143). Berlin: Springer.

Schneider, A. (1997b). *Tonhöhe–Skala–Klang. Akustische, tonometrische und psychoakustische Studien auf vergleichender Grundlage*. Bonn: Orpheus-Verlag für syst. Musikwiss.

Schneider, A. (1998). Klanganalysen bei Aerophonen der Volksmusik. In F. Födermary & L. Burlas (Eds.), *Ethnologische, Historische und Systematische Musikwissenschaft. Oskár Elschek zum 65. Geburtstag* (pp. 51–80). Bratislava: ASCO

Schneider, A. (2000). Inharmonic Sounds: Implications as to «Pitch», «Timbre» and «Consonance». *Journal of New Music Research, 29*, 275–301.

Schneider, A. (2001). Sound, Pitch, and Scale: from 'tone measurements' to sonological analysis in ethnomusicology. *Ethnomusicology, 45*, 489–519.

Schneider, A. (2011). Music Theory: Speculation, Reasoning, Experience. A Perspective from Systematic Musicology. In T. Janz & J. Ph. Sprick (Eds.), *Musiktheorie | Musikwissenschaft. Geschichte—Methoden—Perspektiven* (pp. 53–97). New York: Olms.

Schneider, A., & Frieler, L. (2009). Perception of harmonic and inharmonic sounds: results from ear models. In S. Ystad, R. Kronland-Martinet, & K. Jensen (Eds.), *Computer Music Modeling and Retrieval. Genesis of meaning in sound and music* (pp. 18–44). Berlin: Springer.

Schneider, A., von Busch, R., & Schmidt, L. (2001). Klanganalysen an Arp Schnitger-Orgeln. In N. Ristow, W. Sandberger, & D. Schröder (Eds.), *«Critica Musica.» Studien zum 17. und 18. Jahrhundert* (pp. 247–270). Stuttgart: Metzler.

Schumann, K. E. (1925). *Akustik*. Breslau: Hirt.

Scripture, E. W. (1906). *Research in Experimental Phonetics. The Study of Speech Curves*. Washington, D.C.: Carnegie Institution.
Scripture, E. W. (1927). *Anwendung der graphischen Methode auf Sprache und Gesang*. Leipzig: J. Barth.
Seeger, C. (1951). An instantaneous music notator. *Journal of the International Folk Music Council, 3*, 103–106.
Slawson, W. (1985). *Sound Color*. Berkeley: University of California Press.
Smith, J. O. (2007). Introduction to digital filters with audio applications. Stanford: CCRMA (online book available at https://ccrma.stnford.edu/~jos/filters/).
Stern, W. (1902). Der Tonvariator. *Zeitschrift für Psychologie, 30*, 422–432.
Stevens, S. S. (1934). The Volume and Intensity of Sounds. *The American Journal of Psychology, 46*, 397–408.
Stumpf, C. (1890). *Tonpsychologie*. Bd II. Leipzig: J. Barth.
Stumpf, C. (1898). *Konsonanz und Dissonanz*. Leipzig: J. Barth.
Stumpf, C. (1926). *Die Sprachlaute Experimentell-phonetische Untersuchungen nebst einem Anhang über Instrumentalklänge*. Berlin: Springer.
Sundberg, J. (1997). *Die Wissenschaft von der Singstimme*. Bonn: Orpheus-Verlag für syst. Musikwiss.
Terhardt, E. (1998). *Akustische Kommunikation*. Berlin: Springer.
Tjernlund, P., Sundberg, J., Fransson, F. (1972). Grundfrequenzmessungen an schwedischen Kernspaltflöten. *Studia instr. mus. pop*. II, Stockholm, (pp. 77–96).
Tramo, M., Cariani, P., Delgutte, B., & Braida, L. (2001). Neurobiological Foundations for the theory of harmony in western tonal music. *Annals of the New York Academy of Sciences, 930*, 92–116.
Trendelenburg, F., Thienhaus, E., & Franz, E. (1936). Klangeinsätze an der Orgel. *Akustische Zeitschrift, 1*, 59–76.
Vierling, O. (1936). Der Formantbegriff. *Annalen der Physik,* 5, Folge, Bd 26, 219–232.
Voigt, W. (1975). *Untersuchungen zur Formantbildung in Klängen von Fagott und Dulzianen*. Regensburg: Bosse.
von Qvanten, E. (1875). Einige Bemerkungen zur Helmholtz'schen Vocallehre. [Poggendorfs]. *Annalen der Physik, 230*, 522–552.
Weber, G. (1817/1824/1830). *Versuch einer geordneten Theorie der Tonsetzkunst*. Bd 1, Mainz: Schott (2nd ed. 1824, 3rd ed. 1830).
Wiener, N. (1961). *Cybernetics or control and communication in the animal and in the machine* (2nd ed.). New York: MIT Press.
Willis, R. (1830). On the vowel sounds, and on the reed organ-pipes. *Transactions of the Cambridge Philosophical Society*, *3*, 229–268.
Winckel, F. (1960). *Phänomene des musikalischen Hörens*. Berlin: Hesse.
Zölzer, U. (Ed.). (2003). *DAFX. Digital Audio Effects*. Chichester: Wiley.

Quantal Elements in Musical Experience

Rolf Inge Godøy

1 Introduction

The aim of this chapter is to present a model for understanding unit formation, what we prefer to call *chunking*, at short-term timescales in musical experience, typically in the duration range of approximately 0.5–5 s. The idea is that at these short-term timescales, chunks of sound and associated body motion are conceived and perceived holistically; hence demonstrate what may be called *quantal elements in musical experience*. Very many salient musical features for identifying style, motion, and affect, can be found at such short-term timescales [and sometimes at even shorter timescales as suggested by Gjerdingen and Perrott (2008)]. A better understanding of such quantal elements in musical experience could be useful in the fields of music perception, music analysis, and music information retrieval, as well as in various practical artistic and educational contexts.

Needless to say, from an acoustical point of view, music is something that unfolds linearly in time. This is the case for basic sonic phenomena such as periodicity and frequency, and for more composite phenomena such as timbre, loudness, and pitch, as well as event-level (or note-level) phenomena such as rhythm, texture, and melody. Yet it is also well known from psychoacoustic and music perception research that sequentially occurring elements contribute to holistic, and in a sense 'atemporal', perception of features: Sequentially occurring attack transients and fluctuations in the course of sounds are essential for holistic timbral experience and categorization; sequentially occurring variations in pitch, loudness, and timbre are essential for experiences of expressivity; sequentially occurring tone events are essential for experiences and recognition of motives, textures, motion, and affect, etc. In short, it seems that very many musical features are dependent on some kind of transformation of sequentially occurring elements at short-term timescales to quantal percepts.

R. I. Godøy (✉)
Department of Musicology, University of Oslo, Oslo, Norway
e-mail: r.i.godoy@imv.uio.no

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_4,

Although we have had more than a century of significant research in gestalt theory on unit formation in music, as well as important advances in research on auditory perception and human motion during the last decades, the perception and cognition of such short-term chunks of sound and associated body motion are to my knowledge still not well-researched topics. Drawing on past and present relevant research, the main hypothesis of this chapter is that below the threshold of the short-term timescale (i.e., very approximately in the 0.5–5 s range), there is a qualitative difference in our perception and cognition of continuous sound and continuous body motion, in that what is occurring sequentially is subjectively perceived and conceived as simultaneous, or "in a now", to borrow the expression used by Edmund Husserl more than a century ago (Husserl 1991; Godøy 2010).

The main challenge of this chapter is then to present some kind of understanding of how these quantal elements work in musical experience, in view of the long-term goal of developing a model applicable to sound and motion data. On the way to a sketch of such a model, we shall first have a look at what are the contents and criteria for such quantal elements, as well as present evidence suggesting that these quantal elements are based on various perceptual and motor constraints.

2 Duration Thresholds

One basic principle in both auditory and motion perception, as well as in human motion control, is that of quite distinct qualitative experiences at different timescales, something that has consequences for quantal elements in musical experience. There are the well-known thresholds in both audition and vision at approximately 20 Hz between what may be called the *sub-sonic* and *sonic* domains in audition, and between still pictures and animation in vision, the so-called *flicker-fusion threshold*. Within these domains, there are furthermore thresholds of duration for various features to be perceived, such as that of minimum duration for the perception of pitch, timbre, and location of sound sources, as well as for onset simultaneity and order; however, the time values involved here seem to vary somewhat depending on content and context (Pöppel 1989; Moore 1995). Interestingly, it seems that rather composite and style-related timbral-textural features may be perceived at duration thresholds as low as 250 ms, a finding suggesting that the perception of rich timbral-textural content is indeed both quite fast and robust (Gjerdingen and Perrott 2008).

There are several important auditory features found at the sub-sonic timescale (i.e., below the 20 Hz limit), meaning features that typically take more than 50 ms to unfold. These features include dynamic, timbral, and pitch-related envelopes for sound, ranging in duration from that of various audible fluctuations in pitch, loudness, or timbre, to more composite features such as rhythmical and textural patterns, melodic motives, and figurations. The effect of dynamic envelopes on the holistic perception of sounds have been studied for several decades, such as in the classic investigations of 'plucked' versus 'bowed' sounds (Cutting 1982). We now

have a fairly large collection of research and feature descriptors that try to pinpoint various transients and fluctuations in the sub-sonic range (see e.g., Peeters et al. 2011 for an overview).

Interestingly, envelopes in the sub-sonic range were extensively studied from a subjective perceptual point of view by Pierre Schaeffer and co-workers in the 1950s and 1960s, resulting in what was called the *typology of sonic objects* (Schaeffer 1966; Chion 1983). From our present point of view, an essential feature of this typology is that the sound envelope categories may be linked with body motion categories (Godøy 2006), hence, providing a basis for an embodied and multisensory scheme for understanding quantal elements in musical experience. In this typology, there are three main categories for the overall dynamic envelopes:

- *Sustained*, meaning basically a prolonged and steady sound, reflecting corresponding continuous effort and attention.
- *Impulsive*, meaning a short burst of energy such as in a sound made by hitting or kicking, followed by a longer or shorter decay.
- *Iterative*, meaning a rapid series of impulses such as in a tremolo, a trill, or a vibrato.

Furthermore, it was suggested that there are categorical thresholds between these dynamical envelopes, dependent on the variables of *duration* and *rate*: If the duration of a sustained sound is progressively shortened, there will sooner or later be a transition to the category of an impulsive sound, and conversely, if the duration of an impulsive sound is progressively extended, there will sooner or later be a transition to a sustained sound. Similarly, if the rate of impulsive sounds is increased, this will sooner or later turn into a continuous iterative sound, and conversely, if the rate of onsets in an iterative sound is decreased, this will sooner or later turn into a series of impulsive sounds. Similar time-dependent categorical thresholds for other sonic features were suggested by Schaeffer and co-workers, such as between so-called *grain* (fast fluctuations in the sound) and *allure* (slower fluctuations in the sound).

We find similar qualitative transitions in human motion control, here often referred to as *phase-transitions* (Haken et al. 1985), for instance in the transitions between human walking and running, however details of the underlying biomechanical and/or neurophysiological factors seem not yet to be fully understood. Such categorical changes on the basis of continuous changes in variables (speed, rate, amplitude), contribute to constraint-based quantal elements in music (more on this below). However, in some cases it may be difficult to specify exactly where changes occur, in particular between motion and rest, for the simple reason that a living human body is never completely still, hence the need to define some threshold values for stillness here (Hogan and Sternad 2007). Yet, what emerges from looking at different timescales and qualitative categories in human motion, is a striking similarity with qualitative categories in sonic features, and thus yet another reason for considering sound and motion together in music (Godøy and Leman 2010).

On a slightly longer timescale, on what we could collectively call the *meso-level* timescale, we find some other qualitative thresholds that may all be qualified

as manifesting holistically conceived and perceived events, as opposed to the more continuous features mentioned above. It could be useful again to take Pierre Schaeffer's ideas of the sonic object as point of departure, and also see if more recent research can shed light on features at this timescale. Typically, the sonic object will be in the duration range of 0.5–5 s, however this also depends on content and context. The main idea is that at this meso-level timescale we can perceive a number of salient musical features such as overall dynamic envelope and timbral evolution, as well rhythmical, textural, and melodic patterns, and importantly, the associated body motion patterns. This timescale is probably also the most salient in terms of determining style/genre and affect, e.g., Gjerdingen and Perrott (2008) suggest a 3,000 ms duration as fairly reliable in this respect.

The focus on the sonic object was initially motivated by pragmatic concerns in the early days of the *musique concrete*. Before the advent of the tape recorder, one way to mix sounds was to engrave a closed groove ("sillon fermé") on a phonograph disc, what we today would call a loop, and in the composition process have a number of assistants activate the playback of such sound fragments so that they could be feed into a mix. However, listening to these looped sound fragments, Schaeffer and co-workers discovered that their perceptions changed, that they started to consider the overall dynamic and timbral features of the fragments, and from this practice, more extensive theories of sonic objects emerged (Schaeffer 1966; Chion 1983). Various ideas of gestalt and phenomenological theory were combined with these initially pragmatic experiences of sonic objects, with the main conclusion emerging that the meso-level timescale was the most salient timescale, and that it also allowed for useful qualifications of both the overall shape and the internal, more fine-grained, features of the sonic object.

For timescales significantly longer than this meso-level timescale, it could be convenient to use the term *macro-level* timescale. This timescale is typically on the level of whole sections, tunes, and/or works, and will consist of several meso-level timescale fragments in succession.

In summary, we can think of three main feature timescales for both sound and body motion here:

- *Micro*, meaning the sonic (the above 20 Hz) timescale of basically continuous features, i.e., pitch, stationary timbre, and loudness, as well as fast sub-sonic rate fluctuations such as tremolos, trills, and timbral fluctuations in the sound, and corresponding continuous, smooth body motion and/or very rapid back-and-forth, shaking, or trembling body motion.
- *Meso*, the 0.5–5 s timescale, typically manifesting features such as rhythmical and textural patterns, motives, melodic fragments, modality, tone semantics, sense of motion, affect, and various expressive elements. This timescale is the basis for quantal elements in musical experience.
- *Macro*, typically concatenations of meso-level chunks, but also entailing sensations of longer, and often hierarchical, metrical schemes, e.g., the very often occurring 4 + 4, 8 + 8, etc., measure schemes in Western music.

Clearly, these different feature timescales are concurrent, yet also interdependent: Micro-level features are embedded in meso-level chunks, meso-level chunks are embedded in macro-level segments, and conversely, macro-level segments owe their content to the meso-level chunk features, and these in turn to the micro-level features. This is why Pierre Schaeffer stated that for any sonic object we always have the duality of a larger-scale *context* of the sonic object and an internal *contexture* of the sonic object, as well as "the two infinities" of music, meaning that in listening to musical sound we can intentionally zoom in and out (Schaeffer 1966, 279), similar to Husserl's idea of being able to zoom in and out of any musical excerpt in our memory (Husserl 1991, 49–50).

We may look at these different timescale thresholds from micro to macro as simply reflecting different durations to unfold, e.g., a waltz pattern (depending on the tempo) takes anything from one to a couple of seconds to manifest itself, hence is typically at the meso timescale, whereas e.g., an eight measure pattern of a waltz takes proportionally longer to manifest itself, hence is at the macro timescale. Yet, the subjective experience in our minds and bodies of these different timescale features may not be equivalent to their temporal unfolding, as there may be compression (retrospective accumulation, but also prospective, anticipatory compression) and categorization at work, hence we need to look at some features of continuity versus discontinuity in perception and cognition.

3 Continuity and Discontinuity

The transformation of continuous phenomena, such as sound and motion, into more atemporal and categorical entities is partly a matter of physical attributes, e.g., as manifest in going from the time domain to the frequency domain in signal processing: The temporally unfolding continuous signal, because of its periodicity, is seen as equivalent to the discrete frequency and amplitude values. But also our western musical conceptual apparatus is, needless to say, filled with such continuous to discrete transitions in the form of various categories and schemas. This goes not only for basic elements like pitch, duration, and timbre, but also for more composite phenomena like whole patterns of tones and even more extended works of music. As such, Xenakis' ideas of "outside time" ("hors-temps"), epitomizes our heritage's capability for such notions of discontinuity on the basis of continuity (Xenakis 1992).

However, this continuous to discrete transition in perception of more composite phenomena becomes quite enigmatic when we try to get a better understanding of how it works. Clearly, our event-level perception is also a biological capability common to other species, and involves feature extraction and memory in different forms. Yet it is also an epistemological-philosophical issue regarding time perception in general, and it was a hotly debated topic in the early stages of phenomenological philosophy, as reflected in Husserl (1991), partly dating from 1893 (see Godøy 2010 for details). The basic challenge for Husserl and several of his

contemporaries was this: How is it that we can perceive a melody (or a single tone, for that matter), when we know that the actual physical sound of each tone (or part of any tone) is gone right after its sounding? Husserl discussed and rejected various solutions, and ended up with the famous tripartite model of *retention* (cumulative memory for that which has just passed), *primary impression* (what is happening presently), and *protention* (expectations of what is to come), and with the understanding that perception proceeds by a series of so-called "now-points" that include all three elements of the tripartite model. This has fittingly been called "The Principle of Simultaneous Awareness" by Miller (1982), and could be understood as a quantal element in perception and cognition. Another very interesting element here is that Husserl's "now-points" are *intermittent*, i.e., discontinuous in occurrence, something that we shall return to later.

In more recent research, issues of continuity versus discontinuity in perception and cognition seem still to create debate: On the one hand, there are proponents of continuity who regard most mental activity as a continuous and competitive process with more gradual emergent percepts (e.g., Spivey 2008), and on the other hand, proponents of discontinuity, suggesting that although neuronal processes may be continuous, there is discontinuous, point-by-point awareness and decision-making based on transitions of activation thresholds in the dynamics of neuronal activity (Sergent and Dehaene 2004). A similar moment-by-moment approach has been advocated by Pöppel (1989, 1997), suggesting that these moments are in the roughly 3 s range. Pöppel also suggests that these moments are of similar duration in music, various other arts (e.g., utterance length in poetry), and in very many everyday actions, in turn suggesting a mutual attunement of temporal chunking in consciousness and in body motion.

In other recent memory research, it has been suggested that memory is composite, with components working at different timescales (Snyder 2000): The so-called *echoic memory* being more or less on the continuous signal level, and the *short-term memory* being on the couple of seconds level, i.e., approximately at the chunk-level timescale. In addition, there is the *long-term memory* that has the function of recalling long stretches of events, and also of forming categories and schemas. These categories and schemas, as well as more concrete recollections of past events in long-term memory, are thought to contribute to any ongoing perceptual process, hence are variably reactivated as needed. These various models seem to assume (if not state explicitly) that perception is based on holding continuous sensations in some kind of buffer, hence have a quantal element.

In addition to Schaeffer's pioneering work on sonic objects, there has also been research on *auditory objects*, suggesting that there is indeed a quantal element at work in auditory perception (Griffiths and Warren 2004; Winkler et al. 2006). A modeling of auditory buffering, including some interesting ideas on the holistic and temporally bi-directional working of memory can be found in Grossberg and Myers (2000). It should be commented here that K. N. Stevens' *quantal theory* in linguistics, first presented in 1972, is basically a theory concerned with tolerance

for variation within feature space (i.e., what could be called intra-categorical variation), not with temporal compression as we are here (Clements and Ridouane 2006).

Common to most of the abovementioned research on chunking in perception is that of somehow transforming, or *re-coding*, to borrow the expression from Miller (1956), the continuous into something more discontinuous in our minds. One way to understand these quantal elements in perception could then be that of *constraints* in the sense of a necessity, both in perception (for extracting useful information) and in body motion (for anticipation in motor control), as we shall have a look at in the next section.

4 Constraint-Based Chunking

There are various constraint-based factors at work in the generation and perception of sound that may result in quantal elements in musical experience. For one thing, natural sound is very often continuous within certain duration limits: Sounds typically (but of course variably so) have an envelope with an attack, sustain, and decay. In percussive and plucked sounds, there is an excitation in one point in time followed by a variably long reverberation, thus being a quantal sonic element in the sense that it is one coherent event.

Similarly, sound-producing (and sound-accompanying) body motion is also continuous in the sense that all body motion takes time, i.e., it is physiologically impossible to instantly go from one position to another position. This leads to the phenomenon of *coarticulation*, meaning that at any point in time, the position and speed of an effector (be that finger, hand, arm, etc., or any part of the vocal apparatus) will be determined by the immediately preceding motion as well as the immediately succeeding motion (see Godøy et al. 2010 for details). There is an interesting similarity here with Husserl's tripartite model in that at every instant, at every "now-point" in Husserl's terms, there is the effect of that which has just passed and of that which is to come. Importantly, this temporal coarticulation is a contextual smearing, blurring the boundaries between micro-level events, and effectively creating a quantal element of cohesion within any music-related body motion, as well as often also within the sonic object. The sonic smearing by coarticulation is evident and much studied in linguistics (Hardcastle and Hewitt 1999), and also clearly at work in various musical contexts (Godøy et al. 2010).

Going one step further, it is clear that in motor control, there is a need for anticipation, both in the sense of positioning the effectors for any sound-producing task (e.g., fingers on the keyboard), and for optimal motor control when required motions are too fast to be produced without being preprogrammed into one compact action chunk that is automatically performed in due course. The question of preprogramming of action has been hotly debated for more than a century

(Elliott et al. 2001), and in particular Karl Lashley was an advocate for the need for preprogramming in human behavior (Lashley 1951). However, a general consensus seems to be emerging that preprogramming is necessary, yet that there is also a possibility of feedback and adjustment in the course of one action (Rosenbaum et al. 2007). Notwithstanding any possibilities of ad hoc adjustment, there is substantial evidence in favor of anticipatory and holistic cognition of action gestalts (Klapp and Jagacinski 2011), a preprogramming also evident in the workings of coarticulation as indicated above.

Research on *action hierarchies* attest to such quantal elements in human body motion, where in goal-directed action there is clearly a need for "…a control mechanism in which the different motor elements are not simply linked together in the correct sequence, but are also tuned individually and linked synergistically based on the final goal, with coarticulation observed as an emergent phenomenon." (Grafton and Hamilton 2007, 593–594). And further more, with reference to Nicholai Bernstein's seminal work on motor control: "The fifth aspect of Bernstein's work is now referred to as chunking, the integration of independent motor elements into a single unit. With chunking there can be an increase of coarticulation and reduced cognitive demands, because less elements are organized for a given motor goal. […] Chunking is also a critical element for automatization to emerge." (ibid, 592). And finally: "What Bernstein was trying to identify was a more fluid process for structuring action, based on an internal hierarchical model where elements describing shorter action sequences or motor primitives could be combined. In such a structure a desired outcome is achieved within a cascade of intermediate steps that converge onto a solution. In this situation, the desired outcome is an invariant representation that is held as a reference during planning, when the desired elements are organized." (ibid, 593).

It seems fair to conclude that with both action hierarchies and coarticulation at work in music-related body motion, as well as with musical instrument physical constraints of reverberation (dissipation), we have constraint-based quantal elements in music as follows:

- Sonic events are acoustically often quantal in nature in the sense that we have excitation and reverberation as coherent events that can be holistically conceived and perceived.
- Sonic events and sonic features are included in some kind of sound-producing body motion trajectory.
- Motion trajectories exhibit coarticulation, meaning contextual smearing that creates continuity within a chunk.
- Motion trajectories require anticipatory cognition, hence a compressed overview image of the whole trajectory, constituting a quantal element in motor cognition.

Taking motor elements into consideration, i.e., taking the so-called *motor theory* perspective (Liberman and Mattingly 1985; Galantucci et al. 2006), it becomes quite clear that chunking to a large extent is also determined by motor control, and in particular by the phenomenon of what we call *key-postures*.

5 Key-Postures

One way of understanding human body motion is that it proceeds by a series of goal-directed actions. This is clearly valid for various everyday body motion (Grafton and Hamilton 2007), and it will be argued, is also valid for music-related body motion. In Rosenbaum et al. (2007), a review of research in support of goal-directed action is presented, ending up with a general model of *key-frames* and *interframes* for human motion control. Borrowed from the field of animation, the term "key-frames" designates the images of body postures at salient moments, and the term "interframes" designates the intermediate frames that need to be made (draw by hand or computer generated) between the key-frames in order to create sensations of continuous motion. The generation of interframes in actual human motion is considered less demanding than that of key-frames, and may be left to sub-routines of the motor control system.

Borrowing this idea from Rosenbaum et al. (2007), we prefer to use the term *key-postures*, to avoid potential confusion as to the meaning of key-frames (also used in contexts of video encoding), as well as to indicate that key-postures include both *position* and *shape* (hence also *spread*) of effectors, e.g., the position of hands and spread of fingers on a keyboard. Furthermore, key-postures are *intermittent*, meaning that with key-postures we have a discontinuous, point-by-point, planning and control of music-related motion. We shall assume that these key-postures occur at salient moments in time, typically at downbeats and other accented points in the music, and we will furthermore argue that sound-motion chunks are oriented around such key-postures.

As for downbeats and other accents, we seem to lack a more comprehensive theory as to their cognitive underpinnings. Also, the concept of 'beat' in music is often used, and there seems to be a working consensus as to its meaning, whereas there seems to be less consensus as to its perceptual-cognitive workings (see Todd et al. 2002 for an overview). However, there is an interesting study of the *shape* of the beat trajectory (Elliott et al. 2009), and related to this, studies of *perceptual centers,* so-called *p-centers*, i.e., of the subjective experience of the attack-point in musical sound (Wright 2008). The beat shape research suggests that the shaper (or more pointed) the shape of the effector motion trajectory (e.g., hand motion) for the beat, the more accurate is the resultant timing, and conversely, the more sloped the motion trajectory, the more inaccurate the resultant timing. As for p-centers, available research seems to indicate that the perceived point of an attack is to a large extent a mental phenomenon based on several factors, and not an unambiguous acoustical fact only based on the amplitude peaks.

One plausible candidate for what a downbeat is, or generally, what an accent is, seems to be related to a sensation of force. In discussing beat perception in conducting, Luck and Sloboda (2009) suggest *acceleration* as an important cue for downbeat perception, and that this is related to peaks in muscle contraction: "As for why acceleration along the trajectory might offer a cue for synchronization, there are interesting parallels between this finding and the fact that, when

individuals spontaneously move to rhythmic auditory stimuli (beat-driven music), bursts of instantaneous muscular power tend to be associated with rhythmic elements of the stimuli (the beat of the music). […] That is, people exert muscular energy in synchrony with the beat. If we assume that muscular energy peaks at moments of maximum acceleration of a conductor's hand, we might speculate that perception of a visual beat, like corporeal representation of a musical beat, is closely related to peaks of instantaneous muscular power." (Luck and Sloboda 2009, 472).

Another attribute of the downbeat, as suggested by its etymology, is a downward motion of the hands when conducting: "The explicit or implied impulse that coincides with the beginning of a bar in measured music, by analogy with the downstroke in conducting." (Julian Rushton, "Downbeat", Grove Music Online, accessed 13 October 2012).

From the elements briefly presented above, it seems reasonable to infer that there is *an unequal distribution of attention and effort in music-related body motion*, something that may be summarized as follows:

- Continuous monitoring and adjustment in motor control is problematic for the simple reason that monitoring and adjusting takes time, i.e., the process would be too slow for many musical purposes. This is in accordance with Lashley's classic rejection of so-called *reflex chaining* and his alternative view that there must be preprogramming (anticipative cognition) to enable efficient motion control (Lashley 1951; Rosenbaum et al. 2007). There is also some recent research suggesting that motor control is optimal when intermittent (Loram et al. 2011), i.e., that accuracy of body motion is actually enhanced by having discontinuous control.
- Body motion is centered on certain key-postures, cf. the reasonably well-documented theories of action hierarchies (Grafton and Hamilton 2007) and posture-based motion control (Rosenbaum et al. 2007) mentioned above.
- Between these key-postures, we have continuous motion and coarticulation (Grafton and Hamilton 2007).
- It seems reasonable to assume that also chunks of music-related body motion are centered on key-postures, and furthermore, that key-postures occur typically on downbeats and other accented points. In addition, there are often upbeat motion trajectories, what we call *prefixes*, prior to these downbeat key-postures, and the downbeat key-postures are very often followed by a continuous trajectory until the next downbeat key-posture, what we call *suffixes*, thus creating continuous motion around the key-postures (Godøy 2010).

As key-postures are intermittent, and downbeats (and other beats and accents) are intermittent, the next step is to see if this scheme of intermittent control of music-related motion can somehow be modeled, and in effect demonstrating quantal elements in musical experience.

6 Impulse-Driven Chunking

Given the constraints of key-postures at salient moments in time as well as the emergent coarticulation, it could be possible to think of musical performance (and hence, also of music perception in a motor theory perspective) as a series of impulses with intervening continuous motion trajectories.

As an example, consider the waltz-like texture for piano left hand solo in Fig. 1. Underneath the notation of this fragment, we see the motion capture data for the position, velocity and acceleration of the hand, wrist, and elbow, along the vertical plane. We can see peaks in velocity and acceleration at the downbeat points, indicating salient moments in time (as suggested above), as well as key-postures. The ensuing offbeat tones (the sixth C4–E3) on the *twos* and *threes* of each measure are included in the overall motion trajectories for each measure, and are in effect subsumed as coarticulated events. This means that each measure here is one chunk, centered on the downbeat, a downbeat that coincides with a key-posture.

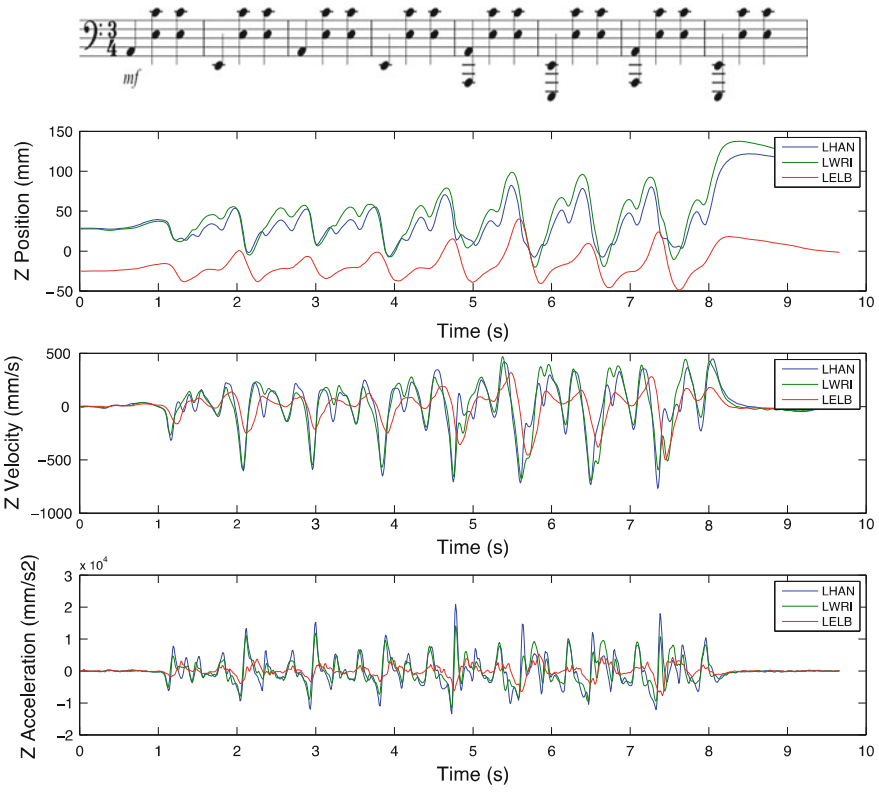


Fig. 1 A waltz-like fragment for the *piano* for *left hand solo*, with notation (*top*), and underneath this, motion trajectories of hand, wrist, and elbow along the *vertical plane*, and underneath this again, the velocity and acceleration plots of this motion data

From the combination of these observations and the abovementioned more general principles of motor control, we could reasonably infer that these chunks are *impulse-driven* in the sense that the downbeat and key-posture are made intermittently present at one point in time, and that the rest of the chunk follows as a result of, and is subordinate to, this impulse. Knowingly setting many and quite complex motor control and biomechanics issues aside for now, it could still be tempting to make a sketch of a general model for impulse-driven chunking in music as an initial hypothesis that would of course need much more elaboration.

In Fig. 2, we see a schematic illustration of such impulse driven chunking. At the top, (a), we have an 'instantaneous' impulse [similar to a Dirac impulse which goes off to a maximal value (or theoretically, to infinity) at one point and is zero elsewhere] combined with a key-posture at that point. However, the effector (finger, hand, elbow, etc.) needs time to get to this key-posture; hence there is a prefix trajectory up to the peak, followed by a suffix trajectory from this peak back to equilibrium (i.e., a state of relative stillness). This looks similar to convolving and impulse with an impulse-response, and analogously, we could here think that the impulse is convolved with the body response, resulting in the trajectory we see at (b). We could say that the musical result is an upbeat prefix to the downbeat point followed by a decaying suffix trajectory back to equilibrium after the downbeat point. Furthermore, when we have several such impulses in succession (which we most often will have in music as we have a series of beats and key-postures), this may be depicted as a series of impulses like in (c). Again, due to the constrains of the effector system, there are prefix trajectories to the peaks as well as suffix trajectories from the peaks back to equilibrium, so that we have trajectories as in (d). Here we see that the prefix trajectories and the suffix trajectories may overlap (if they are close enough and are not interleaved with moments of stillness), hence we have the resultant seemingly continuous and undulating trajectory we see in (e). The effect of this is that of having *continuity as the emergent phenomenon of an underlying series of discontinuous impulses*, hence, in a sense reconciling discontinuity and continuity by such impulse-driven chunking. A similar schematic model of overlapping trajectories of phonemic gestures has been suggested in research on coarticulation in linguistics, see e.g. (Hardcastle and Hewlett 1999, 52).

Modeling impulse-driven chunking in music-related motion could be extended from this conceptual sketch to a more detailed mathematical implementation, in the line of what Edward W. Large has called a "normal form dynamical system" for his modeling of rhythm perception (Large 2000, 534). But initially, impulse-driven chunking has attractive features to explore further also in other directions. For one thing, it treats music as a series of overlapping chunks with superordinate, coarticulatory motion trajectories, i.e., it does not primarily rely on internal gestalt features for coherence, but instead on the overall impulse-driven motion trajectories. Musically (and acoustically), impulse-driven chunking could thus account for the fusion of sonic events into holistically experienced chunks as may also be simulated by so-called *diphone* synthesis. As such, impulse-driven chunking offers a different scheme for the understanding of gestalt formation, in that gestalt

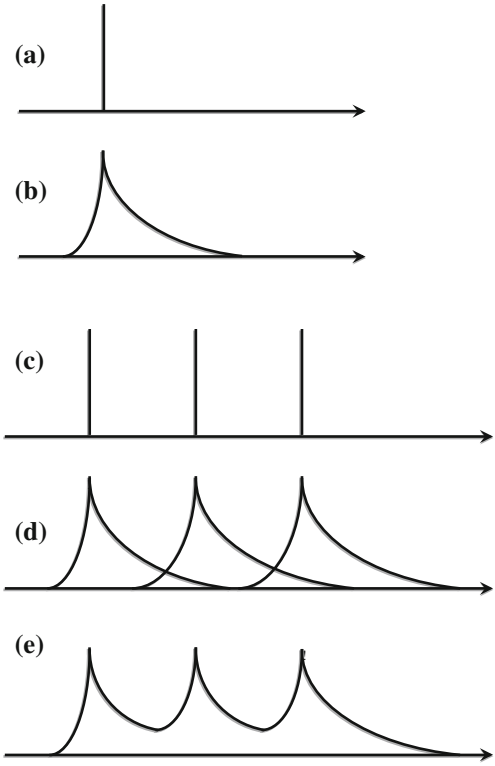


Fig. 2 Schematic illustration of impulses and trajectories. First is shown a *single impulse* (**a**), and this impulse as embedded in a trajectory (**b**). In (**c**) there is a series of *three impulses*, and in (**d**) these impulses are embedded in trajectories. However, with overlap, the emergent result in (**e**) is of an undulating motion trajectory that does not go down to equilibrium (stillness) before at the end

coherence here is not primarily understood as a bottom-up, feature-based phenomenon, but just as much as a top-down impulse-driven, motor schematic, quantal element in music.

Furthermore, impulse-driven chunking has the advantage of shedding light on the problem of determining chunk boundaries: By focusing on impulse-points rather than first of all trying to find the precise start and end points of a chunk, chunk boundary issues can be determined by classifying continuous motion as either belonging to a prefix (upbeat) trajectory or to a suffix (after downbeat) trajectory, with the chunk boundaries determined as the (not so critical) border between these two.

Impulse-driven chunking can be seen as the source of rhythmical grouping, in some cases (but by no means always), also at the so-called *beat-level* in music, e.g., as in the waltz-like fragment in Fig. 1 when it is played so fast that the measure level becomes equal to the beat level of the music. Notably, impulse-driven chunking need not be restricted to isochronous beats, and beats may very well be non-periodic, e.g., as in various folk music, as it is the overall motion trajectory that controls the length of the beats. Impulse-driven chunking could potentially give us interesting views on other 'irregular' rhythmic phenomena as well as syncopation. Some very interesting ideas on beat-level rhythmical

chunking have been presented in Sethares (2007, p. 279), with the significant question "Which Comes First, the Notes or the Beat?" And furthermore, ornaments of all kinds, and various other figures, may be considered to be impulse-driven chunks, making this useful in musical analysis and studies of styles, cf. David Cope's "signatures", i.e., small, style-typical fragments (Cope 1991). Last but not least, expressivity and rhythmical articulation, timing, grooving, etc., could all likewise benefit from being understood as impulse-driven chunks.

7 Conclusions

In summary, it seems clear that we usually have different concurrent timescales, ranging from the micro-level to the macro-level, at work in musical experience, but that there is good reason to designate the meso-level timescale (very approximately in the 0.5–5 s duration range) as highly significant for very many musical features. At the meso-level, there are also clear indications of quantal elements in the holistic perception and production of chunks of sound and music-related body motion.

Several signal-based and/or auditory feature-based arguments for this quantal nature of meso-level chunks may be presented, such as in various gestalt-related lines of inquiry (Tenny and Polansky 1980; Bregman 1990; Winkler et al. 2006), or in other auditory research (Grossberg and Myers 2000; Griffiths and Warren 2004). But as a supplement, there is the *motor theory* perspective taken in this chapter, implying that the quantal nature of meso-level chunks are related to body motion trajectories, be that in the production of sound or in various sound-accompanying motion, or in combinations of these. The basic tenet here is that all within-chunk features are subordinate to the chunk-level motion trajectories, trajectories that in turn are centered on key-postures and driven by impulses, hence that we have a quantal element of musical experience here.

The scheme of impulse-driven chunking turns things around, in the sense that chunk boundaries are considered secondary to chunk impulse-points, so that what comes before the impulse point is a prefix (upbeat) trajectory and what comes after the impulse point is a suffix, i.e., a trajectory returning to equilibrium (unless being 'interrupted' by a new upbeat trajectory before equilibrium as in Fig. 2d, e).

The next step now will be to develop this model sketch into a generic mathematical model, and to test it out on real sound-motion data. Later, more developed versions of this model could be tested in music perception experiments, in music analysis, and in music information retrieval, all with the basic tenet that the 'building blocks' of music are not just tones/notes, but just as much quantal element of impulse-driven sound-action chunks.

Acknowledgments Many thanks to my co-workers Kristian Nymoen for help with the motion capture recordings and Alexander Refsum Jensenius for the Matlab plottings.

References

Bregman, A. (1990). *Auditory scene analysis*. Cambridge and London: The MIT Press.

Chion, M. (1983). *Guide des objets sonores*. Paris: INA/GRM Buchet/Chastel.

Clements, G. N. & Ridouane, R. (2006). Quantal phonetics and distinctive features: A review. In *Proceedings of ISCA. Tutorial and Research Workshop on Experimental Linguistics, 28–30 August 2006*. Athens, Greece.

Cope, D. (1991). *Computers and musical style*. Madison, Wisconsin: A-R Editions, Inc.

Cutting, J. E. (1982). Plucks and bows are categorically perceived, sometimes. *Perception and Psychophysics, 31*(5), 462–476.

Elliott, D., Helsen, W., & Chua, R. (2001). A century later: Woodworth's (1899) two-component model of goal-directed aiming. *Psychological Bulletin, 127*(3), 342–357.

Elliott, M. T., Welchman, A. E., & Wing, A. M. (2009). Being discrete helps keep to the beat. *Experimental Brain Research, 192*, 731–737.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review, 13*(3), 361–377.

Gjerdingen, R., & Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research, 37*(2), 93–100.

Godøy, R. I. (2006). Gestural-sonorous objects: Embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound, 11*(2), 149–157.

Godøy, R. I. (2010). Thinking now-points in music-related movement. In R. Bader, C. Neuhaus & U. Morgenstern (Eds.), *Concepts, experiments, and fieldwork. Studies in systematic musicology and ethnomusicology* (pp. 245–260). Frankfurt am Main: Peter Lang.

Godøy, R. I., & Leman, M. (2010). *Musical gestures: Sound, movement, and Meaning*. New York: Routledge.

Godøy, R. I., Jensenius, A. R., & Nymoen, K. (2010). Chunking in music by Coarticulation. *Acta Acustica United with Acustica, 96*(4), 690–700.

Grafton, S. T., & Hamilton, A. F. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science, 26*, 590–616.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience, 5*, 887–892.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review, 107*(4), 735–767.

Haken, H., Kelso, J., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51*(5), 347–356.

Hardcastle, W. J., & Hewlett, N. (1999). *Coarticulation : Theory, data and techniques*. Cambridge: Cambridge University Press.

Hogan, N., & Sternad, D. (2007). On rhythmic and discrete movements: Reflections, definitions and implications for motor control. *Experimental Brain Research, 181*, 13–30.

Husserl, E. (1991). *On the phenomenology of the consciousness of internal time, 1893–1917 (trans: Brough, J.B)*. Doredrecht-Boston-London: Kluwer Academic Publishers.

Klapp, S. T., & Jagacinski, R. J. (2011). Gestalt principles in the control of motor action. *Psychological Bulletin, 137*(3), 443–462.

Large, E. W. (2000). On synchronizing movements to music. *Human Movement Science, 19*, 527–566.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–131). New York: Wiley.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Loram, I. D., Gollee, H., Lakie, M., & Gawthrop, P. J. (2011). Human control of an inverted pendulum: Is continuous control necessary? Is intermittent control effective? Is intermittent control physiological? *The Journal of Physiology, 589*(2), 307–324.

Luck, G., & Sloboda, J. (2009). Spatio-temporal cues for visually mediated synchronization. *Music Perception, 26*, 465–473.

Miller, G. A. (1956). The magic number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Miller, I. (1982). Husserl's account of our temporal awareness. In H. Dreyfus (Ed.), *Husserl, intentionality, and cognitive science* (pp. 125–146). Cambridge and London: The MIT Press.

Moore, B. C. J. (Ed.). (1995). *Hearing*. San Diego: Academic Press.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America, 130*(5), 2902–2916.

Pöppel, E. (1989). The Measurement of music and the cerebral clock: A new theory. *Leonardo, 22*(1), 83–89.

Pöppel, E. (1997). A hierarchical model of time perception. *Trends in Cognitive Science, 1*(2), 56–61.

Rosenbaum, D., Cohen, R. G., Jax, S. A., Weiss, D. J., & van der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science, 26*(4), 525–554.

Schaeffer, P. (1966). *Traité des objets musicaux*. Paris: Éditions du Seuil.

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science, 15*(11), 720–728.

Sethares, W. A. (2007). *Rhythm and transforms*. Berlin-Heidelberg: Springer.

Snyder, B. (2000). *Music and memory: An introduction*. Cambridge and London: The MIT Press.

Spivey, M. (2008). *The continuity of mind*. New York: Oxford University Press.

Tenny, J., & Polansky, L. (1980). Temporal gestalt perception in music. *Journal of Music Theory, 24*(2), 205–241.

Todd, N. P., O'Boyle, D. J., & Lee, C. S. (2002). A sensorimotor theory of temporal tracking and beat induction. *Psychological Research, 66*, 26–39.

Winkler, I., van Zuijen, T. L., Sussman, E., Horváth, J., & Näätänen, R. (2006). Object representation in the human auditory system. *European Journal of Neuroscience, 24*(2), 625–634.

Wright, M. J. (2008). *The shape of an instant: Measuring and modeling perceptual attack time with probability density functions*. PhD Thesis, Stanford University.

Xenakis, I. (1992). *Formalized music (revised edition)*. Stuyvesant: Pendragon Press.

Part II
Neurocognition and Evolution

Strong Emotions in Music: Are they an Evolutionary Adaptation?

Eckart Altenmüller, Reinhard Kopiez and Oliver Grewe

1 Introduction: The Difficult Question About the Origins of Music

There is general agreement that all human cultures possessed and still possess music. Here, we understand music as intentionally created, non-linguistic, acoustical events, structured in time, and produced in social contexts (Altenmüller und Kopiez 2005). Amongst the oldest cultural artefacts, musical instruments such as bone and ivory flutes have been discovered in the Hohle Fels cave and the Geissenklösterle cave in the region of Swabia, South-West Germany (Conard und Malina 2008). These flutes, dating back to about 35,000 years, indicate a paleolithical musical tradition at the time when modern humans colonized Europe. Intriguingly, they are tuned in line with a "modern" diatonic scale: the grip holes of the flute are arranged in such a way that an octave is divided into five whole steps and two half steps, with the half steps being separated by at least two whole steps. The tuning is so "modern" that the main theme of J.S. Bach's Kunst der Fuge (The art of the Fugue) can be played on a reconstructed Geissenklösterle-flute (Münzel et al. 2002, see Fig. 1). Nicholas Conard, the archaeologist who is in charge of the excavation in the Hohle Fels cave therefore speculates that there might have existed cultural traditions, which persisted from the paleolithic ages until our times and preserved this diatonic scale locally in Central Europe (Conard et al. 2009). This is a strong claim, since performance parameters, i.e. the embouchure and the speed and width of the air-jet used to blow, may yield pitches, which vary more than a quarter tone (Liang 2002).

A modified version of this article has been published in February 2013 in: E. Altenmüller, S. Schmidt & E. Zimmermann (Eds.), Evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man. Oxford, UK: Oxford University Press.

E. Altenmüller (✉) · R. Kopiez · O. Grewe
Hanover University of Music, Drama and Media, Hanover, Germany
e-mail: eckart.altenmueller@hmtm-hannover.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_5,

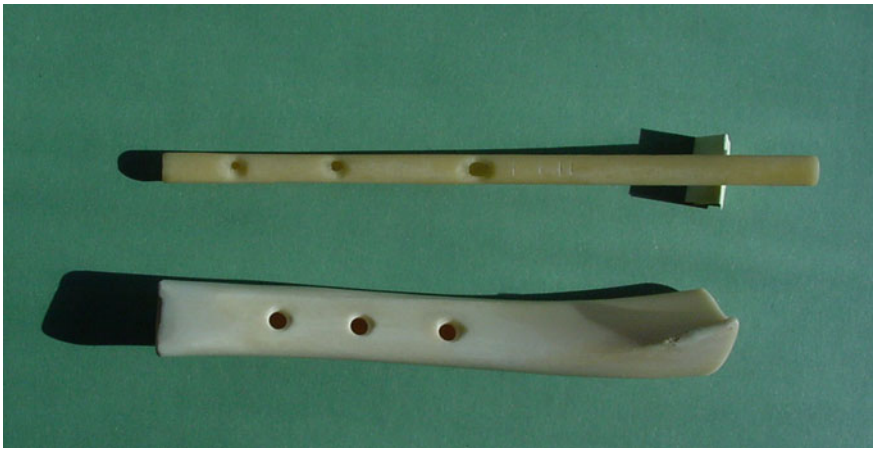


Fig. 1 Replicas of the Geissenklösterle- and the Grubgraben-flutes dating back to about 35,000 and 20,000 years respectively. The 22 cm *long* Geissenklösterle flute is made from the radius bone of a swan wing. The *grip holes* are arranged in a way that five notes of a perfectly tuned diatonic scale can be played. It is unclear, whether the horizontal carvings are ornaments or were used to determine the position of the grip holes. The 16 cm long Grubgraben flute is made from a reindeer tibia. It is similarly tuned in a diatonic scale, however easier to play. The replicas are manufactured by Wulf Hein, paleo-technician. Photo by E. Altenmüller

Furthermore, such a presumed tradition is generally difficult to prove due to lacking continuity of records used in different sites at different times. It might belong to one of those romanticisms, frequently encountered when dealing with speculations about music and its evolutionary roots. We still do not understand the exact function of these flutes, since it is even unclear whether they were regarded as musical instruments and aesthetic objects or for example as signalling instrument, used by hunters or gatherers to indicate a temporary station or to require a specific action. Conard and Malina (2008) claim that the emotional life of palaeolithic individuals was not different from ours. They therefore suggest that these flutes were indeed used for playing expressive tunes and designed to influence early humans' well-being, emotions, group cohesion, and sense of beauty. In favour of this hypothesis is the fact that the manufacturing of these flutes was extremely time consuming and required fine manual skills and technical expertise (Münzel and Conard 2009). Earlier musical activities are likely to have existed although they are not documented in artefacts or as cave art. Here, we consider instruments made from less durable materials, i.e. reed and wood, and furthermore of joint singing, hand clapping or drumming as being connected to motor activities such as rhythmic movements and dancing. It is an open question though, to why these musical activities did emerge or persist, despite them being labour intensive and therefore costly in an environment of constant struggle for survival.

From a scientific viewpoint the question of the origin of music is difficult, if not impossible, to answer. There is too little information available about the nature of musical activities in prehistoric times. Music does not fossilize and we rely on

sparse documents, mainly artefacts such as the above-mentioned flutes. There are remarkably few cave paintings depicting musicians. Probably the earliest—though still debated—depiction of the use of a musical instrument in rituals is the "Schaman with the Mouth Bow" in the Cave "Le Trois Frères" dating back to about 14,000 years (Anati 1996) (see Fig. 2).

Further indirect information concerning the origins of music can be obtained either from a comparative approach, for example when analysing the acoustic communication of nonhuman mammals or from cross-cultural studies, especially when comparing music production and appreciation in humans who have been isolated from westernized cultures, such as the Mafa in the North of Cameroun (Fritz et al. 2009). Finally, conclusions can be drawn from considering ontogenesis, observing the individual developments of vocalizations and responses to music in infants (e.g. Mampe et al. 2009; Zentner and Eerola 2010). Undoubtedly, as many animal vocalizations do in conspecifics, music can evoke strong emotions and change the state of arousal when listened to attentively (Grewe et al. 2007a, b, see also Panksepp and Bernatzky 2002, for the role of attention see Kämpfe et al. 2011). These strong emotions can even have effects on physiological functions, for example on heart beat frequency (Nagel et al. 2008) and brain neurotransmitter production (e.g. Salimpoor et al. 2011). Thus, it is not far-fetched to speculate that

Fig. 2 The Schaman with the mouth bow from the cave "le Trois Frères". Redrawing from Henri Bégouen and Henri Breuil, Les cavernes du Volp. Trois-Frères—Tuc d'Audoubert à Montesquieu-Avantès (Ariège). Paris, 1958, Figure 63. Redrawing by E. Altenmüller

our love for music may have an adaptive value based on evolutionary old mechanisms, linked to the very nature of humans.

In the following, we will strive to find an answer to whether there is sufficient evidence supporting the claim that music is an evolutionary ingrained characteristic of humans. Furthermore we will discuss if making and listening to music as a means to produce and experience strong emotions is a fitness relevant behaviour or not. Finally, we will propose a tentative model attributing the origins of music to a variety of either biologically relevant sources or culturally "invented" activities, thus reconciling the opposing adaptationist—non-adaptationist standpoints of "music as part of the evolutionary founded endowment of humans" (e.g. Brown 2000) versus music as an "invention of humans, a transformative technology of the mind" (Patel 2010).

2 Is Music an Evolutionary Adaptation?

For the sake of brevity we will only summarize the discussion on a potentially adaptational value of music in human cultures. There are several recently published articles and books, reviewing this on-going discussion in more detail (e.g. Patel 2010; Grewe et al. 2009a; Special Issue of Musicae Scientiae 2009/2010). Furthermore, we refer to the excellent classic "The origins of Music" edited by Wallin et al. (2000).

The adaptationist's viewpoint posits that our capacity to produce and appreciate music has an evolutionary adaptive value; it is the product of a natural selection process and contributes to the "survival of the fittest". It implies that it is biologically powerful and based upon innate characteristics of the organism, for example specialized brain networks refined by acculturation and education. Historically Darwin (1871, p. 1209), who proposed in his book "The Descent of Man" an analogy of human music to bird-song, has been the most prominent exponent of this view. He wrote: "Musical tones and rhythm were used by half-human progenitors of man, during a season of courtship, when animals of all kinds are excited by the strongest passions". He further argued that the use of music might have been antecedent to our linguistic abilities, which evolved from music. This thought has been recently elaborated in the musilanguage model of Brown (2000). Indeed, the idea that music or at least musical elements producing strong emotions could be precursors of our language capacity has been already developed in 1770 by Herder in his "Treatise on the Origin of Language" (Abhandlung über den Ursprung der Sprache, Voss, Berlin, 1772), which received the price of the Royal Academy of Berlin. Here, Herder states that language may have evolved from a "natural" affective sound system, common to humans and animals, which aimed at the communication of emotions: "since our sounds of nature are destined to the expression of passions, so it is natural that they will become elements of all emotions". According to Darwin and Herder, music is an acoustic communication system conveying information on emotions and inducing emotions, thus either

(according to Darwin) promoting success in reproduction or (according to Herder) improving social cohesion. These two adaptationists arguments are still discussed: Miller (2000) has explored the sexual selection hypothesis, arguing that making music was a demonstration of hidden qualities in the struggle for mates. Playing a musical instrument means that resources for building such an instrument and investing time in practicing it are available. Furthermore, the performance itself requires self-confidence, creativity, emotionality and, frequently, bodily features (such as skilled use of fingers) which can be conceived as the display of otherwise hidden qualities.

With respect to the social coherence hypothesis, recent research convincingly demonstrates that making music together promotes prosocial behaviour in Kindergarten children (Kirschner and Tomasello 2010). Furthermore, music seems to have the potential to initiate or reinforce the social bonding among individuals in a group by means of "emotional resonance" and shared emotional experiences. McNeill (1995, p. viii) assumed that "moving our muscles rhythmically and giving voice consolidate group solidarity by altering human feelings." In other words, keeping together in time creates social cohesion.

Cross (2009) extended this theory to the effects of music on the human capacity for entrainment. According to Cross, listening to music and making music increases "… the likelihood that participants will experience a sense of shared intentionality. […] Music allows participants to explore the prospective consequences of their actions and attitudes towards others within a temporal framework that promotes the alignment of participant's sense of goal. As a generic human faculty music thus provides a medium that is adapted to situations of social uncertainty, a medium by means of which a capacity of flexible social interaction can be explored and reinforced" (Cross 2009, p. 179). In going further, he ascribes music a role "as risk-free medium for the exercise and rehearsal of social interaction" (Cross 2008).

Besides sexual selection and group-cohesion, adaptationists frequently propose the role of musical and music-like interactions during parental care as a third major group of evolutionary adaptive behaviours. Motherese, for example, is a specific form of vocal-gestural communication between adults (mostly mothers) and infants. This form of emotional communication involves melodic, rhythmic, and movement patterns as well as communication of intention and meaning and, in this sense, may be considered to be similar to music. Motherese has two main functions: to strengthen bonding between mother and infant, and to support language acquisition. Lullabies are universal musical activities designed to manipulate the infant's state of arousal, either by soothing overactive children or by arousing passive children (Shenfield et al. 2003). All of these functions enhance the infant's chances of survival and may therefore be subject to natural selection.

The importance of making music and listening to music as a potentially adaptational feature of humanity is underlined by neurobiological findings linking our sense of music to hard wired neuronal networks and adaptations of neurotransmitters. Humans possess specialized brain regions for the perception of melodies and pitches. This is impressively demonstrated by the selective loss of

the sense of melody and pitch in congenital and acquired amusia, the former being a genetically transmitted deficit in fine grained pitch perception, probably due to a dysfunctional right fronto-temporal neuronal network (Ayiotte et al. 2002). Furthermore, humans have specific sensory motor networks to adapt to and entrain with rhythmic stimulation. These networks are an almost unique feature in vertebrates, with only few exceptions such as the dancing Cacadu Snowball (Patel et al. 2009).

Strong emotions whilst listening to music have been shown to affect various neurotransmitters, predominantly the serotonergic and dopaminergic systems. Serotonin is a neurotransmitter commonly associated with feelings of satisfaction from expected outcomes, whereas dopamine is associated with feelings of pleasure based on novelty or newness. In a study of neurochemical responses to pleasant and unpleasant music, serotonin levels were significantly higher when participants were exposed to music they found pleasing (Evers and Suhr 2000). In another study with participants exposed to pleasing music, functional and effective connectivity analyses showed that listening to music strongly modulated activity in a network of mesolimbic structures involved in reward processing including the dopaminergic nucleus accumbens and the ventral tegmental area, as well as the hypothalamus and insula. This network is believed to be involved in regulating autonomic and physiological responses to rewarding and emotional stimuli (Menon and Levitin 2005).

Blood and Zatorre (2001) determined changes in regional cerebral blood flow (rCBF) with positron emission tomography (PET)-technology during intense emotional experiences involving chill responses accompanied by goose bumps or shivers down the spine whilst listening to music. Each participant listened to a piece of their own favorite music to which a chill experience was commonly associated to. Increasing chill intensity correlated with rCBF decrease in the amygdala as well as in the anterior hippocampal formation. An increase in rCBF correlating with increasing chill intensity was observed in the ventral striatum, the midbrain, the anterior insula, the anterior cingulate cortex, and the orbitofrontal cortex: again, these latter brain regions are related to reward and positive emotional valence.

In a recently published study by the same group, the neurochemical specificity of [11C]raclopride PET scanning was used to assess dopamine release on the basis of the competition between endogenous dopamine and [11C]raclopride for binding to dopamine D2 receptors (Salimpoor et al. 2011). They combined dopamine-release measurements with psychophysiological measures of autonomic nervous system activity during listening to intensely pleasurable music and found endogenous dopamine release in the striatum at peak emotional arousal during music listening. To examine the time course of dopamine release, the authors used functional magnetic resonance imaging (fMRI) with the same stimuli and listeners, and found a functional dissociation: the caudate was more involved during the anticipation and the nucleus accumbens was more involved during the experience of peak emotional responses to music. These results indicate that intense pleasure in response to music can lead to dopamine release in the striatal system. Notably,

the anticipation of an abstract reward can result in dopamine release in an anatomical pathway distinct from that associated with the peak pleasure itself. Such results may well help to explain why music is of such high value across all human societies. Dopaminergic activation furthermore regulates and heightens arousal, motivation and supports memory formation in the episodic and the procedural memory (Karabanov et al. 2010) and thereby will contribute to memorization of auditory stimuli producing such strong emotional responses.

3 Is Music a Human Invention?

The non-adaptationists theory postulates that music is a human invention and has no direct adaptive biological function. However, it can still be useful in terms of manipulating emotions, synchronizing group activities, supporting wellbeing and promoting health. An elegant analogy is the comparison of the ability to make and appreciate music to the ability of humans to control fire, which emerged probably some 150,000 years ago (Brown et al. 2009). Clearly, there is no "fire-making" gene and no neurological syndrome such as an "apyretia"—the inability to make and control fire—but nobody would deny that making fire had an enormous impact not only on human wellbeing (heating, cooking, lighting) and nutrition (better digestion of protein-rich diets from animal meat), but also on physiological parameters such as the configuration of our gut and teeth. Why not considering music as such an ingenious invention in humans?

Historically, this viewpoint dates back to Spencer (1857) and his essay "On the origins and functions of music". Spencer argued that music developed from the rhythms and expressive prosody of passionate speech. The eminent psychologist William James followed this line of arguments by stating that music is a "mere incidental peculiarity of the nervous system" (James 1890, vol. 2, p. 419), which has "no zoological utility" (vol. 2, p. 627).

Two decades later, German music psychologist Stumpf (1911) elaborated the non-adaptationist viewpoint. According to his theory, music is the result of correlative thinking, which allowed transgressing from sliding melodic contours to discrete pitches and intervals.

The most prominent modern protagonist of a non-adaptationist position is Pinker (1997, p. 528), who stated in his book entitled "How the mind works": "Music appears to be a pure pleasure technology, a cocktail of recreational drugs that we ingest through the ear in order to stimulate a mass of pleasure circuits at once". With respect to the biological significance Pinker comes to the same conclusion as James by stating: "As far as biological cause and effects are concerned, music is useless" (ibid., p. 534).

An elegant way to conceptualize music as a human invention, while taking into account how human musicality can shape brain functions (Münte et al. 2002) and even influence our genetic information (be it by selection or by epigenetic features) is the "transformative technology of the mind theory" (TTM-Theory)

proposed recently by Patel (2010). Basically, this theory has developed from a comparative approach, stating that there are aspects of music cognition rooted in other non-musical brain functions, which are shared with other animals. The logic behind it is as follows: if music relies on other brain functions developed for other purposes, then it is NOT music which has shaped our genetic material by natural selection. As in fire making, which relies on skilled motor hand functions developed as a consequence of upright gait and adept use of tools, our ancestors invented music by transforming previously acquired abilities (e.g. refined pitch processing, ability to keep in time with an external beat, see also Patel 2008, p. 207 f). This "invention of music", once established and tested for usefulness in several domains, does not preclude the later development of more specialised brain regions which may be of adaptational value—as is the length of the guts in fire making humans. For example, the potential to memorize melodies and harmonies critically relies on superior right temporal lobe functions, which are shaped by musical expertise (e.g. Schneider et al. 2002; Hyde et al. 2009). The same holds for the sensory-motor hand cortex, which adapts in function and structure to the breath-taking virtuosic skills of the hands in professional violinists and pianists (Bangert and Schlaug 2006, see also Hyde et al. 2009 for the discussion of the famous "hen-egg-problem").

Aniruddh Patel's argumentation in favour of the TTM theory is based on two general lines which we only briefly delineate here: firstly, he focuses on tonality processing and on the differential use of scale pitches such that some are perceived as more stable or structurally significant than others. He argues that this "musical" feature leading to implicit formation of tonal hierarchies is not domain specific, but shared with cognitive processing of syntactic hierarchies in language. Support for this theory is derived from the neuroscientific research on brain networks, which serve both processing of musical hierarchies and language syntax (see for a review Koelsch and Siebel 2005). Accordingly, music tonality processing shares the same resources as syntactic language processing; and both rely on much more basic cognitive operations, namely the general building of mental hierarchies or cognitive "reference points" (Krumhansl and Cuddy 2010) and the mechanisms of statistical learning (McMullen and Saffran 2004).

The second line of argumentation comes from Patel's work on entrainment to a musical beat. Musical beat perception and synchronisation is intrinsic to dance and to many other musical activities, such as synchronizing work songs or choir singing. It does not appear to be rooted in language, since at least prose language does not have temporally periodic beats and does not elicit periodic rhythmic movements from listeners (Patel 2008, Chap. 3). Furthermore, the ability to flexibly synchronize to changing tempi of musical beats seems to be unique to humans and to a few parrot species, who share with humans their excellent vocal learning ability. Here, it is important to note that adaptive rhythm and beat perception is essential for language acquisition. It is already present in prenatal intrauterine auditory learning (Patel et al. 2009). Summarizing, Patel argues that synchronisation to a musical beat relies on the brain systems designed for vocal learning, involving specialized auditory-motor networks not restricted to the

cortex, but also to midbrain structures such as the periaqueductal grey and its homologue in parrots. Thus, the TTM ability to keep in time with an external beat is a by-product of vocal learning and its neuronal prerequisites.

In summary, valid arguments are in favour of music as a human invention, based upon–or transformed from—already pre-existent cognitive and motor capacities of our brain. However, the TTM-theory neglects the strong impact of music on emotions and its possible origin and consequence with respect to its adaptational value. It is interesting to note that the emotional value of music has always been central to the adaptationists' viewpoint, beginning with Herder's and Darwin's ideas quoted above.

In the following two paragraphs we will demonstrate how music can elicit different types of emotions, namely aesthetic emotions and strong emotions. Aesthetic emotions are based on complex feelings with less salient physiological correlates, whereas strong emotions lead to shivers down the spine and chills or thrills accompanied by physiological reactions of the autonomous nervous system. We will argue that the former constitute parts of a TTM, invented in later times, whereas the latter may point towards an evolutionary old acoustic communication system we share with many other non-human mammals.

4 Emotions Induced by Music

Although most listeners agree that music can sound happy or sad, there is fewer consensus whether music truly evokes emotions. It is beyond the scope of this article to review the issue in detail, or furthermore whether and how music induces emotions. This discussion has recently been thoroughly reviewed by Hunter and Schellenberg (2010). Basically, two main theoretical standpoints are held: the cognitivist and the emotivist position. In brief, cognitivists argue that happy- and sad-sounding music does not evoke true happiness and sadness in listeners. Rather, affective responses stem from listener's evaluation of the music (Kivy 1990). However, such an evaluation or "appraisal" of music can clearly induce emotions, and is in itself a constituent of emotions according to the "component theory of emotions" by Scherer (2004). For example, a boring and inaccurate rendering of a musical masterpiece might well induce feelings of anger and frustration in a music lover based on his or her knowledge of other more adequate interpretations.

In contrast, emotivists posit that music directly evokes and induces emotions. Several mechanisms accounting for such a role of music are discussed. Amongst them, cognitive appraisal is only but one of them. Juslin and Västfjäll (2008) have proposed six other mechanisms, namely (1) Brainstem reflexes, (2) Conditioning, (3) Episodic memory, (4) Contagion, (5) Visual imagery, and (6) Expectancies that are fulfilled or denied.

With respect to brainstem reflexes, Juslin and Västfjäll consider automatic reactions of individuals towards highly dissonant sounds as such an emotional effect of music, acting via a hardwired neuronal network of the brainstem.

Although this phenomenon clearly exists the labeling is debatable, being in contrast with "true" brainstem reflexes, for example the constriction of the pupil following exposure to light, these reactions to music are inter-individually variable, adapting to repetitions and strongly depending on learning.

The emotional power of conditioning and episodic memory has been masterly portrayed by Marcel Proust in the chapter "Swann in love", part of the novel "In search of lost time": The hero Swann falls in love with a lady whilst a tune of Vinteuil, is played in the background. Subsequently, the piece becomes the "national anthem of their love", strongly linked to positive emotions of tenderness and longing. After breakup of the liaison, however, listening to the piece produces intense negative emotions in Swann, feelings of distress, melancholy and hatred (Proust 2004). Here, associations of music with significant non-musical life events cause contrary emotions induced by the same piece of music. It should be mentioned, however, that the associative memories linked to music are less frequent than usually assumed: In a retrospective autobiographic study, Schulkind et al. (1999) could demonstrate that only 10 % of the greatest hits of the last 60 years were linked to specific episodic memories.

Emotional contagion is based on the idea of a sympathetic response to music invoking sad feelings in the presence of sad music (e.g. Levinson 1996). Music induced emotions via visual imagery can best be exemplified in opera and film-music, linking specific melodies or instruments to emotionally charged scenes or personalities. A suitable example is the mouth organ melody in the movie "Once upon a time in the west" by Sergio Leone, linking the chromatic tune to emotions of gloomy suspense and revenge personified by the actor Charles Bronson.

Finally, with respect to the expectancies that are fulfilled or denied, Meyer (1956) identifies the building up of tension and subsequent relaxation as a major component of emotional appreciation of music. Recently, Huron (2006) has refined this idea in his book "Sweet anticipation". Here, he develops the ITPRA (Imagination-Tension-Prediction-Response-Appraisal) theory. He identifies five expectancy responses towards music. Two occur before the onset of the event and three afterwards. The first is the "imagination response", which consists of the prediction of what will happen and how will the listener feel when the musical event takes place. The second is the "tension response", which refers to the mental and physiological preparation immediately before the onset of the event. After the event, the "prediction response" is based on the pleasure or displeasure depending on the degree of accuracy of the prediction. Furthermore, listeners evaluate the pleasantness or unpleasantness of the outcome in the "reaction response". Finally, in the "appraisal response", the conscious evaluation of the events occurs. According to Huron, the entire process can lead to specific affective responses. When expectancies are met, music listeners get a certain degree of pleasure which is reinforced if the event and its evaluation are considered positive. If expectancies remain unfulfilled, this does not necessarily lead to negative emotions; rather the result may be laughter, awe or chill-responses: strong emotions that are frequently accompanied by physiological responses of the autonomous nervous system as will be specified below (for a recent update of this theory see also: Huron and Margulies 2010).

Coming back to the question of the adaptational value of music induced emotions, it is reasonable to distinguish strong emotions, leading to the above-mentioned physiological responses from "aesthetic" emotions (Scherer 2005). Scherer groups emotions into two classes, namely utilitarian emotions, such as for example the prototypical emotions anger, disgust, fear, happiness, sadness, surprise (Ekman 1994) and aesthetic emotions. While the former can be objectively assessed by psychophysiological measures and have clear adaptational value in terms of fitness relevant behavior, the latter are characterized by a strong subjective feeling component. Their behavioral and physiological components remain frequently obscure and the emotional responses are highly individual. Zentner et al. (2008) have analyzed the vocabulary used in self-reports of aesthetic emotions induced by music. They were able to group the common affective responses into one of nine categories: Wonder, Transcendence, Tenderness, Nostalgia, Peacefulness, Joyful activation, Tension, and Sadness. It is difficult to attribute an adaptational value to these highly elaborated feelings, although they clearly are beneficial for human wellbeing, adding meaning, consolation and security to our lives. Thus, aesthetic emotions are good candidates as a human invention forming parts of a TTM.

5 The Chill-Response in Music as an Example of Emotional Peak Experience: Phenomenology and Contributing Factors

"Chills", "thrills", or "Shivers down the spine", terms used inter-changeably, combined with goose-bumps occur in many contexts and are elicited in different sensory domains. Physiologically, the chill-response is a consequence of the activation of the sympathetic nervous system. This activation induces the hair to erect, event caused by a contraction of the minuscule arrectores pilorum muscles in the skin. Furthermore, chills are frequently accompanied by other reactions of the sympathetic nervous system, for example increase in heart rate, blood pressure, breathing rate, and sweat production measured by the galvanic skin response. As already mentioned above, chills are linked to dopaminergic activation (Salimpoor et al. 2011), increase in arousal and motivation thus supporting memory formation (for a review see Mc Gaugh 2006). In this way, events leading to a chill response will be memorized more precisely and for longer time. This fact is important as we will be later considering the adaptational value of chill-responses in music.

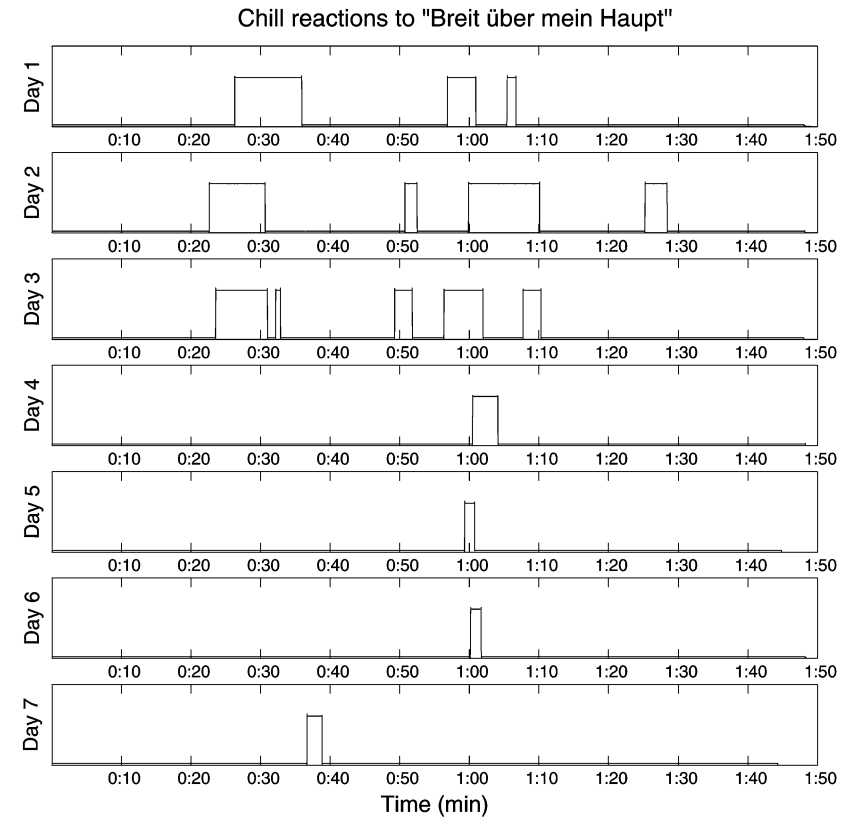
The chill-response seems to be common in furred mammals and occurs as a response to cold or to anger and fear. In the former case, erect hairs trap air close to the warm body surface and create a layer of insulation (Campbell 1996). In the latter, erect hairs make the animal appear larger in order to frighten enemies. This can be observed in the intimidation displays of chimpanzees, in stressed mice and rats, and in frightened cats, but also in the course of courtship of male chimpanzees

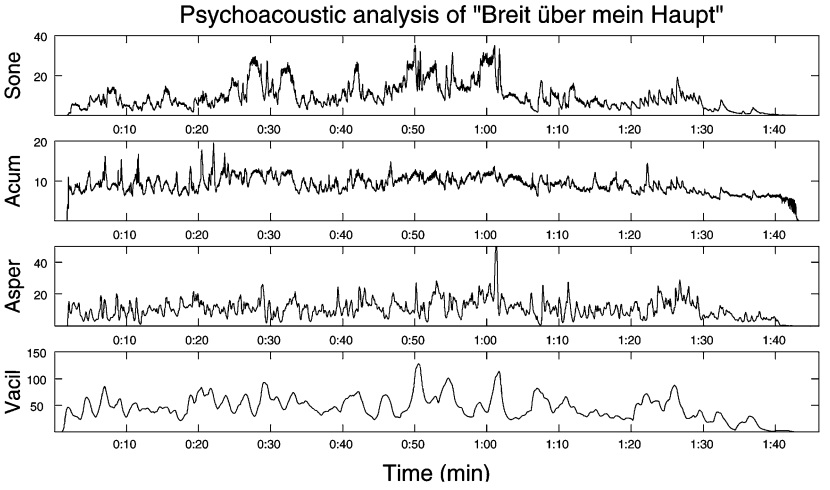
(Nishida 1997, for a review see Kreibig 2010). A special case of acoustically invoked chills in mammals seems to be in response to maternal separation calls in some monkey species. Panksepp (1995) argues that feelings of social loss and social coldness in the offspring could thus be soothed by maternal vocalizations. In his opinion this could explain why, in humans, chills are frequently perceived in the presence of sad or bitter-sweet emotions (Benedek and Kaernbach 2011). Critically, it should be noted that no systematic study on the frequency, time course, and intensity of chill-responses to separation calls in non-human primates exists. Therefore, evidence for such a mechanisms remains, albeit frequently quoted, anecdotal and scientifically ill founded.

In humans, chills can be induced through aural, visual, somatosensory, gustatory, and enteroceptive stimulation (Grewe et al. 2010). Although most research has focused on the above mentioned music evoked chills—which are in most instances linked to pleasurable and joyful, albeit sometimes nostalgic feelings (Guhn et al. 2007, Grewe et al. 2007b) it should not be forgotten that aversive acoustic stimulation, such as the scraping sound of chalk on a blackboard or a dentist's drill, can induce such chill responses even more reliably (Grewe et al. 2010). These aversive sounds are characterized by high intensity, high pitch, and frequently high degree of roughness in psychoacoustic terms.

In the somatosensory domain chills are evoked by cold, as a thermoregulatory reflex, and by tactile stimuli. The latter are frequently perceived as pleasurable and are probably linked to grooming and sexual arousal, although research on this topic is lacking. Gustatory chills are evoked by sour and spicy food, and visual chills by aesthetic objects and feelings of awe (Konecni 2011), but also by viewing highly aversive pictures (Grewe et al. 2010). Finally, chills are frequently elicited by mere mental self-stimulation, thoughts of pleasure and emotional memories, including musical ones. All these highly diverse chill responses have similar physiological correlates, as assessed by measurements of skin conductance response, increases in heart rates and breathing rates (Grewe et al. 2010), and thus cannot be distinguished by psychophysiological laboratory parameters. In the following, we will focus on the "positive" chill response to music linked to pleasurable feelings. We will briefly summarize our findings on musical parameters and listeners' characteristics contributing to evoking these bodily reactions.

Positive chill responses and emotional peak experiences in music are rare events. According to Goldstein (1980), about 70 % of the general population are familiar with these reactions. Interestingly, there are differences between occupational groups. Music students are with up to 90 % more susceptible to chills as medical students (80 %), and employees of an addiction research center (53 %). In a preselected and susceptible group of avid music lovers and amateur choir singers, only 72 % had a chill response when listening to emotionally arousing music for half an hour in a laboratory setting (Grewe et al. 2009a). It should be noted that these responses are fragile and not perfectly reproduced when playing the same musical passages on different days, even in individuals with high "chill-susceptibility" (see Fig. 3).


Chill reactions to "Breit über mein Haupt"
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Day 7
0:10 0:20 0:30 0:40 0:50 1:00 1:10 1:20 1:30 1:40 1:50
Time (min)


Psychoacoustic analysis of "Breit über mein Haupt"
Sone
Acum
Asper
Vacil
0:10 0:20 0:30 0:40 0:50 1:00 1:10 1:20 1:30 1:40
Time (min)

◀ **Fig. 3** Chill response data (*top panel*) for one participant across seven days in response to Strauss' "Breit über mein Haupt" with accompanying psychoacoustic analysis presented in the *bottom panel*. Chill response consistency is evident at t = 1 min. Psychoacoustic parameters loudness, roughness, and fluctuation show peaks at this point in time. However, the other chill responses vary considerably the seven days with a general tendency to habituation. From Grewe et al. 2007a, with permission

Furthermore, they strongly depend on the context. For example in an experiment comparing the effects of listening to favorite "chill-producing" music alone or in a group with friends, less chills occurred in the group condition, pointing to another interesting facet of the chill-response, at least in our western culture: Chills are frequently perceived as intimate and even linked to a sense of shame (Egermann et al. 2011).

In a series of studies we attempted to identify musical factors such as structural characteristics, harmonic progressions, timbre of instruments/voices and loudness developments contributing to elicit a chill response. The results were quite disillusioning. First, there was no simple stimulus–response relationship, i.e. even in music believed to be highly emotionally arousing the chill-responses remain rather casual and not simply reproducible. Secondly, there was no combination of musical factors producing chills in a fairly reliable manner. This was already demonstrated by Guhn et al. (2007), who strived to maximize chill-responses in listeners by experimenter selected "chill-music", unfamiliar to the subjects. Only 29–35 % of the subjects perceived chills in the respective passages from works of Mozart, Chopin and Bruch. In our experiments, the only general factor identifiable as a necessary, however not sufficient, condition to induce a chill-response was a change in musical structure, or, in the terminology of David Huron's ITPRA-model, a non-fulfillment of expectancies (Grewe et al. 2007b).

In a group of 38 quite heterogeneous subjects (age-range 11–72 years, 29 females, five professional musicians, 20 amateur musicians, and 13 non-musicians) we analyzed musical parameters of their favorite music, producing a chill-response in the laboratory. In 29 % of the pieces, the entry of a voice, irrespectively of whether it was human or of an instrument, could be identified (Grewe et al. 2007b). Furthermore, in 19 % a peak in loudness and in 14 % a peak in sharpness was found. When looking closer to the latter parameters, the increase in loudness was prominent in the high register (between 920 and 4,400 Hz), thus contributing directly to the parameter "sharpness". Less salient was the increase in roughness. In 12 % of the chill responses an increase in roughness linked to a reduced tone/noise ratio was observed. The latter reflects an increase in acoustic "density" (Grewe et al. 2007b; Nagel et al. 2008). Transferred to music, this occurs for example when more instruments are playing or more voices are singing with higher loudness and tempo. Behaviorally, all these acoustic changes are accompanied by an increase in arousal, which was confirmed in real time self-reports using the device "EmuJoy" that allowed to monitor the self-declaration of felt valence and arousal on a two dimensional coordinate system (Nagel et al. 2007). A typical example including all the above-mentioned criteria is the

"Barrabas-Call" in St. Matthew's passion of Johann Sebastian Bach. Interestingly, this example was the most frequently quoted in John Sloboda's first pilot study on strong emotions when listening to music (Sloboda 1991). However, the chill response in the "Barrabas-Call" is not reflex-like, since it varies depending on many factors, such as the listening situation of the individual, overall wellbeing, attentional factors, and day-form (Grewe et al. 2009b).

With respect to the listeners' factors in the above mentioned heterogeneous group, strong chill responders differed from those not perceiving chill-responses in several respects: they were more familiar with classical music, rated music as more important for their lives, identified more with the music they preferred, and listened more readily to music in everyday life (Grewe et al. 2007b). Of course, it could be discussed further, whether these features are a consequence of the participant's proneness to pleasurable experiences in music or whether they contribute a priori to a higher susceptibility for chill-responses in music. Concerning psychological traits, chill responders showed a general tendency for less intensive stimuli, as operationalized by Zuckerman's sensation seeking questionnaire (Litle and Zuckermann 1986) and were more reward dependent, i.e. they especially liked approval and positive emotional input from their environment.

Since familiarity with musical genre and personal emotional memories seemed to be important factors in the production of chill responses, we addressed the role of the individual musical biography in a further experiment (Grewe et al. 2009b). The goal of this study was to induce chill responses more reliably and to gain further insights into the factors influencing it. We recruited 54 subjects from three different amateur choirs who had performed Mozart's requiem, further referred to as "Mozart-group", and 41 participants from gospel and pop choirs, further referred to as "control-group", who were unfamiliar with the Mozart requiem and with classical music in general. We exposed these subjects to emotionally moving excerpts from Mozart's requiem (Lacrimosa, Confutatis, Rex tremendae, Tuba mirum, Dies irae), which were either recordings of themselves or of professionals. Furthermore, excerpts from the Requiem of Puccini ad from the Bach Motet "Our live is a shadow", which had been sung in each case by only one of the three choirs of the Mozart-group, were played. As measurements, subjective real-time rating of the intensity of the feelings, and perceived chill-responses were recorded using the software EMuJoy (Nagel et al. 2007). Additionally psychophysical measures such as skin conduction response (SCR), skin conduction level (SCL), heart rate (HR), and breathing rate (BR) were assessed. Figure 4 shows the time course of psychophysiological data 10 s prior, during and 10 s after the chill response in a grand average. The two most salient features of the physiological responses are (a) the increase of SCR about 2 s before the chill reaction and (b) the response after the chill of about 4–6 s which has recently been called the "afterglow effect" (Schubert in press).

Overall, comparable to previous results of Goldstein and Guhn, only about two thirds of the participants reported a minimum of one chill during the experiment. There was high variability in chill responses, ranging from a maximum of 88 chill-responses in one subject to no chills at all in others. On average, each participant

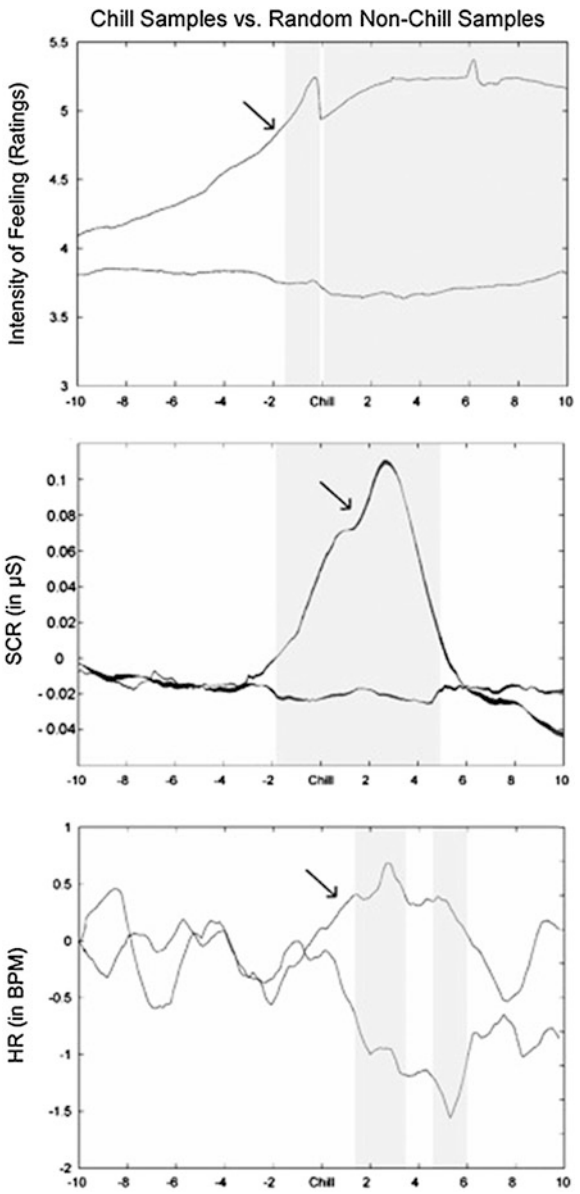


Fig. 4 Comparison of 622 chill samples (*upper line* marked with *arrow*) with 622 random non-chill samples (*lower line*). *Grey shaded* areas indicate significant differences (Random Permutation Test). From the curves it becomes clear that self-declared intensity of feeling, skin response (*SCR*) and heart rate (*HR*) start about two seconds before the individual chill-response, which is marked by an *arrow*. A salient characteristic of the physiological response (SCR and HR) is the afterglow effect, which lasts about 4–6 s after the chill event. Modified from Grewe et al. 2009a

experienced 9 chills during the experiment. Interestingly, chill-responses showed no relation to age, gender or knowing and liking of classical music. However, familiarity with the stimuli influenced the frequency of chill-responses. Chills occurred more frequently in the Mozart-group than in the control-group (72 vs. 56 % of the participants), and the overall number of chills was much higher in the former than in the latter (679 vs. 173 chill-responses). Furthermore, whilst listening

to the Bach motet and the Puccini excerpts, chill-responses were significantly more frequent in the choir-members familiar with the respective pieces. However, it seems not to be very important to listen to one's own interpretation, since only the Confutatis interpretation of one choir produced slightly more chill responses in the choir-members as compared to the professional version (in average 0.95 chills vs. 0.11 chills). Thus, obviously, familiarity with the stimulus is an important factor in eliciting chill-responses. Musical biography and individual associations, for example the remembrances of a successful performance in an awe inspiring gothic cathedral, may well promote emotional susceptibility.

Summarizing this paragraph with respect to the overall topic of this review, namely the evolutionary adaptive value of music, the chill-response is biologically grounded in an ancient reflex-like response of the sympathetic nervous system related to thermoregulation and intimidation displays. It is biologically linked to arousal and facilitates memory formation. In humans, the chill response occurs in the auditory domain in the context of negative arousal and alarm, mainly linked to aversive loud and high frequency noise, and in the context of highly pleasurable events leading to activation of the dopaminergic reward system in the brain. Factors facilitating these positive chill-responses include: structural changes, beginning of something new, increase in loudness in the high register and personal emotional memories linked to positive associations and liking of the music. Chill-responses are more frequent in more sensitive and social personalities. In the following last paragraph we will demonstrate how the chill-response may be linked to an adaptive value of music in human evolution. Lastly, we will develop our model of "mixed origins of music" in human evolution.

6 The Mixed Origins of Music Theory (MOM-Theory): Evidence from the Chill-Response

The evolutionary adaptive value of the chill-response is at hand when considering the above-mentioned biological concomitants. Negative chill-responses may have been direct, arousing reactions towards the piercing sounds of a hunting predator or by the shrieking calls of conspecifics attacked by an enemy or predator (Owren and Rendall 2001). They may be part of an evolutionary ancient inter- and intraspecific affective signaling system of alarm calls and pain shrieks, observable today in many socially living mammals, for example in tree-shrews and vervet monkeys. These sounds furthermore support avoidance behavior in order to increase the distance to the sound source. In this way, close contact to a potentially dangerous predator is prevented, but also the delicate sensory organ of the cochlea with its hair cells, susceptible to high sound pressure levels damage, is protected (Subramaniam et al. 1995). Finally, in agonistic contexts, an intimidation display is activated to frighten the predator. In human evolution, the roots of such a behavior may well date back to some 3.2 million years when we, or better our

about one meter tall evolutionary ancestors Australopithecus afarensis, prowled through the high grass of the Central African dry steppes, haunted by the piercing sounds of some large African eagles, hunting for prey.

Explaining the evolutionary adaptive value of the positive chill response is more difficult. There are two proven and one hypothetical features, which can be related to an adaptive value of music. First, it is the fact that "surprise", for example the above-mentioned "not fulfillment of expectancies", contributes regularly to the chill-response. Since accompanying arousal and activation of the reward system improves memory formation, this kind of acoustic stimulation may well have enlarged the repertoire of auditory patterns remembered by our ancestors and furthermore fuelled curiosity to detect novel auditory stimuli. This, in turn, was of evolutionary relevance, since fast and precise classification of acoustic stimuli was a prerequisite for optimal adaptations of behavior (for example avoidance of a sneaking predator at night and perception of subtle nuances of intraspecific affect vocalizations). We therefore speculate that one of the driving forces of the development of our superior auditory memory was the reward gained by identification of novel acoustic patterns. We will even go further claiming that the first song-like vocalizations, the first artificial sounds produced by primitive instruments (for example wooden drumsticks hit on hollow stumps) may have constituted a safe playground to train auditory discrimination. Furthermore, vocal production abilities improved and reinforced curiosity to detect novel sounds long before language emerged, thus establishing prerequisites to develop the latter.

The second feature of the positive chill response implicating an evolutionary adaptive value is pleasure induction by music. Rooted in the activation of the sympathetic nervous system and of central nervous reward circuits, music as a transformative technology of the mind (TTM) could add moments of happiness and comfort to the hard lives of early modern humans living in hostile environments. The Hohle Fels and Geissenklösterle caves for example were located in alpine tundra at the time the flutes were constructed, 35,000 years ago. Average temperatures were comparable to present day Greenland. Albeit food was readily available, due to the rich wildlife, musculoskeletal diseases, gastro-intestinal infections, parasites, toothache and the omnipresent cold rendered life cumbersome. Music could provide moments of wellbeing, of forgetting the daily hardship not only by producing aesthetic emotions but also by giving rise to occasional emotional peak experiences, which then reawakened love of life.

Finally, the third potential feature with evolutionary adaptive value is the frequently quoted "separation call" theory by Panksepp (1995). It proposes that the evolutionary origin of music induced positive chill-responses is a soothing and "warming" function of maternal monkey vocalizations on the offspring. Unfortunately, this theory has not yet been verified by empirical research. An argument speaking against such a mechanism is the lacking evidence of acoustically evoked chill-responses in infants and toddlers, for example when listening to soothing lullabies. Possibly, such a phenomenon may have been overlooked up to now. However, in informal interviews with children and adolescents, it seems that the

earliest descriptions of positive chill-responses emerge just before reaching puberty. Admittedly, systematic empirical research on this interesting topic is missing.

In short, when hypothetically summarizing the long history of chill-responses, we argue that in the beginning it was predominantly related to a reflex-like mechanism, involved in thermoregulation. This was based on neuronal networks of the autonomous nervous system, involving thermoceptive afferents from the skin and efferent activation of the sympathetic nuclei. These reflexes were additionally activated by exposure to aversive stimuli, such as shrieking sounds, sour food, or enteroceptive pain, producing a threatening display and enlarging the appearance by hair-raising. Due to conditioned reflexes, the trigger of such a threatening display could be modified by learning and vice versa the chill-response activated memory formation. Finally, after the human-specific loss of fur, chill-responses in the context of thermoregulation and threatening displays became biologically meaningless and could be used for other purposes. The acoustically mediated positive chill-responses, previously reflex-like and conditioned, could be powerfully used for auditory learning. Rewarding new and surprising acoustic stimuli with chill-responses accompanied by endorphin- and dopamine-release and subjective feelings of wellbeing could be the most important driving force of auditory learning, constituting a prerequisite of differentiated communication behavior of the socially living early humans.

What are then the origins of music and when did music start to be part of our human condition? In the following, we will expose our "mixed origins of music-theory" or in short "MOM-theory". This will be achieved by integrating several aspects of ancient evolutionary adaptive, or later acquired, or recently refined properties of music. We are aware that this theory, as many other theories in evolution and anthropology cannot be directly proven, since there are no records of musical activities until the first human made instruments appeared. However, we strive to strengthen our arguments by drawing on a comparative approach when possible and by referring to physiological and neuronal adaptations which most probably date back long in our phylogenetic history.

As we have exemplified above with the chill-response, we argue that music may have several roots in human evolution, some dating back to many millions of years and some acquired in later times possibly as a TTM, comparable to the invention of fire (Brown et al. 2009). All these possible origins of music are not mutually exclusive; rather, they demonstrate its richness and multi-faceted nature. The many roots of music may well explain the many effects music can have in humans. In Fig. 5, we provide a scheme of the putative development of music out of an ancient affective signaling system, elaborating our MOM-theory.

In the very beginning, intraspecific and interspecific affective communication amongst pre-humans included shrieking calls of threatened conspecifics and of alarm calls of threatening predators, producing a heightened arousal, which may have been accompanied by aversive chill-reactions as it can be witnessed today in many socially living mammals. An important step in evolution must have been the generalization of these chill-reactions to affiliative sounds and vocalizations with positive emotional valence. These may have been related to parent-offspring

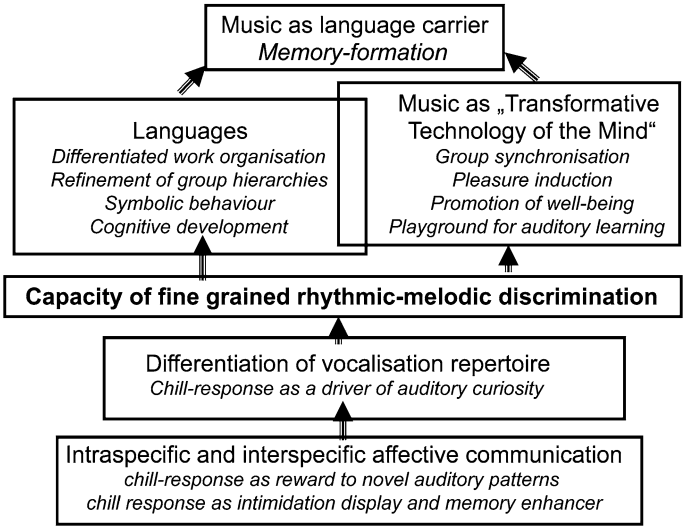


Fig. 5 Schematic display of the mixed origins of music theory (*MOM* theory). For detailed explanation see text

communication, although experimental evidence for the separation-call theory is still lacking. However, irrespectively of the emotional valence induced, the chill-reaction fostered auditory memory formation by activating the brain's reward systems when new acoustic patterns were perceived. It is unclear whether vocalizations, probably some hundred thousand years ago, were related to music in the narrow sense of the word or whether they were part of an ancient affective signaling system comparable to Mithens (2005, p. 172) "hmmmm"- proto language, which means that music was part of a communication system which was holistic, multi-modal, musical and mimetic. However, the brain's reward system activation and improved memory consolidation linked to detection of violations of auditory expectancies may have triggered the superior auditory discrimination faculty of early humans, which in turn constituted the prerequisite of language acquisition and production. Here, we argue that chill-responses to novel melodic contours, timing subtleties, timbre variations and structural breaks lead to superior classification abilities and in turn to a large repertoire of language like vocalizations, apt to replace for example manual gesturing as a means to organize distributed labor in complex social groups (e.g. Corballis 1992).

In parallel with this very fundamental aspect of human auditory learning, music also contains aspects of a TTM, which may have developed much later than language. Besides the above mentioned positive chill-response as a source of pleasure when exposed to refined tunes and playful manipulation of increasingly complex melodies, harmonies and rhythms, other facets of music are good candidates for a TTM: As Patel (2010) exemplified, group cohesion and co-operative behavior are supported by joint clapping and dancing, relying on the human ability

to synchronize in time with an external beat. According to Patel, this capacity is related to vocal learning, and may thus be an epiphenomenon of our language abilities. However, as demonstrated above, our language capacity in turn may be grounded in our auditory classification abilities facilitated by the chill-responses.

A more convincing musical TTM aspect is the processing of musical hierarchies in tonality. These cognitive processes rely partly on the same neuronal resources as syntactic language processing. Brain activation studies showed overlapping neuronal networks when violating either harmonic musical rules or linguistic syntactic rules (for a review, see Kölsch and Siebel 2005). According to Patel, both tonal music and language rely on the same basic cognitive operations, namely the general building of mental hierarchies. Thus, our ability to identify and create tonal hierarchies in music and syntax in language emerges from a common evolutionary old capacity, which is on one side transformed into linguistic syntax and on the other into musical tonal hierarchies. Here, a word of caution is necessary since many forms of music such as Minimal Music, Rap, or many types of Ethnic music in Africa, do not contain tonal hierarchies (for a review, see Stevens and Byron 2010). On the other hand, hierarchies and rules, whether tonal or temporal, are almost universally found in both music and language and thus may indeed share a common "rule detector" mechanism in the brain, which is ancient and evolutionary adapted (Brown and Jordania 2011).

Many other effects of music may additionally be considered as constituents of an evolutionary late acquired TTM. To name but a few, the role of music in improving health status and rehabilitation in stroke patients and patients with basal ganglia disorders may serve as example (e.g. Thaut et al. 2001; Särkämo et al. 2008; Schneider et al. 2010). Similarly, improving memory functions in Alzheimer patients (e.g. Vink et al. 2003) and supporting memory consolidation of complex linguistic contents in healthy individuals (Wallace et al. 1994) may be regarded as a positive side effect of such a musical TTM.

In summary, we argue that on the basis of a very ancient affective communication system, auditory learning was rewarded and lead to an increasing refinement of auditory discrimination abilities in pitch and timing. This may have laid the ground to acquisition of language and also to our love of music, which in turn constituted a safe playground for new auditory experiences. Later, music was adapted for many social functions, increasing our chances of survival by better organising groups and by adding pleasure and aesthetic emotions to our hard lives.

This review is not exhaustive and for the sake of brevity many aspects of our MOM-theory could not be considered sufficiently and may be discussed critically. For instance, we did not comment on the phenomenon of congenital amusia, a condition that strongly supports the existence of evolutionary old, specialized neuronal networks, designed for refined pitch discrimination (for a review see Ayotte et al. 2002). With respect to the positive chill-response we admit that in present times this phenomenon is highly individual, linked to personal memories, and even unknown to about 30 % of the western population. Furthermore, it is usually elicited by highly complex acoustic patterns, such as a Bruckner symphony or a Beatles song. We do not know how flute tunes were experienced by our

ancestors in the Hohle Fels and Geissenklösterle, but we believe it is reasonable to assume that at times of low exposure to music, and artificially produced sound in general, even a simple monophonic tune could have a strong emotional impact. Another open question is whether the positive chill-response is a musical universal or whether it is predominantly linked to variations in the pitch domain, restricted to a limited number of music cultures. If positive chills were not universally found, this would weaken our argument of an evolutionary ancient emotional reaction. This brings us to another perspective, namely that the chill-response may be a consequence of our modern way to listen to music seated in a chair without the possibility to move bodily to the rhythms, comparable to a "sublimation" of our natural urge to move to music. Systematic research on the impact of bodily movements on the positive chill-response is still lacking.

To end with, music as an immensely rich human experience contains many facets and may have many effects:

Orpheus with his lute made trees,
And the mountain tops that freeze,
Bow themselves, when he did sing:
To his music plants and flowers
Ever sprung; as sun and showers
There had made a lasting spring.
Every thing that heard him play,
Even the billows of the sea,
Hung their heads, and then lay by.
In sweet music is such art,
Killing care and grief of heart
Fall asleep, or hearing, die.
(William Shakespeare, Henry VIII, 3.1.4-15).

Acknowledgments This work was supported by the DFG (Al 269-6). Furthermore we would like to thank the many friends and colleagues who gave valuable input to this paper in many discussions on the origins of music. Here, we would like to thank especially Dr. Andrè Lee, Dr. Thomas Fritz, M.Sci. Floris van Vugt, Prof. Dr. Elke Zimmermann, PD. Dr. Sabine Schmidt and the members of the IMMM. We furthermore would like to thank Marta Beauchamp for careful language editing.

References

Altenmüller, E., & Kopiez, R. (2005). Schauer und Tränen: zur Neurobiologie der durch Musik ausgelösten Emotionen. In: C. Bullerjahn, H. Gembris & A.C. Lehmann (Eds.), *Musik: gehört, gesehen und erlebt. Festschrift Klaus-Ernst Behne zum* 65. *Geburtstag*, 159–180. Monografien des IfMPF, 12, Verlag der Hochschule für Musik und Theater Hannover: Hannover.

Anati, E. (1996). Die Felskunst in Europa. In E. Anati (Ed.), *Die Höhlenmalerei* (pp. 238–240). Düsseldorf: Patmos Verlag.

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain, 125*, 238–251.

Bangert, M., & Schlaug, G. (2006). Specialization of the specialized in features of external brain morphology. *European Journal of Neuroscience, 24*, 1832–1834.

Benedek, M., & Kaernbach, C. (2011). Physiological correlates and emotional specificity of human piloerection. *Biological Psychology, 86*, 320–329.

Blood, A. J., & Zatorrre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *PNAS, 98*, 11818–11823.

Brown, S. (2000). The 'music language' model of music evolution. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 271–300). Cambridge: MIT Press.

Brown, S., & Jordania, J. (2011). Universals in the world's music. *Psychology of Music,*. doi:10.1177/0305735611425896.

Brown, K. S., Marean, C. W., Herries, A. I. R., Jacobs, Z., Tribolo, C., Braun, D., et al. (2009). Fire as an engineering tool of early modern humans. *Science, 325*, 859–862.

Campbell, N. A. (1996). *Biology* (4th ed.). Menlo Park CA: Benjamin/Cummings Publishers.

Conard, N.J., & Malina, M. (2008). New evidence for the origins of music from caves of the Swabian Jura. In: A. A. Both, R. Eichmann, E. Hickmann & L.-CH. Koch (Eds.). *Orient-Archäologie Band 22. Studien zur Musikarchäologie VI* (pp. 13–22). Rahden:Verlag Marie Leidorf GmbH.

Conard, N. J., Malina, M., & Münzel, S. C. (2009). New flutes document the earliest musical tradition in southwestern Germany. *Nature, 460*, 737–740.

Corballis, M. C. (1992). On the evolution of language and generativity. *Cognition, 44*, 197–226.

Cross, I. (2008). Musicality and the human capacity for culture. *Musicae Scientiae*, (*Special Issue*: *Narrative in Music and Interaction*), *12*, 147–167.

Cross, I. (2009). The evolutionary meaning of musical meaning. *Musicae Scientiae, 13*, 179–200.

Darwin, C. (1871). The descent of man, and selection in relation to sex. In: E.O. Wilson (Ed.). *From so simple a beginning: The four great books of Charles Darwin*. Reprint 2006. New York: W.W. Norton.

Egermann, H., Sutherland, M. E., Grewe, O., Nagel, F., Kopiez, R., & Altenmüller, E. (2011). The influences of a group setting on the experience of music: A physiological and psychological perspective on emotion. *Musicae Scientiae, 15*, 307–323.

Ekman, P., & Davidson, R. J. (1994). *The nature of emotion (fundamental questions)*. Oxford: Oxford University Press.

Evers, S., & Suhr, B. (2000). Changes of the neurotransmitter serotonin but not of hormones during short time music perception. *European Archives in Psychiatry and Clinical Neurosciences, 250*, 144–147.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., et al. (2009). Universal recognition of three basic emotions in music. *Current Biology, 19*, 573–576.

Goldstein, A. (1980). Thrills in response to music and other stimuli. *Physiological Psychology, 8*, 126–129.

Grewe, O., Nagel, F., Kopiez, R., & Altenmüller, E. (2007a). Listening to music as a re-creative process: Physiological, psychological and psychoacoustical correlates of chills and strong emotions. *Music Perception, 24*, 297–314.

Grewe, O., Nagel, F., Kopiez, R., & Altenmüller, E. (2007b). Emotions over time. Synchronicity and development of subjective, physiological and mimic affective reactions to music. *Emotions, 7*, 774–788.

Grewe, O., Kopiez, R., & Altenmüller, E. (2009a). The chill parameter: Goose bumps and shivers as promising measures in emotion research. *Music Perception, 27*, 61–74.

Grewe, O., Altenmüller, E., Nagel, F., & Kopiez, R. (2009b). Evolutionary-based universals? A discussion of individual emotional reactions towards music. *Musicae Scientiae, 13*, 261–287.

Grewe, O., Katzur, B., Kopiez, R., & Altenmüller, E. (2010). Chills in different sensory domains: Frisson elicited by acoustical, visual, tactile and gustatory stimuli. *Psychology of Music, 39*, 220–239.

Guhn, M., Hamm, A., & Zentner, M. R. (2007). Physiological and musico-acoustic correlates of the chill response. *Music Perception, 24*, 473–483.

Herder, J.G. (1772). *Über den Ursprung der Sprache*. Berlin: Christian Friedrich Voss.

Hunter, P., Schellenberg, G. (2010). Music and emotion. In: A.N. Popper, R.R. Fay & M.R. Jones (Eds.). *Music perception.* Handbook of auditory research (Vol. 36, pp. 129–164). New York: Springer.

Huron, D. (2006). *Sweet anticipation: music and the psychology of expectation*. Cambridge, MA: A Bradford Book.

Huron, D., & Hellmuth Margulis, E. (2010). Musical expectancy and thrills. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 575–604). Oxford: Oxford University Press.

Hyde, K. L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A. C., et al. (2009). Musical training shapes structural brain development. *Journal of Neuroscience, 29*, 3019–3025.

James, W. (1890). *The principles of psychology*. New York: Dover Publications.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioural and Brain Sciences, 31*, 559–621.

Kämpfe, J., Sedlmeier, P., & Renkewitz, F. (2011). The impact of background music on adult listeners: A meta-analysis. *Psychology of Music, 39*, 424–448.

Karabanov, A., Cervenka, S., de Manzano, O., Forssberg, H., Farde, L., & Ullén, F. (2010). Dopamine D2 receptor density in the limbic striatum is related to implicit but not explicit movement sequence learning. *PNAS, 107*, 7574–7579.

Kirschner, S., & Tomasello, M. (2010). Joint music making promotes prosocial behavior in 4-year-old children. *Evolution and Human Behavior, 31*, 354–364.

Kivy, P. (1990). *Music alone: Philosophical reflections on the purely musical experience*. Ithaca: Cornell University Press.

Konecni, V. J. (2011). Aesthetic trinity theory and the sublime. *Philosophy Today, 5*, 64–73.

Koelsch, S., & Siebel, W. (2005). Towards a neural basis of music perception. *Trends Cogn. Sci., 9*, 578–584.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology, 84*, 394–421.

Krumhansl, C. L., & Cuddy, L. L. (2010). A theory of tonal hierarchies in music. In A. N. Popper, R. R. Fay, & M. R. Jones (Eds.), *Music Perception: Springer handbook of auditory research* (36th ed., pp. 51–87). New York: Springer.

Levinson, J. (1996). *The pleasures of aesthetics: Philosophical essays*. Ithaka: Cornell University Press.

Liang, T. Z. (2002). Prähistorische Knochenflöten und ihre Bedeutung für die Revision der chinesischen Musikgeschichte. In E. Hickmann, A. D. Kilmer, & R. Eichmann (Eds.), *Studien zur Musikarchäologie* (Vol. III, pp. 155–160). Rahden: Verlag Marie Leidorf GmbH.

Litle, P., & Zuckerman, M. (1986). Sensation seeking and music preferences. *Personality and Individual Differences, 7*, 575–577.

Mampe, B., Friederici, A. D., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology, 19*, 1–4.

McGaugh, J. L. (2006). Make mild moments memorable: add a little arousal. *Trends Cogn. Sci., 10*, 345–347.

McMullen, E., & Saffran, J. R. (2004). Music and language: A developmental comparison. *Music Perception, 21*, 289–311.

McNeill, W. H. (1995). *Keeping together in time. Dance and drill in human history*. Cambridge: Harvard University Press.

Menon, V., & Levitin, D. J. (2005). The rewards of music listening: response and physiological connectivity of the mesolimbic system. *Neuroimage, 28*, 175–184.

Meyer, L. B. (1956). *Emotions and meaning in music*. London: The University of Chicago Press.

Miller, G. (2000). Evolution of human music through sexual selection. In N. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 315–328). Cambridge: MIT-Press.

Mithen, S. (2005) *The singing Neanderthals*. London: Weidenfeld & Nicholson.

Münte, T. F., Altenmüller, E., & Jäncke, L. (2002). The musician's brain as a model of neuroplasticity. *Nature Neuroscience, 3*, 473–478.

Münzel, S.C., & Conard, N. (2009). Flötenklang aus fernen Zeiten. Die frühesten Musikinstrumente. In: *Eiszeit. Kunst und Kultur. Begleitband zur großen Landesausstellung* (pp. 317–321). Hrsg. Archäologisches Landesmuseum Baden Württemberg.

Münzel, S. C., Seeberger, F., & Hein, W. (2002). The Geißenklösterle Flute–discovery, experiments, reconstruction. In: E. Hickmann, A. D. Kilmer, & R. Eichmann (Eds.). *Studien zur Musikarchäologie III; Archäologie früher Klangerzeugung und Tonordnung; Musikarchäologie in der Ägäis und Anatolien. Orient-Archäologie, Bd. 10* (pp. 107–110). Rahden: Verlag Marie Leidorf GmbH.

Nagel, F., Kopiez, R., Grewe, O., & Altenmüller, E. (2007). EMuJoy: software for continuous measurement of perceived emotions in music. *Behavior Research Methods, 39*, 283–290.

Nagel, F., Kopiez, R., Grewe, O., & Altenmüller, E. (2008). Psychoacoustic correlates of musically induced chills. *Musicae Scientiae, 12*, 101–113.

Nishida, T. (1997). Sexual behavior of adult male chimpanzees of the Mahale Moun- tains National Park, Tanzania. *Primates, 38*, 379–398.

Owren, M. J., & Rendall, D. (2001). Sound on the Rebound: Bringing form and function back to the forefront in understanding nonhuman primate vocal signalling. *Evolutionary Anthropology, 10*, 58–71.

Panksepp, J. (1995). The emotional sources of "chills" induced by music. *Music Perception, 13*, 171–207.

Panksepp, J., & Bernatzky, G. (2002). Emotional sounds and the brain: The neuro-affective foundations of musical appreciation. *Behavioural Processes, 60*, 133–155.

Patel, A. D. (2008). *Music, language, and the brain*. Oxford: Oxford University Press.

Patel, A. (2010). Music, biological evolution, and the brain. In M. Bailar (Ed.), *Emerging disciplines* (pp. 91–144). Huston: Huston University Press.

Patel, A. D., Iverson, J. R., Bregman, R. R., & Schultz, I. (2009). Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current Biology, 19*, 827–830.

Pinker, S. (1997). *How the mind works*. London: Allen Lane.

Proust, M. (2004). *Auf der Suche nach der verlorenen Zeit. Band 1*. Übersetzung von Eva-Rechel-Mertens (pp. 499–555). Frankfurt: Suhrkamp Taschenbuch.

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience, 14*, 257–262.

Särkämö, T., Tervaniemi, M., Laitinen, S., Forsblom, A., Soinila, S., Mikkonen, M., et al. (2008). Music listening enhances cognitive recovery and mood after middle cerebral artery stroke. *Brain, 131*, 866–876.

Scherer, K. R. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *Journal of New Music Research, 33*, 239–251.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*, 695–729.

Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. *Nature Neuroscience, 5*, 688–694.

Schneider, S., Münte, T. F., Rodriguez-Fornells, A., Sailer, M., & Altenmüller, E. (2010). Music supported training is more efficient than functional motor training for recovery of fine motor skills in stroke patients. *Music Perception, 27*, 271–280.

Schubert, E. (in press). Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music. *Psychology of music*. February 8, 2012, Doi:10.1177/0305735611430079.

Schulkind, M. D., Hennis, L. K., & Rubin, D. C. (1999). Music, emotion, and autobiographical memory: They're playing your song. *Memory and Cognition, 27*, 948–955.

Shenfield, T., Trehub, S., & Nakata, T. (2003). Maternal singing modulates infant arousal. *Psychology of Music, 31*, 365–375.

Sloboda, J. (1991). Music structure and emotional response: Some empirical findings. *Psychology of Music, 19*, 110–120.

Spencer, H. (1857). On the origin and function of music. Fraser's Magazine, Oct. 1857.

Stevens, C., & Byron, T. (2010) Universals in music processing. In: S. Hallam, I. Cross, & M. Thaut (Eds.). *Oxford handbook of psychology of music, Ch.* 2 (pp. 53–78). Oxford: Oxford University Press.

Stumpf, C. (1911). Die Anfänge der Musik. Leipzig: Barth. (trans: C. Stumpf, D. Tripett (2012)). The origins of music. Oxford: Oxford University press.

Subramaniam, M., Henselman, L. W., Spongr, V., Henderson, D., & Powers, N. L. (1995). Effect of high-frequency interrupted noise exposures on evoked-potential thresholds, distortion-product otoacoustic emissions, and outer hair cell loss. *Ear and Hearing, 16*, 372–381.

Thaut, M. H., McIntosh, K. W., McIntosh, G. C., & Hömberg, V. (2001). Auditory rhythmicity enhances movement and speech motor control in patients with Parkinson's disease. *Functional Neurology, 16*, 163–172.

Vink, A. C., Birks, J., Bruinsma, M.S., & Scholten, R.J. (2003) Music therapy for people with dementia. *Cochrane database of systematic reviews*, (Vol. 4). No.: CD003477. Retrieved February 1, 2012, DOI: 10.1002/14651858.

Wallace, W. T., Siddiqua, N., & Harun-ar-Rashid, A. K. M. (1994). Memory for music: Effects of melody on recall of text. *Journal of Experimental Psychology Learning, Memory, and Cognition, 20*, 1471–1485.

Wallin, N. L., Merker, B., & Brown, S. (2000). *The origins of music*. Cambridge: MIT Press.

Zentner, M., & Eerola, T. (2010). Rhythmic engagement with music in infancy. *PNAS, 107*, 5768–5773.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification and measurement. *Emotion, 8*, 494–521.

Music and Action

Stefan Koelsch and Clemens Maidhof

Music in the broadest sense can be defined as intentionally structured sounds (for design features of music see Fitch 2006; Koelsch 2012). This requires, that at least one individual somehow creates sounds. In most cultural practices, this is achieved by singing or playing an instrument. Such music performance includes planning, initiation, execution, monitoring, and correction of actions. This chapter deals with action-related processes during music performance. Because the production of actions and the perception of actions and their effects are not separable, the first part of this chapter deals with perception-action mediation (or "mirror mechanisms") during music listening, focussing in particular on neural correlates of perception-action mediation. The second part then reviews studies using event-related potentials (ERPs) to investigate neural correlates of music production.

1 Perception–Action Mediation

The perception of events can give rise to action-related processes. With regard to music, simply listening to music can automatically engage action-related processes. The term *action* as used here implies that an action (a) consists of at

Beitrag zum Jahrbuch Musikwissenschaft Uni-HH.
Author's note: This is an updated and adapted version from Chap. 11 (Music and Action) of Koelsch (2012).

S. Koelsch (✉)
Cluster Languages of Emotion der Freien Universität Berlin,
Habelschwerdter Allee 45, 14195 Berlin, Germany
e-mail: koelsch@cbs.mpg.de

C. Maidhof
Cognitive Brain Research Unit, Institute of Behavioural Sciences, University of Helsinki, Finland and Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Jyväskylä, Finland

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_6,

least one movement, or a chunk of movements (such as playing triad arpeggios across several octaves), (b) has a goal, (c) can be voluntarily executed (or inhibited), (d) can be corrected during execution (if necessary), and (e) is modulated by the anticipation of a specific action effect.[1] Actions can be chained into action sequences, each action having a sub-goal (and related action effects), and the action sequence having a superordinate goal (and a related action effect).

In his *common coding* approach to perception and action, Wolfgang Prinz (1990) described that actions are represented in terms of their perceptual consequences, and that the late stages of perception overlap with the early stages of action in the sense that they share a common representational format.[2] Such a common format can, e.g., be a common neuronal code. 'Common coding' is supposed to be involved when an individual perceives a movement, as well as when an individual perceives the effects of an action produced by another individual (and due to common coding, such perception evokes movement representations). Similarly, Liberman and Mattingly (1985) proposed in their *motor theory of speech perception* that, during speech perception, speech is decoded in part by the same processes that are involved in speech production. With regard to the observation of movements, Giacomo Rizzolatti and colleagues published in the 1990s reports about neurons located in the area F5 of the premotor cortex of macaque monkeys, which were not only active when the monkeys performed a movement, but also when the monkeys simply observed that movement (the so-called *mirror neurons*, reviewed in Rizzolatti and Sinigaglia 2010). For example, when a monkey observed an experimenter grasping a piece of food with his hand, neural responses were evoked in neurons located in area F5. These neurons ceased to produce action potentials when the experimenter moved the food toward the monkey, and they produced action potentials again when the monkey grasped the food. The "mirror function" of these premotor neurons is a physiological correlate of 'common coding', and fundamental for perception–action mediation. This section will provide an overview of studies on perception–action mediation during listening to music.[3]

The first neuroscientific study on auditory perception–action mediation was a study by Jens Haueisen and Thomas Knösche (Haueisen and Knösche 2001) using magnetoencephalography (MEG). In that study, both non-musicians and pianists listened to piano melodies. Compared to non-musicians, musicians showed neuronal activity in (pre)motor areas that was elicited simply by listening to music (the task was to detect wrong notes, and those trials with wrong notes were

[1] An action effect can, e.g., be a single tone, or the sounds of a triad arpeggio across several octaves.

[2] The common coding approach follows the ideomotor approaches of Hermann Lotze (1852) and Willam James (1890). For a summary of the Lotze-James account and the ideomotor framework see, e.g., Prinz (2005), p. 142–143.

[3] For an fMRI study investigating pianists and non-musicians *observing* finger-hand movements of a person playing piano see Haslinger et al. (2005).

excluded from the data analysis).[4] Interestingly, the centre of neuronal activity for notes that would usually be played with the little finger was located more superiorly than activity for notes that would usually be played with the thumb (according to the somatotopic representation of the fingers), supporting the notion that the observed neural activity was premotor activity. Similar activations were observed with functional magnetic resonance imaging (fMRI) when violinists listened to violin music (Dick et al. 2011).

One year later, Evelyne Kohler et al. (2002) investigated neurons in the area F5 of macaque monkeys that discharged not only when the monkeys performed a hand action (such as tearing a piece of paper), but also when the monkeys saw, and simultaneously heard the sound of this tearing action (similar to the mirror neurons mentioned above, that are active during both observation and execution of actions). Importantly, simply hearing the sound of the same action (performed out of the monkey's sight) was equally effective in evoking a response in these neurons. Control sounds that were not related to action (such as white noise, or monkey calls) did not evoke excitatory responses in those neurons. Thus, this study showed that (in monkeys) some premotor neurons are active during both hearing and execution of actions.

Further evidence comes from behaviorial studies, which demonstrated a close coupling between action and perception (Drost et al. 2005a, b): In a series of experiments, pianists and non-musicians were required to play different intervals or chords following corresponding visual stimuli. Simultaneously with the imperative visual stimuli, task-irrelevant sounds were presented, which could be either congruent or incongruent with the visual stimulus. For example, participants were visually instructed to play a C–E interval, but heard concurrently a different interval. Results showed that pianists, but not non-musicians, reacted more slowly to the visual stimuli when the distracting sound was incongruent with the imperative stimulus. In addition, perceived intervals could induce incorrect responses, e.g. when pianists played the heard interval instead of the instructed interval. This indicated that due to music (piano) training, pianists have acquired strong associations between movements and their resulting auditory effects (Drost et al. 2005a, b).

As mentioned above, the study by Haueisen and Knösche (2001) showed perception–action mediation in musicians (pianists). Music-related perception–action mediation in non-musicians was shown by Callan et al. (2006). In that study, activation of premotor cortex was observed not only when participants (non-musicians) were singing covertly, but also when they simply listened to song. Interestingly, premotor activity in the same area was also observed during both covert speech production and listening to speech (Fig. 1a). This showed that neural correlates of mirror mechanisms overlap strongly for music and speech perception.

[4] Neural sources were located on the crown of the precentral gyrus, thus presumably in the premotor cortex, rather than in the motor cortex, contrary to what the title of the article says.

Fig. 1 Premotor activation during listening to musical information in non-musicians. The *top panel* **a** shows areas that were activated during listening to singing, covert singing, listening to speech, and covert speech (conjunction analysis) in the study by Callan et al. (2006). Both *left* and *right premotor* activity was observed in all four conditions, showing that this area is active not only during perception of speech or music, but also during the production of speech or music. **b** shows areas that were more active during listening to trained melodies compared to listening to melodies consisting of untrained (and different) tones in the study by Lahav et al. (2007). *aSTG1* = anterior superior temporal gyrus; *PMC* = premotor cortex; *PP* = planum polare; *PT* = planum temporale; *Stp* = superior temporal plane. Modified with permission from Callan et al. (2006) and Lahav et al. (2007)

In a study on the effects of musical training on perception–action mediation in non-musicians by Lahav et al. (2007), non-musicians were trained over the course of five days to play a piano melody with their right hand. After this training period, simply listening to the trained melody activated premotor cortex (Fig. 1b). Listening to an untrained melody did not activate premotor cortex, suggesting that in the early stages of learning, perception–action mediation relies on fairly specific learned patterns. Dick et al. (2011) reported that in trained musicians, on the other hand, such activity does not differ between familiar and unfamiliar music (and similar training effects were observed in trained actors listening to dramatic speech). Bangert et al. (2006) measured BOLD signals during both listening to melodies and producing simple melodies with the right hand on a keyboard (without auditory feedback). In pianists, activation was observed during both perception and production of melodies in the premotor cortex, the pars opercularis (corresponding to BA 44), the planum temporale, and the supramarginal gyrus

(BA 40).[5] Activations within the premotor cortex (PMC) (and BA 44) during both perception and production of melodies were clearly left lateralized.

Interestingly, perception–action mediation appears to be modulated by emotional processes: In an fMRI experiment on music and emotion (in which pleasant and unpleasant music was presented to the participants; Koelsch et al. 2006) the contrast of listening to pleasant versus listening to unpleasant music showed an increase in BOLD signal in premotor areas, as well as in the Rolandic, or 'central', operculum during listening to pleasant music. During listening to unpleasant music, a decrease of BOLD signal in these areas was found. That is, premotor activity during listening to music was modulated by the emotional valence of the music, suggesting that perception–action mediation is modulated by emotional processes. It is likely that the Rolandic operculum contains, at least partly, the representation of the larynx, and therefore it seems that participants were quasi-automatically (that is, without being aware of this, and without intentional effort) singing subvocally along with the pleasant, but not with the unpleasant music. The activation of the Rolandic operculum during singing is different from the one reported by Callan et al. (2006), perhaps because the former study (Koelsch et al. 2006) used instrumental music, whereas the latter (Callan et al. 2006) used songs. The notion that mirror mechanisms can be modulated by emotional factors is consistent with findings showing that auditory mirror mechanisms elicited by emotional vocalizations can be modulated by the emotional valence of these vocalizations (Warren et al. 2006).[6]

With regard to temporal aspects of music, both cortical (supplementary motor area, SMA, and PMC) and subcortical structures (basal ganglia and cerebellum) are active during both perception and production of tactus, metre, and rhythm (e.g., Grahn and Brett 2007; Grahn and Rowe 2009; Grahn 2009). Moreover, functional connectivity between the basal ganglia, SMA, and PMC increases during the perception of tone sequences based on an isochronous pulse (Grahn and Rowe 2009). Finally, patients with Parkinson's disease show increased difficulties in discriminating changes in such sequences (compared to healthy controls; Grahn 2009), corroborating the notion that the basal ganglia (in addition to SMA, PMC, and cerebellum) play an important role for both the generation and the perception of rhythm and metre.

2 Neural Correlates of Music Production

The last section described action-related neural processes activated by music perception. This section will report studies investigating neural correlates of playing music, that is, of the execution of actions during music production. A few

[5] In addition, *performing* the melodies on a piano (without auditory feedback) elicited activity in auditory areas.

[6] That study used vocalizations such as "yuck" or "yippee" expressing triumph, amusement, fear, or disgust.

neuroimaging studies investigated the networks underlying music performance (Limb and Braun 2008; Parsons et al. 2005; Meister et al. 2004; Zatorre et al. 2007). Here, we will focus on recent studies using electroencephalography (EEG), in particular event-related brain potentials (ERPs). ERPs are brain-electric responses that can reflect with a millisecond resolution a number of cortical processes in the brain. Because of its high temporal resolution, EEG (and magnetoencephalography, MEG) is particularly suited to capture the temporal dynamics of neural processes underlying music performance.

When playing a musical instrument (or when singing), alone or together in a group, a player continuously establishes action goals, forms the corresponding motor programs to execute the right movements at the right time with the right strength, monitors ongoing movements by relating information such as proprioceptive feedback of the actual movements to the planned movements, and initiates corrective movements (when necessary). Corrective movements are also required when synchronizing movements with movements of other players. While playing, such corrections are memorized, and integrated with the execution of simultaneous movements. The perception of the action effects completes the action, and can modify selection, programming, execution, and control of new actions. All these processes overlap in time, making the investigation of these different processes challenging. One approach to investigate music production with ERPs is to examine neural correlates of error-related processes.

Even tens of thousands of hours of deliberate practice cannot prevent musicians from making errors ("error" means that the auditory outcome of an action does not match the intended tone). Errors can have various reasons (e.g., lack of attention, memory, misreading of a score). Recently, rare pitch errors occurring when a pianist hits the wrong key, have attracted some interest. Such errors can be good opportunities to gain insights into mechanisms operating during music performance, and error processing in general. Questions that arise in this context are at what point in time errors are actually detected by the sensorimotor system, whether they are detected already prior to execution, and—if so—at what point in time potential errors can still be corrected. Using music, an ERP-study by Maidhof et al. (2009) investigated whether errors are detected already *before* a movement is fully executed (for a similar study see Herrojo-Ruiz et al. 2009). That study (Maidhof et al. 2009) investigated expert pianists playing scales and scale patterns (bimanually) in a relatively fast tempo (Fig. 2). These stimuli were chosen to provoke pianists committing errors, with the aim to compare the brain-electric potentials related to incorrect with those related to correct keystrokes (in time intervals preceding, and following the onsets of keystrokes).

Results showed that, behaviourally, pianists pressed incorrect and correct keys with different velocities: Participants pressed incorrect keys with a lower velocity than (a) correct keypresses, and (b) the simultaneous correct keypresses (and the velocity of these simultaneous correct keypresses was not influenced by the lower velocity of the erroneous keypress of the other hand). Moreover, correct and incorrect keypresses were produced with different inter-onset intervals (IOIs): The IOI between an incorrect keypress and the preceding keypress was prolonged

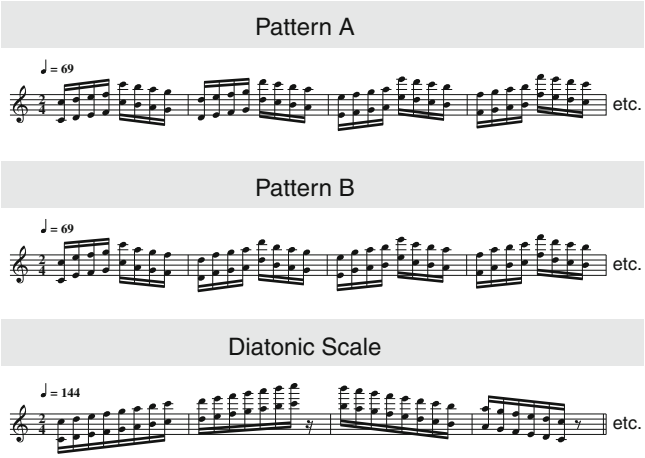


Fig. 2 Illustration of the patterns used in the study by Maidhof et al. (2009) (patterns are shown in *C* major; in the experiment, the stimuli had to be produced in different major keys). The instructed tempo for the scales was 144 bpm, and for the patterns 69 bpm

(compared to the IOI between successive correct keypresses), indicating that an upcoming error slowed down the keypresses (*pre-error slowing*). IOIs of simultaneous keystrokes performed by the other hand (without errors) were influenced by the error; that is, in contrast to the velocity, both hands showed a similar pre-error slowing during an error (even if the error occurred only in one hand), presumably due to the integration of several movements during executing an action.[7] In addition to pre-error slowing, Herrojo-Ruiz et al. (2009) also reported *post-error slowing* following the commission of an error. Those data (Herrojo-Ruiz et al. 2010) also indicate that pre-error and post-error slowing effects are limited to the trials directly preceding and following the errors, i.e., these temporal

[7] The integration of bimanual movements is also referred to as *bimanual coupling*: bimanual movements begin and end synchronously, even when they have different parameters (e.g. amplitudes), and even when movement times differ when the respective movements are performed in isolation by one hand (Marteniuk et al. 1984; Spijkers et al. 1997; Swinnen and Wenderoth 2004; Diedrichsen et al. 2010). In addition, musicians specifically train to play synchronously with both hands. Such integrative processes (including bimanual coupling) are not only due to low-level processes of motor execution, for "the symmetry constraint observed in bimanual coordination […] depends on perceptual variables and task demands […]. More generally, many demonstrations of constraints in bimanual coordination appear to reflect limitations in the simultaneous estimation of high-level, task-relevant states […], rather than hard-wired coordination constraints between the two hands. The human coordination system has evolved to achieve single goals flexibly using many effectors rather than to achieve multiple goals simultaneously" (Diedrichsen et al. 2010, p. 38).

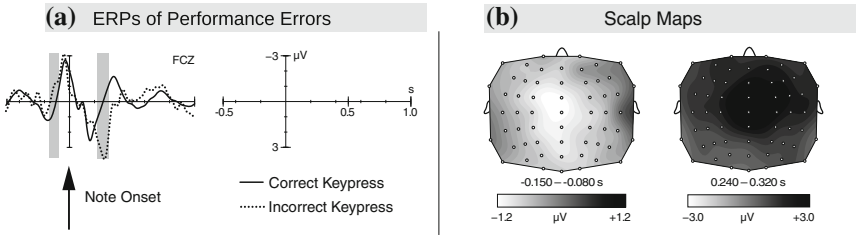


Fig. 3 Grand-average ERPs (recorded from ten pianists) elicited by correctly and incorrectly performed keypresses, time-locked to the onset of keypresses. **a** The *arrow* indicates the note onset and thus the onset of the auditory feedback. The grey areas highlight the pre-error negativity (occurring prior to the keypress), and the later positive effect (occurring after the incorrect keypress). **b** Scalp topographies for the difference potentials for correct keypresses subtracted from incorrect keypresses in the two time windows marked grey in (*A*). Modified with permission from Maidhof et al. (2009)

disruptions are not progressive phenomena, and slowing does not occur several events before or after an error.[8]

The ERPs of the study by Maidhof et al. (2009) showed that, compared to correct keypresses, incorrect keypresses elicited an increased negativity already *before* a wrong key was actually pressed down (Fig. 3). This ERP effect was maximal at central leads and peaked around 100 ms before a key was pressed down (left of Fig. 3B). This *pre-error negativity* was followed by a later positive deflection with an amplitude maximum at around 280 ms after the onset of an incorrect note. This potential had a fronto-central scalp topography and resembles the early *Error Positivity* (Pe) or the P3a (see also right of Fig. 3b). Virtually the same ERP pattern was reported in the study by Herrojo-Ruiz et al. (2009). That study (Herrojo-Ruiz et al. 2009) also employed a condition in which pianists played without auditory feedback. Under that condition, the error positivity was considerably reduced. In addition, the pre-error negativity elicited in the motor condition (without auditory feedback) was identical to the pre-error negativity elicited in the audio motor condition. This finding is consistent with findings showing that, after extensive learning of a sequence, auditory feedback is irrelevant for music performance with regard to error-monitoring (Finney 1997; Finney and Palmer 2003; Pfordresher 2003, 2005, 2006). In those studies, even the complete absence of auditory feedback had mostly no effects on the performance

[8] But see also Palmer et al. (2012), who report decreased key press velocities for correct tones immediately preceding incorrect tones.

of extensively trained piano pieces.[9] Note, however, that the vertical bar in Fig. 3 corresponds to the onset of the MIDI-signal. That is, the vertical bar corresponds to the point in time where the key was pressed down, and this point in time is preceded by touching the key and pressing down the key. Because touching the key and pressing down the key take several tens of milliseconds, tactile and proprioceptive feedback could already have contributed to the observed "pre-error negativity"![10]

In the study by Maidhof et al. (2009), left-hand errors and right-hand errors were also analyzed separately. That analysis showed that the ERPs elicited by the errors were not lateralized. Therefore, left- or right-hand errors were probably not simply due to some right- or left-hemispheric neural disturbance that caused the erroneous movement (such hemispheric disturbances would then occur bilaterally when averaged across left- and right-hand errors). That is, it does not appear that the early ERP difference occurring before a key was pressed down was the cause for the error. Instead, it appears that this early ERP difference reflects cognitive processes of error detection, error correction, and/or movement integration. With regard to movement integration, note that IOIs were prolonged before incorrect keypresses in *both* hands (in contrast to velocities, which differed between synchronous erroneous and correct key presses). Such integrative adjustment of bimanual movements perhaps contributed to this ERP effect.

The fact that the early ERP elicited by incorrect movements occurred *prior* to keypresses indicates that errors were detected *before* they were committed (and before auditory feedback was available). Such an error detection process is probably based on internal forward models: Probably during the formation of a motor program, a forward model is prepared which includes an efference copy, or 'corollary discharge'. The formation of a motor program takes into account the action goal, as well as the initial movement conditions, such as the respective locations and movements of body, arm, hand and target. The formation of a motor program appears to involve several areas, including the pre-supplementary motor area (pre-SMA) and SMA proper (see Fig. 4), the (pre-)supplementary eye field, premotor cortex, primary motor cortex (M1), basal ganglia, and parietal areas (e.g., Hoover and Strick 1999; Middleton and Strick 2000; Nachev et al. 2008; Desmurget and

[9] Specific alterations of auditory feedback, on the other hand, profoundly disrupt performance: For example, disruptive effects of pitch manipulations (false auditory feedback) occur during learning, or when the perceived feedback resembles an intended sequence (reviewed in Pfordresher, 2006). However, if auditory feedback is random, that is, when the feedback sequence is highly dissimilar to the intended sequence, the auditory feedback does not disrupt performance (presumably because players perceive the feedback as being unrelated to the planned actions).

[10] The exact point in time at which a key was touched could be determined with a motion-capture system.

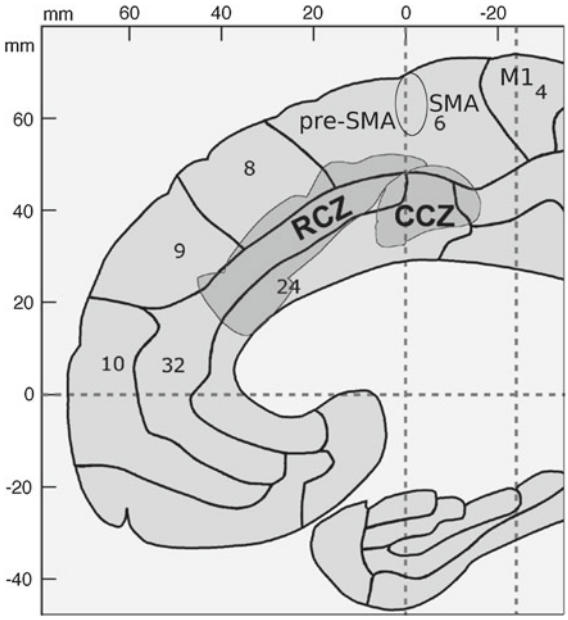


Fig. 4 Anatomical map of the medial frontal cortex. *SMA*: supplementary motor area, *RCZ*: rostral cingulate zone, *CCZ*: caudal cingulate zone. The vertical line through the anterior commissure (*left dashed line*) indicates the approximate border between SMA and pre-SMA. The oval indicates the presumed location of the supplementary eye field according to Amiez and Petrides (2009). The numbers indicate Brodmann areas. Left of the image corresponds to anterior. From Koelsch (2012)

Sirigu 2009; Schubotz 2007).[11] Studies investigating the activity of neurons in M1 of non-human primates showed that the latency between the first activity in M1 and movement onset is variable and can range up to several hundred milliseconds, with the typically assumed latency being around 100–150 ms (Evarts 1974; Porter and Lewis 1975; Thach 1978; Holdefer and Miller 2002; Hatsopoulos et al. 2007).

Then, at the same time as the motor command is sent from M1 to the periphery, an *efference copy* is created (supposedly in brain structures that are also involved in the generation of the movement), and sent to sensory structures. The efference copy is not used to generate the ongoing motor activity, but can be used to predict the outcome (i.e., the sensory consequences) of the motor command. Information of efference copies interacts at several levels of the central nervous system, and

[11] The basal ganglia are either part of the motor planner itself or are in a loop with planning structures. Note that cortico—basal ganglia—thalamo—cortical and cerebellar loops contribute to the programming, initiation, execution, and control of movements. For a role of the basal ganglia (as well as other motor structures) in the perception of tactus and metre see Grahn and Brett (2007). For the role of the basal ganglia in sensorimotor synchronization see Schwartze et al. (2011).

usually modulates sensory processing (see also Poulet and Hedwig 2007; Crapse and Sommer 2008).[12] In the auditory domain, recent studies with tasks in which participants initiate a sound simply by pressing a button show that auditory forward models can modulate auditory evoked potentials and oscillatory brain activity already at very early processing stages (around 30 ms after sound onset, Baess et al. 2008, 2009).

Particularly when actions are carried out quickly, as it is often the case during music performance, it is likely that the efference copy contains more predictive information than the preparation of the efference copy during program formation, e.g. with regard to those sensory consequences of movements that are due to external objects (such as objects that are touched, grazed, or moved during the movement). That is, in the course of the establishment of the efference copy, predictive information about sensory consequences is probably added to the forward model (these consequences are based on knowledge about the nature of external objects, such as their weight, temperature, surface texture, etc.).

During the execution of an action, information about the actual consequences of a movement, is *differentiated* from the information of the efference copy, and related to the predicted consequences of planned movements (Wolpert et al. 1995; 1998; Miall and Wolpert 1996; Desmurget and Grafton 2000; Wolpert and Ghahramani 2000). Information about the actual consequences originates from (a) somatosensory feedback (such as proprioceptive and tactile information), (b) visual feedback, and (c) efferent information (motor outflow).[13] Whenever there is a mismatch between actual and predicted consequences, an error signal is generated. It is likely that such signals are also generated when the sensory consequences of a movement deviate from the predicted consequences, even though the movement itself was correct (for example, when a key of the piano got stuck, or when an experimenter provides false auditory feedback, see also below). The occurrence of the pre-error negativity in the studies by Maidhof et al. (2009) and Herrojo-Ruiz et al. (2009) reflects that this detection mechanism operates *prior* to the full execution of the movement, and thus *before* the perception of the auditory feedback, that is, before the perception of the tone produced by the movement (action effect).

The error signal can, in turn, lead to the *corrective modulation* of motor commands (for details see Desmurget and Grafton 2000). Note, however, that the execution of movements is sometimes faster than the propagation of sensory information to the cortex, and that, thus, sensory feedback cannot always be used

[12] The information of the efference copy also tells the sensory areas about the upcoming sensory perceptions and allows them to prepare for the sensory consequences of the movement, for details see Crapse and Sommer (2008).

[13] Note that, despite some overlap, the networks underlying somatosensory feedback on the one hand, and visual feedback on the other, differ significantly from each other (see, e.g., Swinnen and Wenderoth 2004).

to correct movements (Desmurget and Grafton 2000).[14] Also note that, without any erroneous movement, movements can also be re-programmed during execution (Leuthold and Jentzsch 2002).[15]

In the studies by Maidhof et al. (2009) and Herrojo-Ruiz et al. (2009), the corrective modulation of the motor command might have resulted in the lower velocities of incorrect keypresses. Note that the IOIs were prolonged before incorrect keypresses in both hands (although only one hand performed an erroneous movement). This indicates that the movement of the hand playing the correct key was integrated with the erroneous movement of the other hand. Neural mechanisms of such integration processes during the performance of actions remain to be specified.[16]

The error positivity (Pe) following incorrect keypresses in the study by Maidhof et al. (2009) is probably related to the conscious recognition of a committed error (see also the "error-awareness hypothesis", e.g. Nieuwenhuis et al. 2001). Conscious recognition might also involve the adaptation of response strategy after an error has been perceived, involving remedial performance adjustments following errors ("behavior-adaption hypothesis", Hajcak et al. 2003). Such adaptive processes might also include making up for delays due to pre- and post-error slowing (Herrojo-Ruiz et al. 2010). Recognition of errors might also result in affective processes following the committed error or its consequences, including autonomic responses (such as changes in heart rate and sweat production). Especially in the case of highly-trained musicians, recognition of errors during performance may result in negative affective processes following the committed error or its consequences (like the sounding of a wrong note, which can also be perceived by listeners!), and the corresponding autonomic responses can be particularly obstructive in a concert situation. Recognition of errors during practice, on the other hand, has beneficial effects on learning (to avoid similar errors in the future when aiming to obtain a similar action goal).

Notably, in addition to the experimental design described so far, the mentioned study by Maidhof et al. (2009) provided the participants during playing every so often with manipulated (false) feedback when correct notes were played: Randomly between every 40–60th produced note, the auditory feedback of the digital piano was manipulated in a way that the pitch of one tone was lowered by one semitone (for a similar study see Katahira et al. 2008). This was done to investigate the time course of the neural mechanisms underlying the processing of (manipulated) feedback during music performance, and thus to study the processing of auditory

[14] Or, depending on the movement and the speed of a movement, sensory feedback loops might allow corrections only at the very end of the trajectory.

[15] In that study (Leuthold and Jentzsch 2002), re-programming was reflected electrically in a negative centro-parietal potential that was maximal at around 370 ms after the onset of a cue that required participants to re-program a movement that had already been commenced.

[16] It is likely that a network of numerous (sensori-)motor structures mediates bimanual movement integration. For a study suggesting involvement of the supplementary motor area (SMA) in this network see Steyvers et al. (2003).

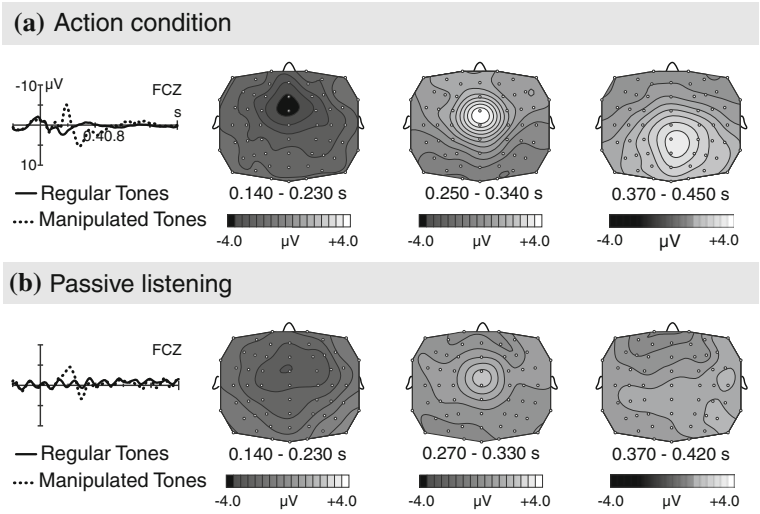


Fig. 5 **a** Grand-average ERPs (recorded from twelve pianists) elicited by correct keypresses with correct (*solid line*) and manipulated (*dotted line*) auditory feedback, time-locked to the onset of tones (i.e., when a key was pressed down). Feedback-manipulated tones elicit a feedback ERN (presumably overlapping with MMN/ERAN potentials), followed by a P3a, and a P3b (best to be seen in the isopotential maps). **b** Grand-average ERPs (recorded from the same twelve pianists) elicited while passively listening to the auditory stimuli of the action condition. Here, the deviants ("manipulated tones") also elicit a negativity peaking around 200 ms (possibly in part due to ERN potentials), and a P3a; no P3b was elicited in this condition (reflecting that the deviants were not task-relevant). Note that in both the action and the passive listening condition, ERN potentials presumably overlapped with MMN/ERAN potentials, making it challenging to disentangle action-related from perception-related brain potentials. Modified with permission from Maidhof et al. (2010)

effects of actions intended by a player or singer. Musicians expect to perceive the auditory feedback of their action, and the intention of musicians to produce specific auditory effects by executing certain actions is a fundamental aspect of music performance. Skilled piano players are trained to produce specific auditory effects with highly accurate movements (Ericsson and Lehmann 1996; Palmer 1997; Sloboda 2000). Accordingly, results of behavioral (Drost et al. 2005a, b), electrophysiological (Bangert and Altenmüller 2003), and neuroimaging studies (see first section of this chapter) consistently show pronounced coupling of auditory and motor systems in individuals with musical training.[17]

ERPs of correct (!) keypresses with and without feedback manipulation are shown in the top of Fig. 5 (reported in Maidhof et al. 2010). Feedback manipulations (i.e., the sound of a wrong note, although the correct key was pressed) elicited a negativity that was maximal at around 200 ms and had a fronto-central

[17] Notably, in the EEG study by Bangert and Altenmüller (2003), musically naive participants showed auditory-sensorimotor co-activity already within 20 min of piano learning.

scalp distribution. As will be discussed further below, this negativity was presumably mainly a *feedback error-related negativity* (feedback-ERN), with additional contributions of MMN/ERAN potentials.[18,19] The negativity was followed by a P3a, and a P3b.[20]

The feedback ERN is a type of *error-related negativity* (ERN or Ne, Botvinick et al. 2001; Yeung et al. 2004; Van Veen and Carter 2006; Falkenstein et al. 1990). Classically, the ERN (or *response-ERN*) is an ERP usually peaking shortly after participants commit an error in a variety of speeded response tasks (although the ERN often begins to emerge already before a button press, shortly after the onset of electromyographic potentials). The ERN typically peaks around 50–100 ms after incorrect responses, regardless of the modality in which the stimulus is presented, and regardless of the modality in which the response is made. The *feedback ERN* is elicited after negative performance feedback (compared to positive feedback), and after feedback stimuli indicating loss (or punishment) in time estimation tasks, guessing tasks and gambling tasks (Miltner et al. 1997; Hajcak et al. 2005, 2007). The feedback ERN is generally taken to reflect expectancy-related mechanisms, probably irrespective of whether the outcome of an event is worse or better than expected (Oliveira et al. 2007; Ferdinand et al. 2008).

Notably, feedback ERN-like components can also be observed in the absence of button-press responses (Donkers et al. 2005; Tzur and Berger 2007, 2009). The N200 is similar to the feedback ERN in latency and scalp distribution, and elicited when a mismatch between an expected and an actual sensory event is detected (Kopp and Wolff 2000; Ferdinand et al. 2008). Because the N2b is considered as a sub-component of the N200 which is elicited only when individuals consciously attend to a stimulus, the N200 in the mentioned studies is synonymous with the N2b (for other subcomponents of the N200 see Patel and Azzam 2005). There is even an ongoing debate as to whether the feedback ERN also reflects a subcomponent of the N200.

Note that musical feedback manipulations most presumably elicit, in addition to action-related processes, cognitive processes related to the perception of acoustic

[18] The "mismatch negativity" (MMN) is a brain electric response that is elicited when repetitive ("standard") auditory information is not repeated anymore, or repeated in a different way. For example, if a single tone is repeated several times, then the presentation of a tone with a different pitch, or loudness, or location, or timbre (or any other physical feature) elicits an MMN. Because the elicitation of an MMN relies on the representation (that is, a "memory trace") of the repetitive information in the auditory sensory memory, the MMN is thought to reflect in part auditory sensory memory operations. For a review see Näätänen et al. (2007).

[19] The "early right anterior negativity" (ERAN) is a brain electric response elicited by music-syntactic irregularities. Note that the elicitation of the ERAN relies on representations of music-syntactic regularities that are stored in a long-term memory format in the brain. This stays in contrast to the MMN: The regularities represented in a sensory memory trace are extracted on-line from the auditory input of the last moments. For reviews see Koelsch (2009) and Koelsch (2012).

[20] Behavioural results showed that feedback-manipulated tones did not cause longer IOIs with regard to succeeding tones.

deviants: Wrong notes (whether produced by the player him/herself, or whether due to false feedback) presumably elicit ERPs such as ERAN or MMN, which overlap with ERPs such as the ERN (or the N2b). A study by Maidhof et al. (2010) offered one approach to deal with this difficulty: In that study, ERPs elicited during music performance (*action condition*) were compared with ERPs elicited when musicians merely listened to such stimuli (*perception condition*). Such comparisons are also interesting because they can perhaps inform us about action-related processes evoked during music *perception*. Figure 5 (bottom panel) shows ERPs elicited in a condition in which pianists passively listened to the tone patterns with and without wrong (feedback-manipulated) tones produced in the action condition. As in the action condition, incorrect tones (compared to correct tones) elicited a negativity that was maximal around 200 ms, but with smaller amplitude than in the action condition. This negativity was followed by a small P3a (being maximal at frontal leads; no P3b was elicited in this condition).

That is, manipulated tones elicited during both the production and the perception of tones negative potentials with maximal amplitudes around 200 ms, with larger amplitude in the action compared to the perception condition. Similarly, the P3a elicited by (wrong) tones was more pronounced during the action condition (when participants were playing) compared to the perception condition (i.e., when participants only listened to the stimuli). The absence of a P3b during the perception condition reflects that the pitch manipulations were task-irrelevant for the participants. Because the N2b, or the ERN, is usually observed in combination with a P3b, it is likely that the observed negative potential is not simply an N2b or ERN.

That is, although the early negativity observed during the *action* condition is presumably in part due to a feedback ERN, it might well be that this ERN effect overlaps with MMN/ERAN potentials related to the processing of acoustic or harmonic-syntactic irregularity. On the other hand, the negativity elicited during the *perception* condition presumably reflects at least in part an MMN/ERAN, but it might well be that these potentials overlap in part with ERN/N2b potentials due to the simulation of action during the perception of music. This illustrates the difficulty to disentangle the different contributions of these components during music performance and music perception. However, there are several aspects that can be addressed to distinguish ERN- from MMN- or ERAN-potentials:

1. A comparison of feedback-manipulated tones with wrong tones produced by players themselves: In contrast to the ERPs elicited by feedback manipulations, ERPs of self-performed errors (Fig. 3) did not show a significant negative effect around 200 ms after the onset of erroneous keypresses (although a small negativity is visible in the ERPs of wrong tones in this time range, cf. Figure 3). Because the auditory deviance is comparable between self-generated errors and feedback manipulations, the ERPs of self-performed errors provide an estimate of possible MMN contributions. Thus, because no clear MMN is visible in the ERPs of self-performed errors, it is unlikely that the negativity elicited by feedback manipulations is simply an MMN or ERAN.

2. A localization of the sources of ERPs. Using current source density, the study by Maidhof et al. (2010) localized the sources of the negativities elicited in both action and perception condition in the rostral cingulate zone (RCZ) of the posterior medial frontal cortex (see also Fig. 4). These results are consistent with an explanation in terms of a feedback ERN: Studies on action monitoring and cognitive control indicate that the RCZ plays a key role in the processing of expectancy violations, performance monitoring, and the adjustment of actions for the improvement of task performance (Veen and Carter 2002; Ridderinkhof et al. 2004; Nieuwenhuis et al. 2004; Veen et al. 2004; Folstein and Van Petten 2008). Therefore, the ERN, the feedback ERN, and the N200/N2b presumably all receive (main) contributions from neural generators located in the RCZ (this also supports the notion that feedback ERN and N2b are subcomponents of the N200, with feedback ERN and N2b possibly being synonymous labels for effects with very similar functional significance).
3. A comparison between musical experts and musical beginners. A study by Katahira et al. (2008) reported an ERN to feedback-manipulated tones in musical experts (similar to the study by Maidhof et al. 2009), but no ERN was observed in participants who had only moderate musical training.[21] This is consistent with results by Maidhof et al. (2010), which showed that the ERN amplitude correlated negatively with the duration of musical training. An absence of ERN potentials (as in the study by Katahira et al. 2008) renders it unlikely that the effect observed in expert musicians is an MMN or ERAN, because frank violations elicit such potentials also in non-experts (although the amplitude of these potentials is also modulated by musical training, Tervaniemi and Huotilainen 2003; Tervaniemi 2009; Koelsch et al. 2002, 2007; Müller et al. 2010; Fujioka et al. 2005).
4. A comparison between diatonic (in-key) and non-diatonic (out-of-key) feedback-manipulations. If ERN potentials partly overlap with ERAN potentials, then non-diatonic feedback-manipulations should evoke larger negative effects than diatonic manipulations (because the ERAN amplitude is related to the degree of violation, Leino et al. 2007). The study by Katahira et al. (2008) reported that the ERN amplitude did not differ between diatonic and non-diatonic feedback-manipulations, suggesting that ERAN-related potentials did not, or only minimally, contribute to the ERN potentials.

It is also worth noting that the MMN is not influenced by the anticipation of deviant tones, nor by prior knowledge of deviant stimuli (e.g. Rinne et al. 2001; Waszak and Herwig 2007; Scherg et al. 1989). Moreover, the MMN amplitude does not differ between a condition in which participants trigger the presentation of

[21] Participants played unfamiliar melodies on a keyboard, and in five percent of the keypresses, the tone was shifted a semitone upwards.

tones, or listen to the same sequence of tones (Nittono 2006).[22] Therefore, different ERN amplitudes in the absence of P3b potentials (as observed in the study by Maidhof et al. 2009) indicate that possible MMN contributions could have been only minor.

The discussed methods provide approaches to illuminate to which extent evoked negativities following the perception of feedback manipulations reflect ERN, N2b, or MMN/ERAN potentials. As mentioned above, the manipulated tones elicited in both the action and the perception condition an early negativity that strongly resembled the ERN (in terms of latency, distribution, and neural generators). In both the study by Katahira et al. (2008) and by Maidhof et al. (2010), this feedback-ERN effect was more pronounced during the performance of music compared to the mere perception of music. Thus, it seems likely that similar expectancy-related mechanisms operate during both performance and perception of music.[23] Importantly, the feedback ERN in the mentioned music studies is influenced by the expectancies generated by the intention and action of the pianists to produce a certain auditory effect. In contrast to these action-related expectancies, pianists could also build expectancies during the perception of the sequences based on the preceding musical context and its underlying regularities. Consequently, the manipulated tones during piano performance were more unexpected than the manipulated tones during the perception of the sequences, resulting in the enlarged feedback ERN in the action compared to the perception condition.

If the feedback ERN reflects the processing of violations of action-related predictions, how are these predictions established during the production and perception of musical sequences? It appears that, during the production of action sequences, pianists anticipate the tone mapped to the particular keypress they are about to perform. After having learned these associations (due to extensive training), the formation of an action plan leads to the establishment of predictions of the sensory feedback using the internal forward model described above. This implies that predictions are formed before a motor command is sent. According to the common coding theory (Prinz 1990), it also seems likely that, when making music, actions are selected, and controlled, using an inverse model of the intended effect, leading to an expectation for a certain effect (the *ideomotor principle*, see e.g. Hommel et al. 2001). As mentioned above, the common coding theory assumes that coding of perception overlaps with the coding of action in the sense that they share a common representational format. Therefore, the anticipated effects of an action should influence its planning, control and execution (the *action-effect principle*). The notion that the prediction of action effects is related to

[22] In that study (Nittono, 2006), participants triggered the presentation of a tone (which was either a standard tone, or one of two pitch-deviants) by pressing buttons. That is, participants had control over the timing of the stimuli, but not over the pitch of the stimuli.

[23] Note that it is conceivable that a feedback ERN can also be elicited during perception (without action), because feedback ERN-like waveforms are also observed when no actions, or responses, are required on the part of the participants (Donkers et al. 2005), and when rules (i.e., expectations) are violated in tasks without overt responses (Tzur and Berger 2007, 2009).

the training of the participants is supported by the correlation between ERN amplitude and amount of training: In the study by Katahira et al. (2008), no ERN was elicited in non-expert players, and in the study by Maidhof et al. (2010) pianists with longer training showed larger ERN amplitudes. While listening to the sequences without playing, predictive mechanisms probably extrapolate from the regularities of the preceding auditory input, and thus generate a prediction towards a specific sound to follow. This expectancy (or prediction) seems to be a fundamental aspect of perception, which is most likely not under the strategic control of individuals (Schubotz 2007).

Even more importantly, the combined data show that the processing of expectancy violations is modulated by the action of an individual. During music performance, players (or singers) expect, based on their intention and their act of performing, to perceive a specific auditory effect. In addition, the preceding musical context induces expectancies for specific tones. Hence, when an unexpected tone is encountered following an action, the detection of the violation of such expectancies elicits a brain response similar to the feedback ERN/N200. A similar effect, although with smaller amplitude, is elicited when pianists merely perceive an unexpected tone (without performing). It is tempting to speculate that these processes are in part also due to action-related mechanisms, such as an effect of simulated action during the perception of music.

The reported studies investigating ERP correlates during music production provided first insights into neural mechanisms operating during music performance. However, several issues remain to be investigated. For example, detailed information about how the movements were executed in terms of their kinetic and kinematic features were lacking, and neural correlates could only be investigated with respect to on- and offsets of the recorded MIDI-signals (i.e., information about the time point when a key was pressed down, and about the key press velocity). Therefore, much of the information about a musical performance could not be quantified. By contrast, using motion capture techniques allows to investigate movements underlying music performance more directly, and enable the analysis of movements of body parts with a high spatial and temporal accuracy. Such studies investigated a huge variety of research questions, e.g. the role of tactile feedback in timing accuracy during piano and clarinet performance (Goebl and Palmer 2008; Palmer et al. 2009), disruptive effects of delayed auditory feedback during rhythm production and the role of the ongoing movement trajectory (Pfordresher and Dalla Bella 2011), the effect of tempo on finger kinematics in pianists (Dalla Bella and Palmer 2011), identification of pianists based on their profiles of finger velocity and acceleration (Dalla Bella and Palmer 2011), the role of anticipatory auditory imagery during music-like sequential tasks (Keller et al. 2010), the relationships between the kinematics of a singer's body movement and their vocal performances (Luck and Toiviainen 2008), the role of different features of conductors' gestures used by ensemble musicians to synchronize (Luck and Toiviainen 2006), to name just a few. It is also worth noting that, in addition, motion capture techniques were used to investigate movements not directly involved in sound production, but

involved in emotional expressions (Livingstone et al. 2009) or as cues for other performers in ensemble performance (e.g., Keller and Appel 2010).

Importantly, obtaining additional detailed information about the movements can lead to a more behaviour-driven analysis of brain activity, and thus to a better understanding of neural processes involved in music performance. In particular, setups combining the recording of EEG, MIDI, and motion capture data could enable researchers to investigate brain activity of natural musical behaviours without many of the limitations mentioned above. A vision for the future is to adapt such setups for the simultaneous acquisition of data from several players to investigate interactions between performance and social factors.

References

Amiez, C., & Petrides, M. (2009). Anatomical organization of the eye fields in the human and non-human primate frontal cortex. *Progress in Neurobiology, 89*(2), 220–230.

Baess, P., Jacobsen, T., & Schröger, E. (2008). Suppression of the auditory n1 event-related potential component with unpredictable self-initiated tones: evidence for internal forward models with dynamic stimulation. *International Journal of Psychophysiology, 70*(2), 137–143.

Baess, P., Widmann, A., Roye, A., Schröger, E., & Jacobsen, T. (2009). Attenuated human auditory middle latency response and evoked 40 hz response to self-initiated sounds. *European Journal of Neuroscience, 29*(7), 1514–1521.

Bangert, M., & Altenmüller, E. O. (2003). Mapping perception to action in piano practice: a longitudinal dc-eeg study. *BMC Neuroscience, 4*(1), 26–39.

Bangert, M., Peschel, T., Schlaug, G., Rotte, M., Drescher, D., & Hinrichs, H. (2006). Shared networks for auditory and motor processing in professional pianists: Evidence from fMRI conjunction. *Neuroimage, 30*(3), 917–926.

Botvinick, M., Braver, T., Barch, D., Carter, C., & Cohen, J. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624–652.

Callan, D., Tsytsarev, V., Hanakawa, T., Callan, A., Katsuhara, M., & Fukuyama, H. (2006). Song and speech: brain regions involved with perception and covert production. *Neuroimage, 31*(3), 1327–1342.

Crapse, T., & Sommer, M. (2008). Corollary discharge circuits in the primate brain. *Current Opinion in Neurobiology, 18*(6), 552–557.

Dalla Bella, S., & Palmer, C. (2011). Rate effects on timing, key velocity, and finger kinematics in piano performance. *PLoS ONE, 6*(6), e20518.

Desmurget, M., & Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. *Trends in Cognitive Sciences, 4*(11), 423–431.

Desmurget, M., & Sirigu, A. (2009). A parietal-premotor network for movement intention and motor awareness. *Trends in cognitive sciences, 13*(10), 411–419.

Dick, F., Lee, H., Nusbaum, H., & Price, C. (2011). Auditory-motor expertise alters "speech selectivity" in professional musicians and actors. *Cerebral Cortex, 21*(4), 938–948.

Diedrichsen, J., Shadmehr, R., & Ivry, R. (2010). The coordination of movement: optimal feedback control and beyond. *Trends in cognitive sciences, 14*(1), 31–39.

Donkers, F., Nieuwenhuis, S., & van Boxtel, G. (2005). Mediofrontal negativities in the absence of responding. *Cognitive Brain Research, 25*(3), 777–787.

Drost, U., Rieger, M., Brass, M., Gunter, T. C., & Prinz, W. (2005a). Action-effect coupling in pianists. *Psychological Research, 69*(4), 233–241.

Drost, U., Rieger, M., Brass, M., Gunter, T. C., & Prinz, W. (2005b). When hearing turns into playing: Movement induction by auditory stimuli in pianists. *The Quarterly Journal of Experimental Psychology Section A, 58*(8), 1376–1389.

Ericsson, K., & Lehmann, A. (1996). Expert and exceptional performance: evidence of maximal adaption to task constraints. *Annual Review of Psychology, 47*, 273–305.

Evarts, E. (1974). Precentral and postcentral cortical activity in association with visually triggered movement. *Journal of Neurophysiology, 37*(2), 373–381.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1990). Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In C. Brunia, A. Gaillard, & A. Kok (Eds.), *Psychophysiological brain research* (pp. 192–195). Tilburg: Tilburg University Press.

Ferdinand, N., Mecklinger, A., & Kray, J. (2008). Error and deviance processing in implicit and explicit sequence learning. *Journal of Cognitive Neuroscience, 20*(4), 629–642.

Finney, S. (1997). Auditory feedback and musical keyboard performance. *Music Perception, 15*(2), 153–174.

Finney, S., & Palmer, C. (2003). Auditory feedback and memory for music performance: Sound evidence for an encoding effect. *Memory and Cognition, 31*(1), 51–64.

Fitch, W. (2006). The biology and evolution of music: A comparative perspective. *Cognition, 100*(1), 173–215.

Folstein, J., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology, 45*(1), 152–170.

Fujioka, T., Trainor, L., Ross, B., Kakigi, R., & Pantev, C. (2005). Automatic encoding of polyphonic melodies in musicians and nonmusicians. *Journal of Cognitive Neuroscience, 17*(10), 1578–1592.

Goebl, W., & Palmer, C. (2008). Tactile feedback and timing accuracy in piano performance. *Experimental Brain Research, 186*(3), 471–479.

Grahn, J. (2009). The role of the basal ganglia in beat perception. *Annals of the New York Academy of Sciences*, 1169(1), 35–45.

Grahn, J., & Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *Journal of Cognitive Neuroscience*, 19(5), 893–906.

Grahn, J., & Rowe, J. (2009). Feeling the beat: premotor and striatal interactions in musicians and nonmusicians during beat perception. *The Journal of Neuroscience, 29*(23), 7540–7548.

Hajcak, G., Holroyd, C., Moser, J., & Simons, R. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology, 42*(2), 161–170.

Hajcak, G., McDonald, N., & Simons, R. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behaviour. *Psychophysiology, 40*(6), 895–903.

Hajcak, G., Moser, J., Holroyd, C., & Simons, R. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology, 44*(6), 905–912.

Haslinger, B., Erhard, P., Altenmüller, E., Schroeder, U., Boecker, H., & Ceballos-Baumann, A. (2005). Transmodal sensorimotor networks during action observation in professional pianists. *Journal of cognitive neuroscience, 17*(2), 282–293.

Hatsopoulos, N., Xu, Q., & Amit, Y. (2007). Encoding of movement fragments in the motor cortex. *The Journal of neuroscience, 27*(19), 5105–5114.

Haueisen, J., & Knösche, T. (2001). Involuntary motor activity in pianists evoked by music perception. *Journal of Cognitive Neuroscience, 13*(6), 786–792.

Herrojo-Ruiz, M., Jabusch, H., & Altenmüller, E. (2009). Detecting wrong notes in advance: neuronal correlates of error monitoring in pianists. *Cerebral Cortex, 19*(11), 2625–2639.

Herrojo-Ruiz, M., Strübing, F., Jabusch, H. C., & Altenmüller, E. (2010). EEG oscillatory patterns are associated with error prediction during music performance and are altered in musician's dystonia. *NeuroImage, 55*, 1791–1803.

Holdefer, R., & Miller, L. (2002). Primary motor cortical neurons encode functional muscle synergies. *Experimental Brain Research, 146*(2), 233–243.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*(05), 849–878.

Hoover, J., & Strick, P. (1999). The organization of cerebellar and basal ganglia outputs to primary motor cortex as revealed by retrograde transneuronal transport of herpes simplex virus type 1. *The Journal of neuroscience, 19*(4), 1446–1463.

James, W. (1890). The principles of psychology. NewYork: Henry Holt and Company.

Katahira, K., Abla, D., Masuda, S., & Okanoya, K. (2008). Feedback-based error monitoring processes during musical performance: An ERP study. *Neuroscience Research, 61*(1), 120–128.

Keller, P., & Appel, M. (2010). Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles. *Music Perception, 28*(1), 27–46.

Keller, P., Dalla Bella, S., & Koch, I. (2010). Auditory imagery shapes movement timing and kinematics: Evidence from a musical task. *Journal of Experimental Psychology: Human Perception and Performance, 36*(2), 508.

Koelsch, S. (2009). Music-syntactic processing and auditory memory: Similarities and differences between ERAN and MMN. *Psychophysiology, 46*(1), 179–190.

Koelsch, S. (2012). Brain and music. Wiley.

Koelsch, S., Fritz, T., Cramon, D. Y., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: An fMRI study. *Human Brain Mapping, 27*(3), 239–250.

Koelsch, S., Jentschke, S., Sammler, D., & Mietchen, D. (2007). Untangling syntactic and sensory processing: An ERP study of music perception. *Psychophysiology, 44*(3), 476–490.

Koelsch, S., Schmidt, B. H., & Kansok, J. (2002). Influences of musical expertise on the ERAN: An ERP-study. *Psychophysiology, 39*, 657–663.

Kohler, E., Keysers, C., Umilta, M., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science, 297*(5582), 846–848.

Kopp, B., & Wolff, M. (2000). Brain mechanisms of selective learning: Event-related potentials provide evidence for error-driven learning in humans. *Biological Psychology, 51*(2–3), 223–246.

Lahav, A., Saltzman, E., & Schlaug, G. (2007). Action representation of sound: audio motor recognition network while listening to newly acquired actions. *Journal of Neuroscience, 27*(2), 308–314.

Leino, S., Brattico, E., Tervaniemi, M., & Vuust, P. (2007). Representation of harmony rules in the human brain: Further evidence from event-related potentials. *Brain Research, 1142*, 169–177.

Leuthold, H., & Jentzsch, I. (2002). Spatiotemporal source localisation reveals involvement of medial premotor areas in movement reprogramming. *Experimental Brain Research, 144*(2), 178–188.

Liberman, A., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36.

Limb, C., & Braun, A. (2008). Neural substrates of spontaneous musical performance: An fMRI study of jazz improvisation. *PLoS ONE, 3*(2), e1679.

Livingstone, S., Thompson, W., & Russo, F. (2009). Facial expressions and emotional singing: A study of perception and production with motion capture and electromyography. *Music Perception, 26*(5), 475–488.

Lotze, H. (1852). *Medicinische psychologie oder physiologie der seele*. Leipzig: Weidmann.

Luck, G., & Toiviainen, P. (2006). Ensemble musiciansâ^{TM} synchronization with conductorsâ^{TM} gestures: An automated feature-extraction analysis. *Music Perception, 24*(2), 189–200.

Luck, G., & Toiviainen, P. (2008). Exploring relationships between the kinematics of a singerâ^{TM}s body movement and the quality of their voice. *Journal of interdisciplinary music studies, 2*(1–2), 173–186.

Maidhof, C., Rieger, M., Prinz, W., & Koelsch, S. (2009). Nobody is perfect: ERP effects prior to performance errors in musicians indicate fast monitoring processes. *PLoS ONE, 4*(4), e5032.

Maidhof, C., Vavatzanidis, N., Prinz, W., Rieger, M., & Koelsch, S. (2010). Processing expectancy violations during music performance and perception: an ERP study. *Journal of Cognitive Neuroscience, 22*(10), 2401–2413.

Marteniuk, R., MacKenzie, C., & Baba, D. (1984). Bimanual movement control: Information processing and interaction effects. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 36*(2), 335–365.

Meister, I., Krings, T., Foltys, H., Boroojerdi, B., Müller, M., & Töpper, R. (2004). Playing piano in the mind an fmri study on music imagery and performance in pianists. *Cognitive Brain Research, 19*(3), 219–228.

Miall, R., & Wolpert, D. (1996). Forward models for physiological motor control. *Neural networks, 9*(8), 1265–1279.

Middleton, F., & Strick, P. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Research Reviews, 31*(2–3), 236–250.

Miltner, W., Braun, C., & Coles, M. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience, 9*(6), 788–798.

Müller, M., Höfel, L., Brattico, E., & Jacobsen, T. (2010). Aesthetic judgments of music in experts and laypersons—an ERP study. *International Journal of Psychophysiology, 76*(1), 40–51.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology, 118*(12), 2544–2590.

Nachev, P., Kennard, C., & Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. *Nature Reviews Neuroscience, 9*(11), 856–869.

Nieuwenhuis, S., Holroyd, C., Mol, N., & Coles, M. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience and Biobehavioral Reviews, 28*(4), 441–448.

Nieuwenhuis, S., Ridderinkhof, K., Blom, J., Band, G., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology, 38*(5), 752–760.

Nittono, H. (2006). Voluntary stimulus production enhances deviance processing in the brain. *International Journal of Psychophysiology, 59*(1), 15–21.

Oliveira, F., McDonald, J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. *Journal of cognitive neuroscience, 19*(12), 1994–2004.

Palmer, C. (1997). Music performance. *Annual Review of Psychology, 48*, 115–138.

Palmer, C., Koopmans, E., Loehr, J., & Carter, C. (2009). Movement-related feedback and temporal accuracy in clarinet performance. *Music Perception, 26*(5), 439–449.

Palmer, C., Mathias, B., & Anderson, M. (2012). Sensorimotor mechanisms in music performance: actions that go partially wrong. *Annals of the New York Academy of Sciences, 1252*(1), 185–191.

Parsons, L., Sergent, J., Hodges, D., & Fox, P. (2005). The brain basis of piano performance. *Neuropsychologia, 43*(2), 199–215.

Patel, S., & Azzam, P. (2005). Characterization of N200 and P300: Selected studies of the event-related potential. *International Journal of Medical Sciences, 2*(4), 147–154.

Pfordresher, P. (2003). Auditory feedback in music performance: Evidence for a dissociation of sequencing and timing. *Journal of Experimental Psychology, 29*(4), 949–964.

Pfordresher, P. (2005). Auditory feedback in music performance: The role of melodic structure and musical skill. *Journal of Experimental Psychology, 31*(6), 1331–1345.

Pfordresher, P. (2006). Coordination of perception and action in music performance. *Advances in Cognitive Psychology, 2*(2), 183–198.

Pfordresher, P., & Dalla Bella, S. (2011). Delayed auditory feedback and movement. *Journal of Experimental Psychology: Human Perception and Performance, 37*(2), 566.

Porter, R., & Lewis, M. (1975). Relationship of neuronal discharges in the precentral gyrus of monkeys to the performance of arm movements. *Brain Research, 98*(1), 21–36.

Poulet, J., & Hedwig, B. (2007). New insights into corollary discharges mediated by identified neural pathways. *Trends in Neurosciences, 30*(1), 14–21.

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.), *Relationships between perception and action* (pp. 167–201). NewYork: Springer.

Prinz, W. (2005). An ideomotor approach to imitation. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: Mechanisms of imitation and imitation in animals* (pp. 141–156). Cambridge: MIT Press.

Ridderinkhof, K., Ullsperger, M., Crone, E., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science, 306*(5695), 443.

Rinne, T., Antila, S., & Winkler, I. (2001). Mismatch negativity is unaffected by top-down predictive information. *NeuroReport, 12*(10), 2209–2213.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Reviews Neuroscience, 11*(4), 264–274.

Scherg, M., Vajsar, J., & Picton, T. W. (1989). A source analysis of the late human auditory evoked potentials. *Journal of Cognitive Neuroscience, 1*, 336–355.

Schubotz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. *Trends in cognitive sciences, 11*(5), 211–218.

Schwartze, M., Keller, P., Patel, A., & Kotz, S. (2011). The impact of basal ganglia lesions on sensorimotor synchronization; spontaneous motor tempo; and the detection of tempo changes. *Behavioural Brain Research, 216*(2), 685–691.

Sloboda, J. (2000). Individual differences in music performance. *Trends in cognitive sciences, 4*(10), 397–403.

Spijkers, W., Heuer, H., Kleinsorge, T., & van der Loo, H. (1997). Preparation of bimanual movements with same and different amplitudes: Specification interference as revealed by reaction time. *Acta Psychologica, 96*(3), 207–227.

Steyvers, M., Etoh, S., Sauner, D., Levin, O., Siebner, H., & Swinnen, S. (2003). High-frequency transcranial magnetic stimulation of the supplementary motor area reduces bimanual coupling during anti-phase but not in-phase movements. *Experimental Brain Research, 151*(3), 309–317.

Swinnen, S., & Wenderoth, N. (2004). Two hands, one brain: cognitive neuroscience of bimanual skill. *Trends in Cognitive Sciences, 8*(1), 18–25.

Tervaniemi, M. (2009). Musicians—same or different? *Annals of the New York Academy of Sciences*, 1169(The Neurosciences and Music III Disorders and Plasticity), 151–156.

Tervaniemi, M., & Huotilainen, M. (2003). The promises of change-related brain potentials in cognitive neuroscience of music. *Annals of the New York Academy of Sciences*, 999(THE NEUROSCIENCES AND MUSIC), 29–39.

Thach, W. (1978). Correlation of neural discharge with pattern and force of muscular activity, joint position, and direction of intended next movement in motor cortex and cerebellum. *Journal of Neurophysiology, 41*(3), 654–676.

Tzur, G., & Berger, A. (2007). When things look wrong: An ERP study of perceived erroneous information. *Neuropsychologia, 45*, 3122–3126.

Tzur, G., & Berger, A. (2009). Fast and slow brain rhythms in rule/expectation violation tasks: Focusing on evaluation processes by excluding motor action. *Behavioural Brain Research, 198*(2), 420–428.

Van Veen, V., & Carter, C. (2006). Error detection, correction, and prevention in the brain: a brief review of data and theories. *Clinical EEG and neuroscience, 37*(4), 330–335.

van Veen, V., & Carter, C. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience, 14*(4), 593–602.

van Veen, V., Holroyd, C., Cohen, J., Stenger, V., & Carter, C. (2004). Errors without conflict: implications for performance monitoring theories of anterior cingulate cortex. *Brain and Cognition, 56*(2), 267–276.

Warren, J., Sauter, D., Eisner, F., Wiland, J., Dresner, M., & Wise, R. (2006). Positive emotions preferentially engage an auditory-motor "mirror" system. *Journal of Neuroscience, 26*(50), 13067–13075.

Waszak, F., & Herwig, A. (2007). Effect anticipation modulates deviance processing in the brain. *Brain Research, 1183*, 74–82.

Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*, 1212–1217.

Wolpert, D., Ghahramani, Z., & Jordan, M. (1995). An internal model for sensorimotor integration. *Science, 269*(5232), 1880–1882.

Wolpert, D., Miall, R., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in Cognitive Sciences, 2*(9), 338–347.

Yeung, N., Botvinick, M., & Cohen, J. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111*(4), 931–959.

Zatorre, R., Chen, J., & Penhune, V. (2007). When the brain plays music: auditory–motor interactions in music perception and production. *Nature Reviews Neuroscience, 8*(7), 547–558.

Music for the Brain Across Life

Teppo Särkämö, Mari Tervaniemi and Minna Huotilainen

1 Foreword

As the French poet Victor Hugo (1802–1885) put it, "music expresses that which cannot be said and on which it is impossible to be silent". Just like spoken language, music has been an essential part of every known human culture and therefore has roots that reach deep into our very selves and into our brains. Thus far, the oldest concrete evidence regarding the early existence of music were obtained a few years ago from southern Germany, where archaeological excavations revealed a 40,000-year-old flute made of bone (Conard et al. 2009). Some scholars believe that a singing-based form of communication, a protolanguage, could be even older, possibly dating back over 200,000 years, and could have formed a basis of the development of modern spoken language (Mithen 2005).

More recently, various cultural trends and technological innovations, such as the karaoke and the choir singing boom, MP3 players, and digital streaming services and players (e.g., Spotify, iTunes), have made music more available and easily accessible than ever before. In its many forms, music has become a popular leisure activity and hobby through which many of us mediate our emotional and

T. Särkämö (✉) · M. Tervaniemi · M. Huotilainen
Cognitive Brain Research Unit, Institute of Behavioural Sciences, University of Helsinki, Siltavuorenpenger 1 B, 9 FI-00014 Helsinki, Finland
e-mail: teppo.sarkamo@helsinki.fi

T. Särkämö · M. Tervaniemi · M. Huotilainen
Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Jyväskylä, Finland

M. Tervaniemi
Department of Psychology, University of Jyväskylä, Jyväskylä, Finland

M. Huotilainen
Finnish Institute of Occupational Health, Helsinki, Finland

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_7,

arousal state, experience creativity and aesthetic pleasure, and interact with others. Thanks to modern brain imaging methods, such as electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET), as well as behavioural and clinical studies, we are now starting to understand better how music affects us and how it can be used to promote well-being and facilitate recovery and rehabilitation. In this article, we aim to provide a brief review of the neural basis of music in both the healthy and the damaged brain, the development of musical skills and the meaning of music in different ages, and the effectiveness of music-based interventions in various somatic, psychiatric, and neurological illnesses.[1]

2 Music in the Healthy Brain

Neuroscience of music is a relatively new, fast-developing field of science, which has during the past 20 years provided a lot of novel information on how music is processed in the brain, how musical activities can shape the brain, and what neural mechanisms underlie the therapeutic effect of music. To date, converging evidence suggests that music activates an extremely complex and wide-spread, bilateral network of cortical and subcortical areas that controls many auditory, cognitive, motor, and emotional functions.

The process of music perception begins in the inner ears where acoustic information is converted to an electric impulse or signal. The signal then travels along the auditory nerve to the brain stem (especially to the inferior colliculus) where certain basic features of the sound, such as periodicity and intensity, are first processed (Pickles 2008). Interestingly, the earliest signs of musical training can be seen as early as 10 ms after sound onset in the auditory brain stem, which in musicians can represent the frequency of the sound with more fidelity than in non-musicians (Kraus and Chandrasekaran 2010). From the brain stem, the auditory information is conveyed to the thalamus and from there primarily to the auditory cortex, but also directly to limbic areas, such as the amygdala and the medial orbitofrontal cortex (LeDoux 2000). All across this pathway, an enormous amount of ascending contacts are active in the process of refining and processing the auditory input. The primary auditory cortex and its neighbouring superior temporal areas analyse the basic acoustic cues of the sound, including frequency, pitch, sound level, temporal variation, motion, and spatial location (Hall et al. 2003). The left auditory cortex has a better temporal resolution and the right auditory cortex a better spectral resolution, which is thought to form one crucial premise for the lateralization of speech to the left hemisphere and music to the right hemisphere (Zatorre et al. 2002).

[1] An extended and updated version of this review entitled "Music perception and cognition: development, neural basis, and rehabilitative use of music" will be published in Wiley Interdisciplinary Reviews: Cognitive Science

Music is, however, much more than just the sum of its basic acoustic features. Upon its initial encoding and perception, music triggers a sequence of cognitive, motor, and emotional processes in the brain that are governed by numerous cortical and subcortical areas. Next, we outline five such processes.

1. The perception of higher-order musical features, such as chords, harmonies, intervals, and rhythms, calls for a rule-based *syntactic analysis* of complex patterns of spectral and temporal fluctuations within the sound stream, enabling the perception of some of the most essential features of music syntax. According to neuroimaging studies, this takes place in a network comprising the inferior and medial prefrontal cortex, the premotor cortex, the anterior and posterior parts of the superior temporal gyrus, and the inferior parietal lobe (Janata et al. 2002a; Koelsch and Siebel 2005; Patel 2003).
2. Continually keeping track of the music, which always unfolds over time, requires the engagement of the *attention and working memory system*, which is spread over many prefrontal areas (especially the dorsolateral prefrontal cortex), the cingulate cortex, and inferior parietal areas (Janata et al. 2002b; Zatorre et al. 1994).
3. Hearing music that is familiar to the listener from past experience triggers processing especially in the hippocampus as well as in medial temporal and parietal areas, which are involved in *episodic memory* (Janata 2009; Platel et al. 2003).
4. Hearing music that touches us emotionally engages a network of many deep limbic and paralimbic areas, including various midbrain areas, striatal areas (especially nucleus accumbens), the amygdala, the hippocampus, the cingulate cortex, and the orbitofrontal cortex (Blood and Zatorre 2001; Koelsch 2010; Menon and Levitin 2005). This dopaminergic network is known as the *mesolimbic or reward system* of the brain and it has been implicated in the experiencing of emotions, pleasure, and reward and in regulating the autonomic nervous system and the endocrine (or hormone) system. Recently, the direct involvement of striatal dopamine in the emotional reaction to music was demonstrated in a combined psychophysiological, PET and fMRI study (Salimpoor et al. 2011).
5. Producing music by singing or playing an instrument, moving to the beat of music, or even just perceiving the rhythm of music involves the *motor network* of the brain, including areas in the cerebellum, the basal ganglia, and the motor and somatosensory cortices (Grahn and Rowe 2009; Zatorre et al. 2007).

3 Musical Disorders in the Brain

Our ability to perceive, process, and appreciate music may become impaired in many neurological illnesses (Goll et al. 2010). The most well-known disorder is *amusia*, which can be either innate (*congenital amusia*) or result from a brain

lesion (*acquired amusia*). The term amusia refers to an inability to perceive and/or produce music, which is not caused by a disorder in another domain, such as hearing, motor, or cognitive functions (Peretz 2006; Stewart et al. 2006). Amusia can be observed in the majority of musical features (perceiving pitch, timbre, or rhythm or recognizing musical emotions or musical pieces) or be specific to one or some of them. The most commonly reported deficit is that of poor pitch discrimination: amusic individuals are typically not able to perceive pitch changes smaller than a semitone (Foxton et al. 2004). As a result, they often have great difficulties in perceiving sequential notes (or tones) and, therefore, in recognizing melodies– for some rare individuals, music may sound more like noise.

It has been estimated that the prevalence of congenital amusia is approximately 2–4 % in the general population (Henry and McAuley 2010). Genetic studies of congenital amusia suggest that the disorder is heritable: in amusic families, 39 % of first-degree relatives have the same deficit, whereas only 3 % have it in the control families (Peretz et al. 2007). Furthermore, dizygotic (identical) twins have more uniform performance in a musical pitch perception test than monozygotic (fraternal) twins (Drayna et al. 2001). Compared to congenital amusia, acquired amusia seems to be a lot more common deficit, at least after a cerebrovascular accident such as stroke. In studies of stroke patients, the reported incidence of amusia has varied between 60 and 69 % in the acute stage (about one week post-onset) and between 35–42 % in the subacute/chronic stage (>3 months post-lesion) (Särkämö et al. 2009; Schuppert et al. 2000; Ayotte et al. 2000). Based on structural and functional MRI studies, the crucial neuroanatomical correlate of congenical amusia appears to be the auditory cortex and the inferior frontal gyrus in the right hemisphere (Hyde et al. 2007, 2011) as well as the subcortical neural matter tracts (arquate fasciculus) connecting these areas (Loui et al. 2009). Correspondingly, acquired amusia is most typically caused by damage to the auditory cortex and its surrounding cortical and subcortical areas (anterior and posterior superior temporal gyrys, insula) or to temporoparietal or inferior frontal areas, especially in the right hemisphere (Stewart et al. 2006).

Interestingly, amusia can occur independently of or in parallel with linguistic disorders, thereby raising an intriguing question of whether the neural mechanisms of music and speech processing are separate or shared. In studies of brain damaged patients, approximately half of the patients with acquired amusia have been documented to have at least minor aphasia (Stewart et al. 2006), although there are also cases of clear double dissociations (amusia without aphasia and vice versa), suggesting that there may be separate neural modules for music and speech (Peretz and Coltheart 2003). Recent studies, however, have found that aphasic patients have difficulties in perceiving musical structures (Patel et al. 2008) and, conversely, that individuals with congenital amusia have difficulties in perceiving the intonation and prosody of speech (Liu et al. 2010; Jiang et al. 2012), thereby supporting the views about the commonalities between speech and music perception at the neural level.

4 Music Across Life Span

Music, especially hearing, singing, and producing musical sounds, appears to evoke the natural interest of infants and children across cultures (Trehub 2003). Indeed, babies seem to be born with innate musical abilities: even small infants can detect the pitch, timbre, and duration of the sounds, recognize familiar melodic and rhythmic patterns, and prefer consonant over dissonant music and singing over speech (Trehub 2003). Infants are also sensitive to prosody, in other words, to changes in the melody, rhythm, stress, and intonation of speech, which are used to communicate emotions and to emphasize word meanings in speech. Intuitively, parents tend to speak to their babies in a manner which utilizes this sensitivity. In fact, infant-directed speech (or motherese or parentese) contains many musical or singing-like elements, such as strong pitch fluctuations and repetitive melodic lines, which help the infant to grasp and acquire the essential structure of natural speech (Nakata and Trehub 2004). Lullabies and play songs are also globally used to modulate the arousal level of infants, as reflected, for example, in salivary cortisol changes (Shenfield et al. 2003). At the age of 6 months, babies start to babble and to "dance", i.e., to adjust their movements with the tempo of music (Zentner and Eerola 2010). For a toddler, musical activity is a playground of sorts, where parents can use reciprocal communication and rhythmic movements to regulate the emotional and attentional state of the child. At the same time, the child him/herself can practice the cognitive, motor, and social skills needed for speech acquisition and communication.

At preschool age, children are often enthusiastic in expressing music with their gestures and movements and in taking part in musical activities as listeners, singers, players, and dancers. In many native cultures, music making or dancing is an integral and natural part of the everyday life of children. For the developing brain, repeated exposure to music in the childhood environment can be beneficial. In developmental animal studies, an enriched auditory environment that contains complex sounds or music, has repeatedly and consistently been shown to improve auditory functions, learning and memory as well as induce neural plasticity, as indicated by changes in neurotransmitter (e.g., dopamine, glutamate) and neurotrophin (e. g., brain-derived neurotrophic factor, BDNF) levels, synaptic plasticity, and neurogenesis (Angelucci et al. 2007; Bose et al. 2010; Rickard et al. 2005). According to studies on children, musical hobbies can improve auditory and motor skills (Hyde et al. 2009) as well as high-level cognitive skills such as logical reasoning, executive functioning, attention, and memory (Hannon and Trainor 2007; Schellenberg 2004; Moreno et al. 2011). Musical skills and music training seem to be also related to enhanced neural processing of speech and improved language skills, such as reading, speech segmentation, perceiving speech in noise, and pronunciation of a foreign language (Besson et al. 2011; Kraus and Chandrasekaran 2010; Milovanov and Tervaniemi 2011). At the neural level, structural changes in the auditory cortex, the motor cortex, and the corpus callosum have been observed already after 15 months of individual piano lessons (Hyde et al. 2009).

During adolescence, music serves as a forum for constructing the developing self-identity, forming interpersonal relationships, experiencing agency and self-control, and in dealing with negative emotions and stress (Saarikallio and Erkkilä 2007; North et al. 2000). Furthermore, a key aspect of all musical activity is emotional expression, which, according to recent theory (Molnar-Szakacs and Overy 2006), is at least partly mediated by the *mirror neuron system*, a set of frontal and parietal cortical structures, which are thought to contribute to the understanding the actions of other people (i.e., empathy), learning new skills by imitation, and to theory of mind, and which continue to develop through adolescence and early adulthood. Musicking, or even simple music listening, can thus form a safe shared and dynamic platform for exploring one's emotional processes with respect to others and for forging relationships through common experiences, chats, and discussions. Some evolutionary theories of music postulate that joint musical activities, such as singing and dancing with others, facilitate the release of endorphins and the experience or reward and pleasure, which in turn promote group cohesion and social bonding (Dunbar 2003). In children, there is already some empirical evidence for the emergence of prosocial behavior after joint music making (Kirschner and Tomasello 2010).

Finally, music has a lot to give also in adulthood and in old age. In most cases, individual musical preferences are formed during adolescence and early adulthood—maybe because of this, music also offers means to refresh and process memories and reflect on prior experiences later in life. During adulthood, music is strongly linked to emotional and self-conceptual processing, mood, and memories (Saarikallio 2010). Music continues to play a vital role as well during aging. Studies suggest that regular musical activities are very important to seniors in maintaining psychological well-being and in contributing to positive aging by providing ways to maintain self-esteem, competence, and independence and in reducing loneliness and isolation (Cohen et al. 2002; Hays and Minichiello 2005). Interestingly, studies of healthy seniors have also showed that regular musical activities, such as dancing or playing an instrument, are cognitively beneficial (Bugos et al. 2007; Kattenstroth et al. 2010) and can also reduce the risk of developing dementia later (Verghese et al. 2003). Similarly, one recent study reported that musicians performed better than non-musicians on tasks of executive functioning, memory, visuospatial judgement, and motor dexterity in the old age (while controlling for general lifestyle activities), suggesting that musical activity is associated with successful cognitive aging (Hanna-Pladdy and Gajewski 2012).

5 Music as a Form of Therapy and Rehabilitation

Broadly defined, *music therapy* is an intervention provided by a trained music therapist where music is used in a therapeutic interaction with the client to achieve individually defined goals. The scientific study of the efficacy of different music interventions has increased rapidly during the past 20 years, and the experimental

evidence for these interventions is accumulating regarding their clinical utility and applicability in the treatment and rehabilitation of many somatic, psychiatric, and neurological illnesses. In the following, we will briefly review what is currently known about the efficacy of music interventions regarding five domains: emotion, attention and sensory functions, memory, communication, and motor functions.

Emotion. Music has a powerful emotional effect, which is manifested both psychologically and physiologically (e.g., in heart rate, respiration, skin temperature, and hormone secretion, limbic/paralimbic brain activation (Juslin and Västfjäll 2008; Lundqvist et al. 2009; Koelsch 2010). Various music interventions have been applied in the rehabilitation of person suffering from various affective disorders, such as depression and anxiety, or from illnesses with more severe neuropsychiatric symptoms, such as schizophrenia or Alzheimer's disease (AD). For depressed patients, studies suggest that music therapy is an applicable method that can alleviate depression and anxiety symptoms and improve mood (Gold et al. 2009; Maratos et al. 2008). Recently, a novel therapeutic technique called *improvisational psychodynamic music therapy*, which utilizes musical improvisation and interaction in a psychodynamic context, was found to be effective on reducing depression and anxiety and improving general functioning in working-age depressed patients (Erkkilä et al. 2011). For schizophrenic patients, music therapy can help improve the global and mental state and the social functioning of the patients and to reduce their negative symptoms, depression, and anxiety (Mössler et al. 2011). For persons suffering from AD or other form of dementia, music therapy may be effective in reducing neuropsychiatric and behavioural symptoms, such as agitation and wandering, as well as in enhancing social and emotional functioning, although more methodologically robust studies are still needed (Vink et al. 2011). Finally, music interventions can be effective in reducing anxiety, improving mood, and influencing various autonomic nervous system functions (heart rate, respiration, blood pressure) also in patients suffering from severe chronic somatic illnesses, including cancer (Bradt et al. 2011) and coronary heart disease (Bradt and Dileo 2009).

Attention and sensory functions. Music has a unique capacity to draw and direct attention and influence arousal and vigilance. Clinically, this attribute has been effectively utilized in the alleviation of pain (Bernatzky et al. 2011). Evidence suggests that musical interventions are able to reduce pain intensity and also to reduce the amount of opioid medication in many pain conditions, especially in post-surgical pain (Cepeda et al. 2006). Another application of music is the treatment of tinnitus using a novel treatment strategy called *tailor-made notched music training*. In this training, patients listen on a daily basis to specially-made music where the energy spectrum of the music is notched around the individual tinnitus frequency. Currently, behavioural and brain-imaging evidence suggests that subjective tinnitus loudness and annoyance and tinnitus-related auditory evoked fields can be significantly reduced by the training (Okamoto et al. 2010). These results are important because tinnitus is a highly common and often very debilitating disorder for which there is no effective drug treatment. Music listening has also been utilized to mediate attention and arousal in children with attention

deficit hyperactivity disorder (ADHD) and in stroke patients suffering from unilateral spatial neglect, a disorder of awareness of the contralesional space. In ADHD children, the presence of background music has been found to reduce distractibility and enhance concentration on a school task (Abikoff et al. 1996; Pelham et al. 2011). In neglect patients, listening to pleasant music can temporarily overcome the bias in spatial attention and reduce neglect (Hommel et al. 1990; Soto et al. 2009).

Memory. Listening to music naturally entails keeping track of the incoming auditory information (auditory sensory memory, working memory) as well as analyzing the musical structure and the identity of the piece (long-term semantic memory) and retrieving the experiences and memories associated with it (long-term episodic memory). By affecting mood and arousal state (Thompson et al. 2001), music can temporarily improve memory performance in healthy subjects (Greene et al. 2010). In persons with AD, exposure to stimulating music can temporarily improve autobiographical recall, especially from the remote life era (Irish et al. 2006; Foster and Valentine 2001), and songs can function as mnemonic aids for recalling verbal material (Simmons-Stern et al. 2010). Also after an acute stroke, daily music listening can enhance the recovery of auditory and verbal memory (Särkämö et al. 2008, 2010). The positive effects on memory were also coupled with reduced depression and confusion during the early recovery stage (Särkämö and Soto 2012), suggesting that the positive effect of music on cognition is at least partly mediated by enhanced mood.

Communication. Both music and speech are forms of communication that make use of the acoustic properties of sound, such as pitch, timbre, and rhythm. Music has been clinically utilized in enhancing verbal communication in persons with developmental or acquired neurological disorder, such as autistic spectrum disorder or stroke. Children with autism typically lack communication skills but may have enhanced auditory and musical abilities (Ouimet et al. 2012). Currently, there is preliminary evidence that music therapy may help autistic children to improve their communicative skills (Gold et al. 2006). Remarkably, a recently developed method called *auditory-motor mapping training*, which utilizes song-like intonation and bimanual motor activities, has been observed to improve the articulation of non-trained words and phrases in autistic children who were entirely non-verbal at the beginning of the training (Wan et al. 2011). Another example of a music-based rehabilitation method that emphasizes the melodic and rhythmic elements of speech, is *melodic intonation therapy*, which has been utilized in the rehabilitation of aphasia caused by a left hemisphere lesion (Norton et al. 2009). Although the clinical efficacy of this therapy has yet to be substantiated, evidence from small case series suggests that an intense course of melodic intonation therapy can lead to improvements in speech production (Schlaug et al. 2010; Zipse et al. 2012) and to various neuroplastic changes in the spared right frontotemporal network (Schlaug et al. 2010; Zipse et al. 2012).

Motor functions. Rhythm and movement are intimately linked to music; in fact, some cultures do not even differentiate "music" and "dance" in their vocabulary. Also in the brain, almost all musical activity, even the passive listening of music,

automatically recruits motor areas, and there is rich connectivity between auditory and motor brain areas (Zatorre et al. 2007). Clinically, our innate tendency to sequence and entrain movements to the beat of music has been utilized in the rehabilitation of walking in many neurological illnesses including stroke, traumatic brain injury, and Parkinson's disease (PD). *Rhythmic auditory stimulation*, in which an external auditory rhythm is provided by a metronome or music tapes and adapted to the movements of the patient, has been shown to be beneficial for improving gait in hemiparetic stroke patients (Bradt et al. 2010) and in PD patients (McIntosh et al. 1997; Hove et al. 2012). Interestingly, some evidence suggests that also hearing familiar stimulating music can temporarily enhance motor coordination in PD patients (Bernatzky et al. 2004; Sacrey et al. 2009). Another way to use music in motor rehabilitation is to utilize active music making in the form of instrument playing. Recently, a method called *music-supported therapy* has been developed where movements of the affected upper extremity are trained by playing a piano keyboard or electronic drum set. Music-supported therapy has been shown to be effective in improving the speed, precision, and smoothness of arm movements, enhancing the recovery of fine and gross motor skills, and facilitating the coherence and functional connectivity of auditory-motor networks in the frontotemporal cortex (Schneider et al. 2010; Altenmüller et al. 2009; Rodriguez-Fornells et al. 2012; Rojo et al. 2011).

In summary, the clinical evidence for the effectiveness of music interventions in alleviating emotional, cognitive, communicative, and motor deficits has increased substantially during the last years. Currently, research in the fields of music therapy, psychology, and cognitive and affective neuroscience is beginning to merge, and there are now on-going multidisciplinary studies in many countries aimed towards determining the clinical impact of music and uncovering its underlying neural mechanisms.

6 Concluding Remarks

In this article, we have reviewed a number of studies which together shed light on the neural basis, development and rehabilitative use of music. Modern neuroimaging has shown that musical activities, ranging from simple music listening to singing and playing a musical instrument, have diverse positive effects on the structure and function of the brain. Musical activities have different roles and meanings in different phases of life: during infancy and early childhood, they can support speech development; during school years, they can develop cognitive and attentional skills; during adolescence, they help to build self-identity and enhance emotional self-regulation; and during adulthood and old age, they help maintain cognitive performance and memory and improve mood. Clinically, the use of music therapy and other music interventions as a form of treatment and rehabilitation has received scientific support especially in somatic, psychiatric, and neurological illnesses involving deficits in emotions, attention, and sensory functions, memory,

communication, and motor functions. In summary, although the research field is still relatively young and more studies are still needed, music can be considered as a viable and promising non-pharmacological form of treatment and rehabilitation and, more generally, as an enriching and useful hobby that can shape the development and maintain the healthy functioning of the brain across life.

References

Abikoff, H., Courtney, M. E., Szeibel, P. J., & Koplewicz, H. S. (1996). The effects of auditory stimulation on the arithmetic performance of children with ADHD and nondisabled children. *Journal of Learning Disabilities, 29*, 238–246.

Altenmüller, E., Marco-Pallares, J., Münte, T. F., & Schneider, S. (2009). Neural reorganization underlies improvement in stroke-induced motor dysfunction by music-supported therapy. *Annals of the New York Academy of Sciences, 1169*, 395–405.

Angelucci, F., Fiore, M., Ricci, E., Padua, L., Sabino, A., & Tonali, P. A. (2007). Investigating the neurobiology of music: Brain-derived neurotrophic factor modulation in the hippocampus of young adult mice. *Behavioural Pharmacology, 18*, 491–496.

Ayotte, J., Peretz, I., Rousseau, I., Bard, C., & Bojanowski, M. (2000). Patterns of music agnosia associated with middle cerebral artery infarcts. *Brain, 123*, 1926–1938.

Bernatzky, G., Presch, M., Anderson, M., & Panksepp, J. (2011). Emotional foundations of music as a non-pharmacological pain management tool in modern medicine. *Neuroscience and Biobehavioral Reviews, 35*, 1989–1999.

Bernatzky, G., Bernatzky, P., Hesse, H. P., Staffen, W., & Ladurner, G. (2004). Stimulating music increases motor coordination in patients afflicted with Morbus Parkinson. *Neuroscience Letters, 361*, 4–8.

Besson, M., Chobert, J., & Marie, C. (2011). Transfer of training between music and speech: Common processing, attention, and memory. *Frontiers in Psychology, 2*, 94.

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences United States of America, 98*, 11818–11823.

Bose, M., Muñoz-Llancao, P., Roychowdhury, S., Nichols, J. A., Jakkamsetti, V., Porter, B., et al. (2010). Effect of the environment on the dendritic morphology of the rat auditory cortex. *Synapse, 64*, 97–110.

Bradt, J., & Dileo, C. (2009). Music for stress and anxiety reduction in coronary heart disease patients. *Cochrane Database of Systematic Reviews, 2*, CD006577.

Bradt, J., Dileo, C., Grocke, D., & Magill, L. (2011). Music interventions for improving psychological and physical outcomes in cancer patients. *Cochrane Database of Systematic Reviews*, 8, CD006911.

Bradt, J., Magee, W. L., Dileo, C., Wheeler, B. L., & McGilloway, E. (2010). Music therapy for acquired brain injury. *Cochrane Database of Systematic Reviews*, 7, CD006787.

Bugos, J. A., Perlstein, W. M., McCrae, C. S., Brophy, T. S., & Bedenbaugh, P. H. (2007). Individualized piano instruction enhances executive functioning and working memory in older adults. *Aging and Mental Health, 11*, 464–471.

Cepeda, M. S., Carr, D. B., Lau, J., & Alvarez, H. (2006). Music for pain relief. *Cochrane Database of Systematic Reviews, 2*, CD004843.

Cohen, A., Bailey, B., & Nilsson, T. (2002). The importance of music to seniors. *Psychomusicology, 18*, 89–102.

Conard, N. J., Malina, M., & Münzel, S. C. (2009). New flutes document the earliest musical tradition in southwestern Germany. *Nature, 460*, 737–740.

Dunbar, R. I. (2003). The origin and subsequent evolution of language. In M. H. Christiansen & S. Kirby (Eds.), *Language evolution* (pp. 219–234). Oxford: Oxford University Press.

Gold, C., Wigram, T., & Elefant, C. (2006). Music therapy for autistic spectrum disorder. *Cochrane Database of Systematic Reviews, 2*, CD004381.

Gold, C., Solli, H. P., Krüger, V., & Lie, S. A. (2009). Dose-response relationship in music therapy for people with serious mental disorders: Systematic review and meta-analysis. *Clinical Psychology Review, 29*, 193–207.

Goll, J. C., Crutch, S. J., & Warren, J. D. (2010). Central auditory disorders: Toward a neuropsychology of auditory objects. *Current Opinion in Neurology, 23*, 617–627.

Erkkilä, J., Punkanen, M., Fachner, J., Ala-Ruona, E., Pöntiö, I., Tervaniemi, M., et al. (2011). Individual music therapy for depression: Randomised controlled trial. *British Journal of Psychiatry, 199*, 132–139.

Grahn, J. A., & Rowe, J. B. (2009). Feeling the beat: Premotor and striatal interactions in musicians and nonmusicians during beat perception. *Journal of Neuroscience, 29*, 7540–7548.

Greene, C. M., Bahri, P., & Soto, D. (2010). Interplay between affect and arousal in recognition memory. *PLoS One, 5*, e11739.

Foxton, J. M., Dean, J. L., Gee, R., Peretz, I., & Griffiths, T. D. (2004). Characterization of deficits in pitch perception underlying 'tone deafness'. *Brain, 127*, 801–810.

Foster, N. A., & Valentine, E. R. (2001). The effect of auditory stimulation on autobiographical recall in dementia. *Experimental Aging Research, 27*, 215–228.

Drayna, D., Manichaikul, A., de Lange, M., Snieder, H., & Spector, T. (2001). Genetic correlates of musical pitch recognition in humans. *Science, 291*, 1969–1972.

Hall, D. A., Hart, H. C., & Johnsrude, I. S. (2003). Relationships between human auditory cortical structure and function. *Audiology and Neurootology, 8*, 1–18.

Hanna-Pladdy, B., & Gajewski, B. (2012). Recent and past musical activity predicts cognitive aging variability: Direct comparison with general lifestyle activities. *Frontiers in Human Neuroscience, 6*, 198.

Hannon, E. E., & Trainor, L. J. (2007). Music acquisition: Effects of enculturation and formal training on development. *Trends in Cognitive Science, 11*, 466–472.

Hays, T., & Minichiello, V. (2005). The meaning of music in the lives of older people: A qualitative study. *Psychology of Music, 33*, 437–451.

Henry, M. J., & McAuley, J. D. (2010). On the prevalence of congenital amusia. *Music Perception, 27*, 413–418.

Hommel, M., Peres, B., Pollak, P., Memin, B., Besson, G., Gaio, J. M., et al. (1990). Effects of passive tactile and auditory stimuli on left visual neglect. *Archives of Neurology, 47*, 573–576.

Hove, M. J., Suzuki, K., Uchitomi, H., Orimo, S., & Miyake, Y. (2012). Interactive rhythmic auditory stimulation reinstates natural 1/f timing in gait of Parkinson's patients. *PLoS One, 7*, e32600.

Hyde, K. L., Lerch, J. P., Zatorre, R. J., Griffiths, T. D., Evans, A. C., & Peretz, I. (2007). Cortical thickness in congenital amusia: When less is better than more. *Journal of Neuroscience, 27*, 13028–13032.

Hyde, K. L., Zatorre, R. J., & Peretz, I. (2011). Functional MRI evidence of an abnormal neural network for pitch processing in congenital amusia. *Cerebral Cortex, 21*, 292–299.

Hyde, K. L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A. C., et al. (2009). Musical training shapes structural brain development. *Journal of Neuroscience, 29*, 3019–3025.

Irish, M., Cunningham, C. J., Walsh, J. B., Coakley, D., Lawlor, B. A., Robertson, I. H., et al. (2006). Investigating the enhancing effect of music on autobiographical memory in mild Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders, 22*, 108–120.

Janata, P. (2009). The neural architecture of music-evoked autobiographical memories. *Cerebral Cortex, 19*, 2579–2594.

Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B., & Bharucha, J. J. (2002a). The cortical topography of tonal structures underlying Western music. *Science, 298*, 2167–2170.

Janata, P., Tillmann, B., & Bharucha, J. J. (2002b). Listening to polyphonic music recruits domain-general attention and working memory circuits. *Cognitive Affective and Behavioral Neuroscience, 2*, 121–140.

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., Chen, X., & Yang, Y. (2012). Amusia results in abnormal brain activity following inappropriate intonation during speech comprehension. *PLoS One, 7*, e41411.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences, 31*, 559–575.

Kattenstroth, J. C., Kolankowska, I., Kalisch, T., & Dinse, H. R. (2010). Superior sensory, motor, and cognitive performance in elderly individuals with multi-year dancing activities. *Frontiers in Aging Neuroscience, 2*, e1–e9.

Kirschner, S., & Tomasello, M. (2010). Joint music making promotes prosocial behavior in 4-year-old children. *Evolution and Human Behavior, 31*, 354–364.

Koelsch, S. (2010). Towards a neural basis of music-evoked emotions. *Trends in Cognitive Sciences, 14*, 131–137.

Koelsch, S., & Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences, 9*, 578–584.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience, 11*, 599–605.

LeDoux, J. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155–184.

Liu, F., Patel, A. D., Fourcin, A., & Stewart, L. (2010). Intonation processing in congenital amusia: Discrimination, identification and imitation. *Brain, 133*, 1682–1693.

Loui, P., Alsop, D., & Schlaug, G. (2009). Tone deafness: A new disconnection syndrome? *Journal of Neuroscience, 29*, 10215–10220.

Lundqvist, L. O., Carlsson, F., Hilmersson, P., & Juslin, P. N. (2009). Emotional responses to music: Experience, expression, and physiology. *Psychology of Music, 37*, 61–90.

Maratos, A., Gold, C., Wang, X., & Crawford, M. (2008). Music therapy for depression. *Cochrane Database of Systematic Reviews*, 1, CD004517.

McIntosh, G. C., Brown, S. H., Rice, R. R., & Thaut, M. H. (1997). Rhythmic auditory-motor facilitation of gait patterns in patients with Parkinson's disease. *Journal of Neurology Neurosurgery and Psychiatry, 62*, 22–26.

Menon, V., & Levitin, D. J. (2005). The rewards of music listening: Response and physiological connectivity of the mesolimbic system. *Neuroimage, 28*, 175–184.

Milovanov, R., & Tervaniemi, M. (2011). The interplay between musical and linguistic aptitudes: A review. *Frontiers in Psychology, 2*, 321.

Mithen, S. (2005). *The singing Neanderthals: The origins of music, language, mind and body*. London: Weidenfeld & Nicolson.

Molnar-Szakacs, I., & Overy, K. (2006). Music and mirror neurons: From motion to'e'motion. *Social Cognitive and Affective Neuroscience, 1*, 235–241.

Moreno, S., Bialystok, E., Barac, R., Schellenberg, E. G., Cepeda, N. J., & Chau, T. (2011). Short-term music training enhances verbal intelligence and executive function. *Psychological Science, 22*, 1425–1433.

Mössler, K., Chen, X., Heldal, T. O., Gold, C. (2011). Music therapy for people with schizophrenia and schizophrenia-like disorders. *Cochrane Database of Systematic Reviews*, 12, CD004025.

Nakata, T., & Trehub, S. (2004). Infants' responsiveness to maternal speech and singing. *Infant Behavior and Development, 27*, 455–464.

North, A. C., Hargreaves, D. J., & O'Neill, S. A. (2000). The importance of music to adolescents. *British Journal of Educational Psychology, 70*, 255–272.

Norton, A., Zipse, L., Marchina, S., & Schlaug, G. (2009). Melodic intonation therapy: Shared insights on how it is done and why it might help. *Annals of the New York Academy of Sciences, 1169*, 431–436.

Okamoto, H., Stracke, H., Stoll, W., & Pantev, C. (2010). Loudness and tinnitus-related auditory listening to tailor-made notched music reduces tinnitus cortex activity. *Proceedings of the National Academy of Sciences United States of America, 107*, 1207–1210.

Ouimet, T., Foster, N. E., Tryfon, A., & Hyde, K. L. (2012). Auditory-musical processing in autism spectrum disorders: A review of behavioral and brain imaging studies. *Annals of the New York Academy of Sciences, 1252*, 325–331.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience, 6*, 674–681.

Patel, A. D., Iversen, J. R., Wassenaar, M., & Hagoort, P. (2008). Musical syntactic processing in agrammatic Broca's aphasia. *Aphasiology, 22*, 776–789.

Pelham, W. E, Jr, Waschbusch, D. A., Hoza, B., Gnagy, E. M., Greiner, A. R., Sams, S. E., et al. (2011). Music and video as distractors for boys with ADHD in the classroom: Comparison with controls, individual differences, and medication effects. *Journal of Abnormal Child Psychology, 39*, 1085–1098.

Peretz, I. (2006). The nature of music from a biological perspective. *Cognition, 100*, 1–32.

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience, 6*, 688–691.

Peretz, I., Cummings, S., & Dubé, M. P. (2007). The genetics of congenital amusia (tone deafness): A family-aggregation study. *American Journal of Human Genetics, 81*, 582–588.

Pickles, J. (2008). *An introduction to the physiology of hearing*. Emerald: Bingley.

Platel, H., Baron, J. C., Desgranges, B., Bernard, F., & Eustache, F. (2003). Semantic and episodic memory of music are subserved by distinct neural networks. *Neuroimage, 20*, 244–256.

Rickard, N. S., Toukhsati, T. R., & Field, S. E. (2005). The effect of music on cognitive performance: Insight from neurobiological and animal studies. *Behavioural and Cognitive Neuroscience Reviews, 4*, 235–261.

Rodriguez-Fornells, A., Rojo, N., Amengual, J. L., Ripollés, P., Altenmüller, E., & Münte, T. F. (2012). The involvement of audio-motor coupling in the music-supported therapy applied to stroke patients. *Annals of the New York Academy of Sciences, 1252*, 282–293.

Rojo, N., Amengual, J., Juncadella, M., Rubio, F., Camara, E., Marco-Pallares, J., et al. (2011). Music-supported therapy induces plasticity in the sensorimotor cortex in chronic stroke: A single-case study using multimodal imaging (fMRI-TMS). *Brain Injury, 25*, 787–793.

Saarikallio, S. (2010). Music as emotional self-regulation throughout adulthood. *Psychology of Music, 39*, 307–327.

Saarikallio, S., & Erkkilä, J. (2007). The role of music in adolescents' mood regulation. *Psychology of Music, 35*, 88–109.

Sacrey, L. A., Clark, C. A., & Whishaw, I. Q. (2009). Music attenuates excessive visual guidance of skilled reaching in advanced but not mild Parkinson's disease. *PLoS One, 4*, e6841.

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience, 14*, 257–262.

Särkämö, T., & Soto, D. (2012). Music listening after stroke: Beneficial effects and potential neural mechanisms. *Annals of the New York Academy of Sciences, 1252*, 266–281.

Särkämö, T., Tervaniemi, M., Laitinen, S., Forsblom, A., Soinila, S., Mikkonen, M., et al. (2008). Music listening enhances cognitive recovery and mood after middle cerebral artery stroke. *Brain, 131*, 866–876.

Särkämö, T., Pihko, E., Laitinen, S., Forsblom, A., Soinila, S., Mikkonen, M., et al. (2010). Music and speech listening enhance the recovery of early sensory processing after stroke. *Journal of Cognitive Neuroscience, 22*, 2716–2727.

Särkämö, T., Tervaniemi, M., Soinila, S., Autti, T., Silvennoinen, H. M., Laine, M., et al. (2009). Cognitive deficits associated with acquired amusia after stroke: A neuropsychological follow-up study. *Neuropsychologia, 47*, 2642–2651.

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science, 15*, 511–514.

Schlaug, G., Norton, A., Marchina, S., Zipse, L., & Wan, C. Y. (2010). From singing to speaking: Facilitating recovery from nonfluent aphasia. *Future Neurology, 5*, 657–665.

Schneider, S., Münte, T., Rodriguez-Fornells, A., Sailer, M., & Altenmüller, E. (2010). Music-supported training is more efficient than functional motor training for recovery of fine motor skills in stroke patients. *Music Perception, 27*, 271–280.

Schuppert, M., Münte, T. F., Wieringa, B. M., & Altenmüller, E. (2000). Receptive amusia: Evidence for cross-hemispheric neural networks underlying music processing strategies. *Brain, 123*, 546–559.

Shenfield, T., Trehub, S., & Nakata, T. (2003). Maternal singing modulates infant arousal. *Psychology of Music, 31*, 365–375.

Simmons-Stern, N. R., Budson, A. E., & Ally, B. A. (2010). Music as a memory enhancer in patients with Alzheimer's disease. *Neuropsychologia, 48*, 3164–3167.

Soto, D., Funes, M. J., Guzmán-García, A., Warbrick, T., Rotshtein, P., & Humphreys, G. W. (2009). Pleasant music overcomes the loss of awareness in patients with visual neglect. *Proceedings of the National Academy of Sciences United States of America, 106*, 6011–6016.

Stewart, L., von Kriegstein, K., Warren, J. D., & Griffiths, T. D. (2006). Music and the brain: Disorders of musical listening. *Brain, 129*, 2533–2553.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2001). Arousal, mood, and the Mozart effect. *Psychological Science, 12*, 248–251.

Trehub, S. E. (2003). The developmental origins of musicality. *Nature Neuroscience, 6*, 669–673.

Verghese, J., Lipton, R. B., Katz, M. J., Hall, C. B., Derby, C. A., Kuslansky, G., et al. (2003). Leisure activities and the risk of dementia in the elderly. *New England Journal of Medicine, 348*, 2508–2516.

Vink, A. C., Bruinsma, M. S., & Scholten, R. J. (2011). Music therapy for people with dementia. *Cochrane Database of Systematic Reviews*, *3*, CD003477.

Wan, C. Y., Bazen, L., Baars, R., Libenson, A., Zipse, L., Zuk, J., et al. (2011). Auditory-motor mapping training as an intervention to facilitate speech output in non-verbal children with autism: A proof of concept study. *PLoS One, 6*, e25505.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences, 6*, 37–46.

Zatorre, R. J., Chen, J. L., & Penhune, V. P. (2007). When the brain plays music: Auditory-motor interactions in music perception and production. *Nature Reviews Neuroscience, 8*, 547–558.

Zatorre, R. J., Evans, A. C., & Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *Journal of Neuroscience, 14*, 1908–1919.

Zentner, M., & Eerola, T. (2010). Rhythmic engagement with music in infancy. *Proceedings of the National Academy of Sciences United States of America, 107*, 5768–5773.

Zipse, L., Norton, A., Marchina, S., & Schlaug, G. (2012). When right is all that is left: Plasticity of right-hemisphere tracts in a young aphasic patient. *Annals of the New York Academy of Sciences, 1252*, 237–245.

The Perception of Melodies: Some Thoughts on Listening Style, Relational Thinking, and Musical Structure

Christiane Neuhaus

Every day, countless numbers of pop songs, ballads, and classical themes are broadcasted via radio stations, sung at schools, or rehearsed by professional ensembles. Given an almost infinite variety of melodies, we may ask which attributes they have in common, and what features we can extract. Composers, however, often decide for the word 'motif' rather than 'melody' to explain their musical ideas, partly due to different sequence length, and partly because both terms point to a different character of tone progression: A 'motif' has the potential for development because of clearly recognizable rhythmic and interval features, enabling us even to build the whole musical architecture. A 'melody', by contrast, is often considered a closed entity (or integrated whole) with elements kept in balance.

In the following chapter we focus on the perception of melodic structure. We first try to define the word 'melody' (Greek: *melōdía*, 'singing, chanting') from the viewpoint of music theory. In my opinion, four structural properties determine the melody's nature, namely 'interval type', 'melodic contour', 'balance between form parts', as well as the 'underlying harmonic framework' (based on the piece's tonality). The first three attributes describe a melody in horizontal direction whereas the fourth refers to its vertical dimension.

Let us quickly go through the details: The word 'interval type' describes the type of tone combination that predominates over several bars. In this regard, most tonal Western melodies either make use of scale segments (or tone steps; in German: 'Skalenmelodik') or of triads in succession (in German: 'Dreiklangsmelodik'), or of a combination of both. The second property, named 'melodic contour' (first mentioned by Abraham and Hornbostel in 1909), refers to the outline of a melody. It describes the curve of melodic movement in rough lines without considering detailed interval progression. For the majority of Western tunes the following contour types are characteristic: continuously ascending or descending, undulating around the pitch axis, or having an arch-like shape (e.g. Adams 1976).

C. Neuhaus (✉)
Systematic musicology, University of Hamburg, Hamburg, Germany
e-mail: chr_neuhaus@t-online.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_8,

From the empirical point of view White (1960) argues convincingly that 'Happy birthday' and other familiar tunes that have been altered with some algebraic procedures can easily be identified when transformations are linear, i.e. when preserving contour while the opposite holds true when transformations are non-linear, i.e. when distorting contour. In detail, subtracting, adding, or multiplying an integer, e.g. i(+)1 or 2i(−)1, keep tunes recognizable but whenever intervals are allowed to change sign, recognition is poor.

Now let us think on 'balance between form parts' which is the third structural property that contributes to the quality of a melody. In this regard lots of tunes are restricted to eight bars which is the common length of the 'musical period'. This type of musical form can further be split up into segments of 4 plus 4, both parts revealing some symmetry due to similar melodic beginnings. (On the empirical findings on musical form perception I will report in detail several pages later).

The fourth structure-determining factor is the assignment of a melody to a latent harmonic framework which is mostly concurrent with the cadence scheme. This assignment allows us to evaluate each tone in its tension towards the harmonic basis, in particular towards inherent vertical anchor points such as the tonic or the dominant, often arousing feelings of tension or relaxation. The involvement of vertical aspects into the horizontal line particularly becomes apparent in the improvisation practice of North Indian music. Here, each tone of a raga becomes established in the listener's mind by constantly playing the drone, i.e. the tonic tone, on the tanpura. This way, distance estimation in vertical direction can be performed easily (see e.g. Danielou 1982).

However, most encyclopedias do not define 'melody' by mere description or analysis of its structural properties but also consider the perceptual side with a particular focus on Gestalt psychology (c.f. van Dyke Bingham 1910; Ziegenrücker 1979). The Gestalt school of psychology (represented by Wertheimer, Krüger, Koffka, and others) has two principles at hand to determine how melodies are perceived, namely the experience of 'unity or completeness' (also known as the 'law of closure'), and the quality of 'structuredness' ('Gegliedertheit') of auditory sequences. The first attribute primarily refers to perceiving as a cognitive process whereas the second is a structural property inherent in the material as such.

Let us briefly explain the points: The impression of completeness arises from cohesive forces regarding the melody's inner structure. Cohesiveness is high when tones are in close distance (i.e. in spatial and temporal proximity) to each other but low when tones are scattered over several octaves. This holistic approach, considering melodies as Gestalt entities, once more becomes obvious through the 'law of transposition', first articulated by the Austrian philosopher Christian von Ehrenfels (1890), the founder of Gestalt theory. According to this perceptual law, melodies remain invariable and can still be recognized when shifting the entire figure to another key.

The second principle that determines how melodies are perceived is based on (pre-)'structuredness' ('Gegliedertheit') as an inherent property of time-based art per se (e.g. West et al. 1991). 'Pre-structuredness' enables the active mind to

subdivide sound patterns into perceivable portions or musical phrases. Interestingly, this perceptual process of segmentation has an equivalent on the neuronal level: For the perception of musical phrase boundaries, a neural correlate has been identified by using event-related potentials as the method of study. This correlate is called Closure Positive Shift (abbreviated CPS), it has a counterpart in the perception of speech. Whenever attention is on the detail, musicians react in the way just described whereas non-musicians respond with an early negativity (cf. Knösche et al. 2005; Neuhaus et al. 2006). This indicates that qualitatively different ERP components (CPS and early negativity) reflect some top-down activation of general but dissimilar phrasing schemata depending on prior musical training.

Besides this issue of (pre-)structuredness which helps to segment and encode the complete musical sequence we should also become aware of another form of experiencing the whole, named 'complex quality' which is a term coined by Krüger in 1926. Complex qualities are indivisible entities of purely sensual character that are evoked by emotions or sensations. This difference in holistic experience—caused by (pre-)structured and perceptually groupable objects on the one hand and indivisible emotions on the other—is indeed the main point that distinguishes the 'Berlin School of Gestaltpsychologie' (represented by Wertheimer, Köhler, and Koffka) from the 'Leipzig School of Ganzheitspsychologie' (represented by Krüger and Wundt).

When further elaborating on pre-structuredness, two comments have to be mentioned here, the first made by the Czech musicologist Karbusicky (1979), the second made by Levitin (2009), an American sound producer and neuroscientist. Karbusicky focusses on the difference between the words 'tone structure' and 'tone row'. As mentioned above, 'tone structure' indicates that notes are grown together into an entity whereas 'tone row' describes that notes are loosely connected so that elements can arbitrarily be replaced. Levitin (2009), by contrast, draws our attention to the emotional aspect of this issue, emphasizing that "a randomized or 'scrambled' sequence of notes is not able to elicit the same [emotional] reactions as an ordered one" (p. 10). Levitin focusses in particular on the structure's temporal aspects and their respective neural correlates. By using functional imaging methods he could prove that two brain regions of the inferior frontal cortex—named Brodmann Area 47 and its right-hemispheric homologue—are significantly activated when tone structure is preserved but do not react to arbitrarily scrambled counterparts, having the same spectral energy but lacking any temporal coherence (Levitin and Menon 2005). In addition, Neuhaus and Knösche (2008) could demonstrate by using event-related potentials that the brain responds to time-preserving versus time-permuting tone sequences from a very early stage of sequence processing on. In detail, lack of time order causes a larger increase of the P1 component already around 50 ms (measured from sequence onset) while no such difference in brain activity beyond 250 ms could be observed. In this regard, the brains of musicians and non-musicians react similarly, as both groups of participants have been tested in this study. From these results we conclude that the establishment of a metrical frame, i.e. the intuitive grasp of the underlying beat and the meter's accent structure is essential to integrate new tone items and for processing pitch and time relations (cf. Neuhaus and Knösche 2008) (Fig. 1).

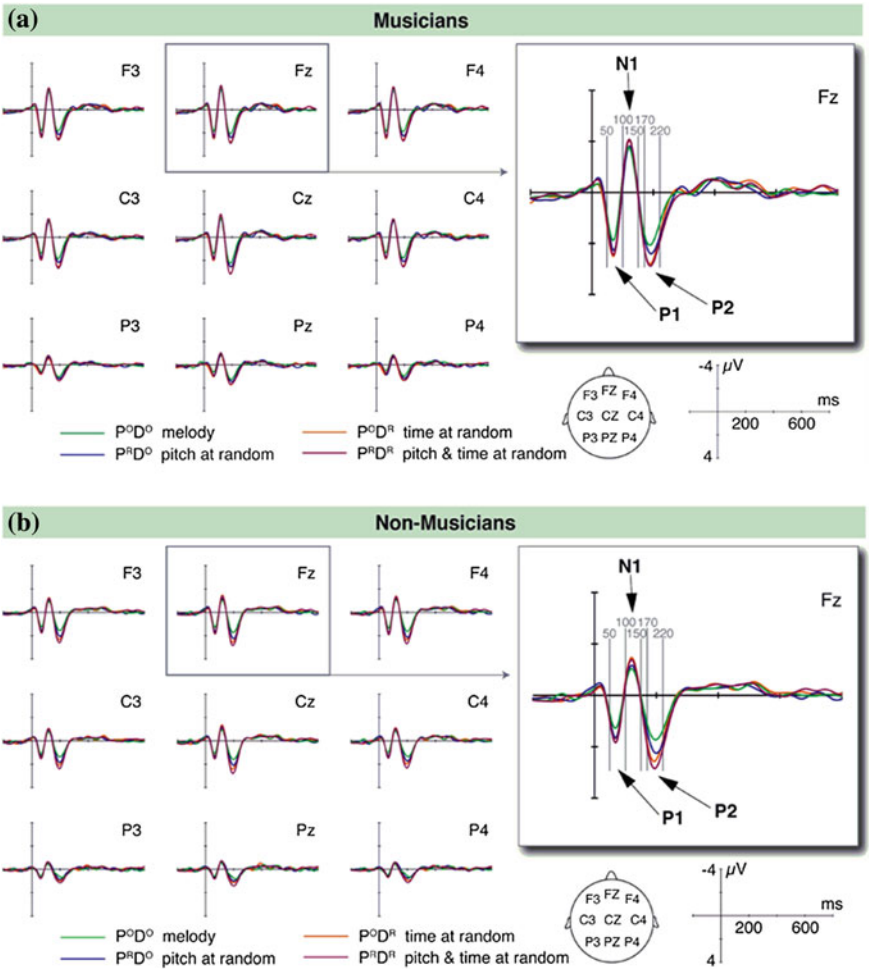


Fig. 1 Grand average event-related potentials of musicians (**a**) and non-musicians (**b**) at selected electrodes. Compressed brain activity over all tones per condition (from tone 6 onward). [*Source* Neuhaus and Knösche (2008)]

Taken non-randomness, i.e. the structured nature of a melody as given, how do we become aware of interval type, melodic contour, and the other properties mentioned above? Which listening mode is the most adequate to grasp these essential musical features? Everyone would agree with me to leave emotional, associative, aesthetic, and distracted styles of listening aside in favor of that one which aims at understanding the logic behind the tones and brings the composer's intention into focus.

The listening style that fits best is called 'structural hearing', a term coined by the Austro-American music theorist Felix Salzer in 1952. According to his younger colleague, the American musicologist Rose Rosengard Subotnik

structural hearing especially stands for modern viewpoints on music and became "the prevalent aesthetic paradigm in Germanic and Anglo-American musical scholarship" around 1950 (cf. dell' Antonio 2004, p. 2). To elaborate on the details, structural hearing "highlights [the] intellectual response to music to the almost total exclusion of human physical presence" (cf. dell' Antonio 2004, p. 8) and should therefore be seen in contrast to some sort of kinesthetic listening style, where the listener may feel the urge to dance, tap his feet, or snap his fingers (cf. Huron 2002; see also Kubik 1973, for the African variant of this motion-inducing listening mode). This 'intellectual response to music', as dell' Antonio puts it, might be restricted to music students, professional instrumentalists, and other people educated in music as it requires that several preconditions such as ear training and self-discipline as well as certain deductive musical abilities like melodic completion and anticipation are fulfilled. Thus, in full consequence, "[structural hearing] gives the listener the sense of composing the piece as it actualizes itself in time." (Subotnik 1988, p. 90). To what extent non-musicians are also capable of doing so, i.e. whether the aforementioned (listening) skills can also be acquired through implicit learning by mere exposure to music remains a question for further empirical research.

In this context we may call to mind that from a sociological point of view music philosopher Adorno (1962) as well as the musicologist and pedagogue Rauhe et al. (1975) made interesting general remarks on music consumption, taking the economic preconditions as well as the different educational standards within society into account. Both worked separately on a specific typology of listening behavior, elaborating carefully on—mostly a priori obtained—categories of reception (in German: 'Rezeptionskategorien').

With regard to structural hearing, Adorno (1962) distinguishes between two types of music listeners. The first is the 'music expert' who is almost exclusively recruited from music professionals, and the second one is the 'good [adequate] listener'. Please note that the 'good listener' should not be confused with another type of recipient, called 'educated listener' (or 'possessor of knowledge'; in German: 'Bildungskonsument'), being familiar with facts and reports about musical pieces and their interprets.

According to Adorno the 'music expert' is able to set past, present, and future moments within a musical piece in relation to each other to comprehend the musical logic behind the piece. Furthermore, while listening to music, the expert is aware of every structural aspect and can report on each musical detail afterwards.

The 'good [adequate] listener', by contrast, possesses these listening skills in a reduced form. Although he can relate musical parts to each other as the 'music expert' does, he is not fully aware of the consequences and implications that specific chords and other pivotal points with regard to musical progression have.

Rauhe et al. (1975) tackle this issue in a slightly different way. From the very start they distinguish between distinct attributes freed from human prototypes. That is, they argue on the basis of listening style rather than on the personnel level. Thus, the distinguishing feature between Rauhe's and Adorno's typologies is a certain flexibility in terms of switching between listening habits.

With regard to structural hearing, Rauhe et al. (1975) differentiate between a 'structural-analytical' and a 'structural-synthetical' listening mode (in German: 'strukturell-analytische Rezeption' and 'strukturell-synthetisches Hören'), meaning that the attentional focus is either on the structural detail, or that skills are quite similar to those possessed by Adorno's 'musical expert'.

In slight contrast to that, modern American musicologists such as Subotnik, dell' Antonio, or Huron put more emphasis on the real-time listening situation as such, i.e. on versatile listening styles beyond any (a priori drawn) boundaries, taking into account that actual listening behavior may change spontaneously throughout a performance. The American music psychologist Aiello (1994) puts it this way:

> I believe that when we listen, consciously or subconsciously, we *choose* what to focus on. This may be because of outstanding features of the music itself, certain features that are emphasized in the particular performance we are hearing, or because we just choose to focus on this or that element during that particular hearing. A piece of classical music is filled with more information than a listener can process in a single hearing. […] Given the complexity and the richness of the musical stimulus, listening implies choosing which elements to attend to. (p. 276).

More importantly, this principle of selection is confirmed with a series of experiments performed by the Belgian music psychologist Deliège et al. (1996). Deliège and their colleagues provide proof that musicians and non-musicians are able to extract several cues such as 'registral shifts' or 'change of density' from the textural surface of unfamiliar pieces, showing that cue abstraction is an effective means to grasp and encode the most obvious musical aspects in any real-time listening situation.

To extend this approach some sort of 'zoom in'/'zoom out' principle seems best to deal with real-time listening situations, giving the listener the perceptual freedom to either consider the whole or focus on the structural detail. This zooming principle corresponds nicely to Rauhe's structural-analytical and structural-synthetical listening styles. However, to put this zooming principle into practice, several preconditions must be fulfilled, each pointing to higher-order perceptual processes, including the involvement of an active mind. Such preconditions are, first, attentional mechanisms, i.e. some kind of conscious perception of the musical piece, second, strength of mind as well as, third, some form of intentionality, i.e. *directing* attention towards the whole pattern or a specific sound feature.

But which attribute is essential to experience a melody as an organic whole? How shall we use our zoom mechanism when listening to a musical piece in real-time? In my opinion, synthesis is more important than selection, i.e. the mental act of setting parts in relation to each other has priority over, what Deliège calls, 'cue abstraction'. Hugo Riemann, the well-known music theorist of the 19th century, already became aware of this principle in 1877, he called 'relational thinking' (in German: 'Beziehendes Denken'). Relational thinking means to create coherence between constituent elements, intervals, and form parts along the horizontal axis, lasting several seconds up to one minute, i.e. occurring within the temporal limits of working memory. However to describe these processes in modern psychology, the term 'chunking' is much more frequently used.

Theodor Lipps, a late German 19th-century philosopher and psychologist, puts emphasis on the *structural* side of relational thinking, i.e. on properties given in melodies and other musical pieces as such. By using some kind of mathematical formula, he called 'the law of number 2', Lipps managed to specify the relations between constituent tones and intervals as follows:

> Whenever tones meet each other that behave as $\mathbf{2^n : 3, 5, 7}$ etc., the latter naturally move towards the former; i.e. they have a natural tendency of inner motion to come to rest in the former. The latter "search for" the former as their natural basis, anchor points, or natural centers of gravity.

See also the original wording in German:

> Treffen Töne zusammen, die sich zueinander verhalten wie $\mathbf{2^n : 3, 5, 7}$ usw., so besteht eine natürliche Tendenz der letzteren zu den ersteren hin; es besteht eine Tendenz der inneren Bewegung, in den ersteren zur Ruhe zu kommen. Jene "suchen" diese als ihre natürliche Basis, als ihren natürlichen Schwerpunkt, als ihr natürliches Gravitationszentrum. (quoted from van Dyke Bingham, 1910, p. 11)

Please note that whenever we argue in favor of *process*, i.e. less in favor of structure, *future-directed* aspects of relational thinking become relevant. They are closely connected to concepts known as 'expectancy' and 'prediction' which either derive from a certain familiarity with the piece, from style-specific knowledge, or, in a statistical sense, from frequent occurrence of some textural features. In this regard, Pearce et al. (2010) distinguish between musical expectations generated on the basis of over learned rules and those built on the basis of associations and co-occurrence of events (with both processes most likely activating different parts of the brain).

From this it follows that processes of 'thinking ahead', be it prediction, anticipation, or 'setting-parts-in-relation-to-each-other', do not only demand the listener's perceptual awareness of the present with a particular focus on the input's acoustical features but also require some sort of re-activation of concepts about style, phrasing, harmonic progression, and other relevant structural issues stored in mind. Accordingly when 'thinking ahead', bottom-up and top-down processes do interact.

How can we 'operationalize' these processes? How can we grasp relational thinking by experiment? We first have to bring to mind that processes such as building expectancies and establishing coherence are two-fold, referring to musical items in horizontal as well as in vertical direction. Regarding the linear dimension, these processes can further relate to small segments or to the large-scale musical form. Let us quickly review the relevant research:

Narmour (1990, 1992), for instance, restricts his elaborations to the detail, i.e. to interval expectation formed within 3-tone segments. To explain how these expectancies for interval progression develop in the listener's mind he proposes a so-called implication-realization model, comprising five principal aspects based on the Gestaltists' perceptual laws, namely proximity, similarity, as well as good continuation. Among these five, the principle of *registral direction* means that in terms of small intervals pitch direction is expected to continue whereas for large intervals, i.e. of a perfect fifth or more, pitch direction is expected to change (see also Krumhansl 1997 for details).

Narmour's model of expectancy formation can simply be tested in a rating experiment, i.e. by asking participants to predict melodic continuation after having interrupted a folksong or some other musical piece at a specific point in time. However, when looking at this testing method a little more closely, Eerola et al. (2002) point out that, since listener's expectations may change in real-time listening situations, it is far better to obtain some continuous data on prediction. Thus, when testing expectation for melodic progression a dynamic approach seems more appropriate, using a coordinate system and a mouse-driven slider to enter the likelihood of forthcoming melodic events on a computer screen.

Krumhansl's probe tone method, in contrast (developed in collaboration with Roger Shepard in 1979) is a valuable means to investigate tone relations in vertical direction. Again, Krumhansl makes use of the rating method, this time by judging the cohesive strength between melodic events and the latent harmonic framework. For this objective, participants have to evaluate on a rating scale how well the twelve probe tones of the chromatic scale fit into the context of the just heard piece, be it tonal, atonal, bi-tonal, or non-Western (e.g. Kessler et al. 1984).

Cook (1987) as well as Eitan and Granot (2008) take a different approach to relational thinking. They study cohesiveness within *large-scale musical forms* along the horizontal axis. The objective of their studies is to find out whether and to what extent participants become aware of key change and of permuted musical sections while listening to sonata movements and other types of compositional form. A second focus is on the aesthetical impressions created by hybrid and original pieces. Again, insights are gained from rating results. By taking two of Mozart's masterworks (i.e. his piano sonata KV 332 and the earlier KV 280) as the original, and comparing it with a mixed version including some equivalent parts of the respective counterpart, Eitan and Granot (2008) could demonstrate that preferences for hybrid versions in musically trained listeners are strong, and become even stronger when exposed to these pieces several times. Since these rating results do not confirm that priority is given to the original, Eitan and Granot (2008) call the musical logic and the piece's inner unity into question. However, any sweeping generalizations based on these findings should be avoided since judgment results might be restricted to a specific idiolect and style, i.e. to Mozart's way of composing and to the Classical period per se.

In addition, Cook (1987) investigated the effect of tonality on cohesion. In this study, music students rated the degree of completeness of musical pieces of different lengths up to 6 min, ending either on the tonic or a distant key. Since modulations were merely perceptible within a time span of one minute or less, Cook concluded that form-building effects of tonality are weak which advocates for the psychological reality of small-scale over large-scale musical structures.

An even more radical view, called 'concatenationism', is hold by the American philosopher Jerrold Levinson (1997; cited by v. Hippel 2000). Concatenationism in its strictest sense describes a "moment-by-moment listening, faintly tinted by memory and expectation" (p. 135). According to Levinson, 'concatenationism'

denies any conscious influence of large-scale musical form on listening at all, meaning that we simply hear musical sections in succession and remain in the musical present.

To put this issue to the test, I performed a neurocognitive experiment on relational thinking at the Max Planck Institute for Human Cognitive and Brain Sciences Leipzig which I would like to re-report in this context here. To my knowledge it is the first neuroscience approach to musical form perception at all. Brain activity was recorded by using event-related potentials (ERPs) as the measuring method. To control the listening result immediately after presenting a melody, short behavioral feedback was given and registered via button press response. In general, event-related potentials (as well as other neuroimaging methods) have the advantage to map physiological processes in real time, enabling us to watch brain reactions—either (synchronized) extracellular current flows in terms of EEG and ERP or oxygen consumption regarding fMRI—while bundles of nerve cells are active.

The objective of the present study was to find out if the brain responds to the mental act of relational thinking and, if so, whether chunking processes can be made visible by specific component reactions. From the structural point of view I thus tested 'balance between form parts', which is a property considered essential when thinking about the melody's particular characteristics (c.f. p. 1f).

For the clarity of the experiment I decided for the small-scale type of musical form, using the eight-bar 'musical period', also called 'Liedform', in two variants, AABB and ABAB (see Fig. 2).

The figure shows that form parts with equal labelling (**AA**, **ABAB**, A**BAB**) have exactly the same rhythm while intervals are only similar, whereas form parts with different labelling (**AB**) are dissimilar in both, rhythm and interval structure. In terms of sequence succession we assume that relational thinking is affected by the sequential order of form parts, meaning that chunking may either be facilitated or made more difficult when A-parts alternate with B-parts as in ABAB, or when A- and B-parts follow in immediate repetition as in AABB. From these considerations the following hypotheses are deduced:

1. H1: Adjacent form parts of *contrasting* structure (AB) are subsumed to higher-level perceptual units. Melodies of form type ABAB are rated as hierarchical.

   H0: The sequential order of form parts has no effecton building higher-level perceptual units at all.

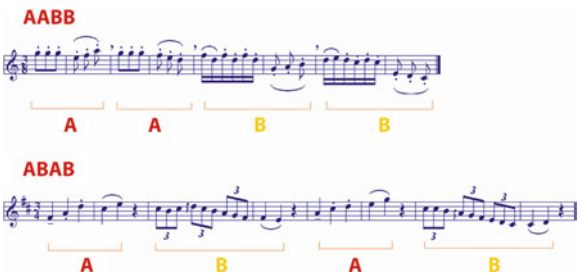


Fig. 2 Example melodies in *AABB* and *ABAB* forms (modified versions of originals). *First example* L. v. Beethoven. Rondo, WoO 48. *Second example* C. Ph. E. Bach. Thema from Sonata III

2. H1: Structuring small-scale musical forms is an online cognitive process taking place in working memory.
   Specific ERP components serve as indicators.

   H0: Structuring form parts happens in retrospect. Post hoc ratings are necessary to evaluate the entire melody as either 'hierarchical' or 'sequential'.

From this it follows that melodies of form type ABAB probably elicit a large arc of suspense, increasing coherence, whereas in melodies of form type AABB the arc of suspense is smaller, and cohesive strength is reduced.

Figure 3 illustrates the task. Participants had to listen to and evaluate the patterns by choosing either a sequential or a hierarchical listening style. Rating results were indicated by press of key buttons. The study was exclusively performed with non-musicians.

1 Methods

1.1 Subjects

Twenty students of different faculties (recruited from the University of Leipzig) participated in the experiment. Age and gender were equally balanced (10 males, 10 females; average age = 25.9 years, *SD* = 2.29).

2 Stimuli and Task

Melodies of form types AABB and ABAB were built on the eight-bar schema of the Classical period split up into 2 + 2 + 2 + 2- phrase-units (see note examples in Fig. 2). Each form type was realized with 25 different melodies randomly presented in three keys (C major, E major, A flat major), resulting in 75 different

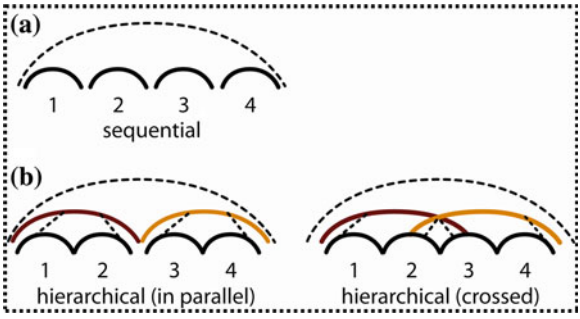


Fig. 3 Illustration of the listening task. Possible segmentations of a melody in real-time. For each given example participants had to decide by individual judgment which listening strategy was best. **a** Patterns are closed sections following in a row. **b** Patterns are related to each other and form higher-order units

versions of AABB and ABAB, respectively. In terms of melodic contour, an arch-like shaping of all four phrases might intensify the impression of closure, and melodies of that shape might be processed sequentially whereas an upward contour followed by a downward one, each extending over 2 plus 2 bars, strengthens coherence, and listening might be more hierarchical. We therefore took heed that example pieces with arch-like and upward/downward contour were equal in number so that modulating effects caused by contour could be excluded. In addition, pause length, length of the pre-boundary tone, and the underlying harmonic schema were kept constant in each melody to avoid interferences with phrase boundary perception, probably yielding the CPS. In each sequence, the chord progression was "tonic (1st phrase)—dominant (2nd and 3rd phrase)—tonic (4th phrase)". Phrase boundaries had an average pause length of 0.27 s ($SD = 0.12$), and the average duration for each pre-boundary tone was 0.42 s ($SD = 0.25$).

The example pieces were played on a programmable keyboard (Yamaha PSR 1000) in the timbre 'grand piano'. They were stored in MIDI format using the music software Steinberg™ Cubasis VST 4.0. For wave-file presentation via soundcard, MIDI files were transformed to Soundblaster™ audio format.

Each musical piece was played with an average tempo of 102.36 BPM ($SD = 27.92$), differing slightly between AABB (108.76 BPM, $SD = 32.23$) and ABAB (95.96 BPM, $SD = 21.63$). However, an independent samples t-test yielded a non-significant result ($t(48) = 1.65$, $p > 0.1$), giving certainty that tempo should not be considered as a disturbing factor.

Figure 3 shows the given options how to deal with each melody. The drawings were also presented during instruction to illustrate the following task: "Listen carefully and evaluate the parts: Whenever you perceive sections as closed, following in a row—press the left button for 'sequential'. Whenever you perceive sections as related to each other, forming higher-order units—press the right button for 'hierarchical'." Participants were also requested to keep head, neck, arms, hands, and fingers as relaxed as possible and to reduce the amount of eye blinks during ERP recording (cf. general guidelines of ERP measurement, Picton et al. 2001). To get familiar with the task, each recording session started with a test block of ten melodies.

3 Experimental Set-up and Recording

The total duration of each recording session was approximately 55 min. Three blocks with stimuli were presented, each consisting of 50 melodies of types AABB and ABAB in pseudo-random order. Each example piece was part of a trial sequence. It ran as follows: "Fix your eyes onto the monitor (2 s), listen to the added type of melody (approximately 11 s), then evaluate whether the melody is sequential or hierarchical (less than 6 s)." Trial sequences were presented automatically using the software package ERTS (Experimental Run Time System,

Version 3.11, BeriSoft 1995). Each subject was comfortably seated in a dimmed and electrically shielded EEG cabin in front of a monitor. For binaural presentation of stimuli, a loudspeaker was placed at a distance of 1 m.

Brain activity was measured with 59 active Ag/AgCl electrodes (Electro Cap International Inc., Eaton, Ohio) placed according to the 10-10 system onto the head's surface (Oostenveld and Praamstra 2001). Brain activity per electrode was referenced to the left preauricular point (A1), and the sternum was used as the ground electrode. EEG signals were recorded with an infinite time constant and digitised with a sampling rate of 500 Hz. Ocular artefacts were measured with a vertical and a horizontal electrooculogram (EOGV, EOGH). The impedance at each electrode channel was kept below 5 kΩ.

4 Data Analysis

4.1 Pre-processing of Signals

The obtained raw signals were high-pass filtered (cut off frequency 0.50 Hz) and carefully examined for eye blinks, muscle activity, and technical artefacts. Artefact-free trials were merged over melodies and blocks, but averaged separately according to form type (AABB vs. ABAB), phrase onset (2nd, 3rd, and 4th phrase), and electrode channel. The time window for averaging was −200–1,000 ms measured from onset of the respective phrase. ERP traces were baseline-corrected, using a pre-onset interval from −30 to 0 ms. The pre-processing procedures were performed for each subject individually. Figure 5a–c shows the grand average ERP over all subjects at nine representative electrodes (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4). The time range for display is the same as for averaging, namely −200–1,000 ms pertaining to phrase onset.

5 Statistical Analysis

Table 1 shows the button press responses of participants rating melodies as hierarchical or sequential. An additional Chi-square test was performed to prove whether rating results correlated significantly with pattern structure. In contrast to that, Table 2 summarizes the attributes of several 'prototype melodies' in AABB and ABAB that 75 % of participants (n = 15) consistently rated as hierarchical or sequential.

To test the significance of participants' *ERP data*, we computed several ANOVAs (repeated-measures analyses of variance). We started with a four-way analysis so that, if significant, effects of factor and level could be analyzed separately. Repeated measures factors were Time Window (five time ranges: 30–80,

Table 1 Participants' rating scores specified according to Form type

Form type	Rating results	
	"hierarchical"	"sequential"
AABB	1009 (67.45 %)	487 (32.55 %)
ABAB	1109 (74.28 %)	384 (25.72 %)
Across form types	2118 (70.86)	871 (29.14)

Sum of raw scores and percentages

80–140, 140–280, 300–600, 600–800 ms), Onset (2nd, 3rd, 4th phrase), Form (AABB and ABAB), and Channel (36 electrodes). Due to lack of space, results reported here are restricted to only one time window (300–600 ms). Further details were provided with a three-factor ANOVA analysis per time window including the repeated measures factors Onset, Form, and Channel. Dependencies between factors and factor levels were specified with several one-way and two-way post hoc tests. Degrees of freedom were adjusted with Huynh and Feldt's epsilon, and results were considered significant at an α-level of 0.05.

6 Results

6.1 Behavioral Results

Table 1 shows the overall rating scores of participants. The distribution assigns the total number of listening judgments ('hierarchical' vs. 'sequential') to form type ('AABB vs. ABAB'). Altogether, approximately two-thirds of example melodies were judged as 'hierarchical', the remaining one-third was judged as 'sequential'. Further specification by 'Form type' revealed that example pieces in ABAB compared to AABB were slightly more often rated as 'hierarchical' (74.28 vs. 67.25 %). This difference reached statistical significance ($t(38) = 2.13$, $p < 0.05$).

Table 2 Prototype melodies, consistently rated as hierarchical or sequential (14 themes in AABB and 19 themes in ABAB)

Frequency distribution	AABB		ABAB	
	Hierarchical	Sequential	Hierarchical	Sequential
Total number of analyzed melodies	11 [44 %]	3 [12 %]	16 [64 %]	3 [12 %]
Rhythmic contrast between part A and part B	11 [44 %]	1 [4 %]	15 [60 %]	1 [4 %]
Contour				
Archlike	1 [4 %]	2 [8 %]	3 [12 %]	2 [8 %]
Upward–downward	10 [40 %]	1 [4 %]	12 [48 %]	1 [4 %]
Downward–upward			1 [4 %]	

Frequency distributions (raw scores and percentages)

We computed a Chi-square test to evaluate if participants' *overall rating results* differed significantly from chance level. The X^2-value of 16.9 is far beyond the critical value of 10.83 at the 0.001 α-level (X^2 (1, N = 2989) = 16.9, $p < 0.001$). To quantify correlation strength between 'Form type' and 'Listening style', we additionally computed the adjusted contingency coefficient C^*, yielding scores on a range between 0 and 1. The result ($C^* = 0.86$) shows that the effect size between 'Form type' and 'Listening style' is large, indicating that correlation between both factors is strong.

Table 2 gives information about a subset of example melodies that three-fourths of participants (n = 15) consistently rated as hierarchical or sequential. Two variables are introduced which explain modification *across form types*: 'rhythmical contrast' and 'melodic contour'. The data show that both, sharp rhythmical contrast between A and B and/or upward-downward (instead of arch-like) contours contribute to a strong impression of higher-level junctions—either of 'AB' with 'AB' or of 'AA' with 'BB'.

7 Electrophysiological Results

We expect that chunking tendencies were strongest at the immediate onset of A- and B-parts. For that reason, three trigger points were set per example piece, one at each phrase onset. Note that at each phrase onset the average tone length was 0.24 s ($SD = 0.17$) while mode values, i.e. tone lengths occurring most frequently, were 0.07, 0.08, and 0.18 s over examples. We therefore proceed on the assumption that each phrase onset displays the average brain response to at least two or three tones rather than to the onset tone alone.

The first idea was that the brain reflects the interrelation between 'Form type' (AABB vs. ABAB) and 'Listening style' (hierarchical vs. sequential). Figure 4 shows the grand average ERP at the onset of the second phrase which is the crucial phrase of the eight-measure theme following the initial phrase.

From 300 ms onward, we observe a splitting of traces according to 'Form type' (AABB vs. ABAB) but no further division for subjective listening style

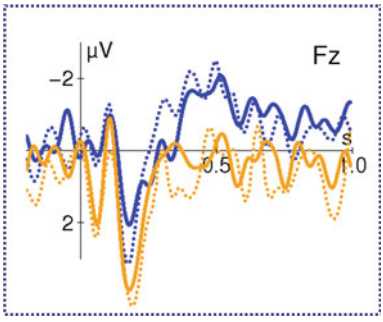


Fig. 4 Grand average ERPs at phrase onset 2. Division of traces by form type and listening style. Cortical activity at electrode *Fz*

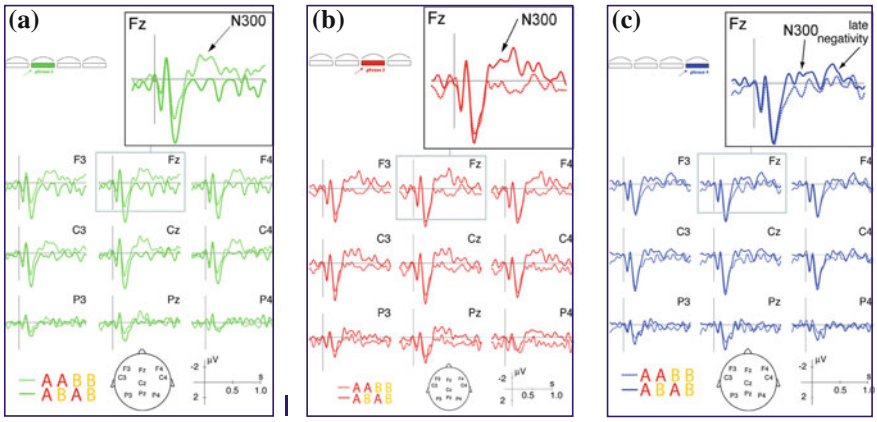


Fig. 5 **a–c** Grand average ERPs at phrase onsets 2, 3, and 4. Curve splittings according to 'Form type' (*AABB* vs. *ABAB*)

(hierarchical vs. sequential). This observation is validated with statistics. The ANOVA analysis yields a main effect of Form type ($F(1,19) = 7.53$, $p < 0.01$, 36 channels) whereas no interaction between 'Form type' and 'Listening style' could be found. The ANOVA results for curve splitting therefore suggest that the brain is merely sensitive to pattern structure (but not to the respective listening style).

Figure 5a–c displays the event-related potentials *as a mere function of form type*. The most interesting result is the broad negative shift between 300 and 600 ms measured from phrase onset. Since the amplitude maximum is fronto-central, we call this negative shift 'anterior negativity' (anterior N300). As you can see from Fig. 5a–c polarity changes between the onsets of the second and the third (and fourth) phrase. At phrase onset 2, the anterior negativity is up for patterns **AA** (immediate repetition of A) compared to AB, while at phrase onset 3 it is up for **ABA** (non-immediate repetition of A) compared to AAB. At phrase onset 4 we observe an anterior N300 for A**B**A**B** compared to AABB, although amplitude is reduced. Due to this pattern reversal, a main effect of 'Form type' is only marginally significant, whereas the interaction 'Form type x Onset' is highly significant. The anterior N300 also decreases from anterior to posterior. This is validated by a highly significant interaction 'Onset x Form type x Channel', mostly pronounced at phrase onset 2. (Due to lack of space ANOVA results are not displayed here).

8 Discussion

What are the study's main results in particular with regard to relational thinking?

Let us briefly call them to mind:

1. [*behavioral*]. Sequences with form parts in alternating order (ABAB) are more often judged as 'hierarchical' than sequences with form parts in immediate succession (AABB) (cf. Table 1).
2. [*behavioral*]. Judgment results are more pronounced when rhythmical contrast between A- and B-parts is sharp and when melodic contours are upward-downward (cf. Table 2).
3. [*electrophysiological*]. The brain seems to react to pattern similarity at the onsets of adjacent (**A**|**A**) and of non-adjacent (e.g. **AB**|**A**) form parts. It does *not* respond to contrasting pattern structure (**A**-**B**).
4. [*behavioral and electrophysiological*]. Decisions on how form parts are perceived are not reflected in the ERP as a 'real-time protocol of brain activity'. Instead, the impression of the entire melody seems necessary.

The *behavioral results* suggest that participants are able to distinguish between example pieces of strong and weak cohesion. As already mentioned several sentences before, cohesiveness is strong when the following textural properties are given: (1) Form patterns alternate with each other as in ABAB, (2) melodic contours are upward-downward, and (3) rhythmic contrast between A- and B-parts is sharp. Among these points causing the impression of hierarchy, 'sequential order' obviously is the essential one while features regarding rhythm and contour are merely supplementary. Thus, the choice of the adequate listening style and the ease of evaluation seem largely to depend on the extent to which these properties are developed.

Note that due to the given task the primary focus was on listening strategy and on stimulus evaluation, i.e. on the *pre-requisites* for relational thinking. That is, chunking, defined as a cognitive process of 'subsuming form parts to higher-order units' probably requires a still larger amount of mental activity as needed by the current task.

Please note further that this experiment was exclusively performed with non-musicians, meaning that the process of 'setting-form-parts-in-relation-to-each-other' obviously belongs to a number of implicitly acquired musical skills. In this regard, Bigand and his colleagues were able to show that untrained listeners process tension and relaxation as well as several structural aspects above chance level and often with same results as professional musicians (Bigand 2003; Bigand and Poulin-Charronnat 2006). In this context Tillmann and Bigand (2004) write the following:

> Nonmusician listeners tacitly understand the context dependency of events' musical functions and, more generally, the complex relations between tones, chords, and keys. (p. 212).

Now let us try to explain the electrophysiological data. First, form parts with equal labelling (**AA**, **AB**A**B**, A**B**A**B**) are of exactly the same rhythm while intervals are only similar whereas form parts with different labelling (**AB**) differ in both, interval size and tone duration. We therefore suggest that the brain detects pattern similarity *on the basis of identical rhythmic structure*. In this context, the

anterior N300 may serve as a neural correlate, indicating the amount of mental effort which is necessary to identify the structural properties at phrase onsets. For these cognitive processes working memory resources are needed that are mainly found in the brain's frontal part. Whenever working memory is active, we observe amplitude maxima at anterior parts of the brain (cf. Patel 2003; and Fig. 5a–c).

Leaving considerations about rhythmical sameness aside, we can also explain the data from the viewpoint of expecting and predicting musical structure. In a recent study on pitch expectations in church hymns Pearce and colleagues (2010) managed to demonstrate that high-probability tones (perceived as expected) compared to low-probability tones (perceived as unexpected) elicit a fronto-central negativity in the time range from 300 to 600 ms similar to the anterior N300 of the present study but slightly different in shape. However, on closer examination, we have to keep in mind that the anticipated events of both studies are of different character and size, using highly probable single tones on the one hand and highly probable form parts (stretching over 2 bars) on the other. Even so, the similarity between the anterior brain components of both studies is striking.

With regard to 'relational thinking'—what conclusions should we draw from this study?

In my opinion the key aspect is that this 'setting-parts-in-relation-to-each-other' is not part of a superior process named 'structural listening' occurring in the same time range, but rather a separate process taking place several seconds later. This means that, initially, the anterior N300 indicates the processing of structural features whenever the brain becomes aware of rhythmical sameness in real-time listening situations, whereas relational thinking, or, more exactly, the choice of the adequate listening style combined with subsequent rating decisions, needs a complete perceptual unit to evaluate the coherence of form parts. Thus, so far, no specific neural correlate for pure active, higher-order mental acts such as chunking or Gestalt perception has been found, showing that relational thinking seems measurable only in retrospect by using traditional behavioral methods. The present ERP study therefore makes clear that a new paradigm seems more appropriate in order to investigate the idealized form of 'expert structural listening' referring to real-time listening skills such as reconstructing and anticipating the score.

Having elaborated on Riemann's 'relational thinking' and 'balance between form parts' from a structural point of view, we should close this chapter with some general remarks on structural listening and on grasping the musical idea, this time by also including the composer's point of view. The first point we should think about is that spatio-temporal concepts depend on the type of cognitive process, i.e. there are conceptual differences between imagining and creating on the one hand and simply listening to music on the other.

Let us first consider the composer's point of view: When describing the creative act, they often report on aspects of simultaneity and spatiality, i.e. on the overall-structure of a musical piece. Mozart, for instance, obviously sees the complete musical architecture in his mind's eye (c.f. a letter written by J. F. Rochlitz 1813) whereas, Schönberg uses the term "timeless entity" to describe this overall framework (quoted from Cook 1990, p. 226). Thus, a main problem when

composing music is to transfer spatial concepts into time-based sequences, i.e. to unfold basic ideas successively step by step. Listening, by contrast, proceeds in the opposite direction in that the overall idea has to be grasped (and compressed) from the sequential 'spread out', i.e. from tones presented in succession. Thus, in terms of transferring sound information, processes of composing/imagining and listening are diametrically opposite (see also Mersmann 1952).

In sharp contrast to that, Riemann completely omits this spatio-temporal aspect. He skips also every facet regarding structure and the sound material as such. Instead, he puts emphasis on the immediate mental exchange between the composer's and the listener's minds for rapid transfer of thought and direct flow of ideas. See the following quotation (Riemann 1914/15):

> The Alpha-Omega in music is not the sound, i.e. real music, or the audible tones as such, but rather the idea of tone relation. Before putting them into notation, these relations live in the imagination of the creative artist and then in the listener's imagination ….In other words, the essential approach to music, i.e. the key to its inner nature is neither provided by acoustics nor by tone physiology and tone psychology but rather by imagination of tone. (p. 15f)

See here the original wording in German:

> daß nämlich gar nicht die wirklich erklingende Musik, sondern vielmehr die in der Tonphantasie des schaffenden Künstlers vor der Aufzeichnung in Noten lebende und wieder in der Tonphantasie des Hörers neu erstehende Vorstellung der Tonverhältnisse das Alpha und Omega der Tonkunst ist. ….. Mit anderen Worten: den Schlüssel zum innersten Wesen der Musik kann nicht die Akustik, auch nicht die Tonphysiologie und Tonpsychologie, sondern nur eine "Lehre von den Tonvorstellungen" geben. (Riemann 1914/15, p. 15f)

However, several constraints are imposed on Riemann's idea of communicating from mind-to-mind, mainly caused by attitudes, experiences, expectations, wishes, and situational needs on the side of the listener (cf. Rauhe et al. 1975). In other words, encoding and decoding or the sending and receiving of the overall musical idea are largely dissimilar processes, even more when they are split between two people, since they are modified by some mental acts and attributes in-between. In this regard, Jackendoff distinguishes between a 'compositional grammar' and a 'listening grammar', although admitting in the spirit of Riemann that "the best music arises from an alliance of a compositional grammar with the listening grammar" (1988, p. 255; quoted from Cook 1994, p. 87). Even so, it makes more sense to assume that one person is real and the other fictional, meaning that a composer constructs his 'ideal listener' to enhance creativity and shape the (creative) work like poets or journalists do when having their 'ideal reader' in mind (cf. dell' Antonio 2004).

Altogether we may conclude that the perception of musical structure as well as related issues such as structural listening and relational thinking need an interdisciplinary approach, i.e. the move from music theory to music psychology to Gestalt psychology and back, focusing on both, the sound object, i.e. the tone material, and the respective perceptual process.

Let me close with a quotation from Miller and Gazzaniga (1984) on the interrelationship between structure and process in its broadest sense:

> There seems to be general agreement that the objects of study are cognitive *structures* and cognitive *processes*, although this distinction is drawn somewhat differently in different fields. In American psychology, for example, a long tradition of functional psychology has made it easier to think in terms of processes—perceiving, attending, learning, thinking, speaking—than in terms of structures. Yet, something has to be processed. … Linguists, on the other hand, generally find it easier to think in terms of structures—morphological structures, sentence structures, lexical structures—and to leave implicit the cognitive processes whereby such structures are created or transformed. Computer scientists seem to have been most successful in awarding equal dignity to representational structures and transformational processes. … In principle, however, it is agreed that both aspects must be considered together, but that once the process is understood the structure is easier to describe. Thus, cognition is seen as having an active and a passive component: an active component that processes and a passive component that is processed. (p. 8f)

References

Abraham, O., & v. Hornbostel, E. M. (1909). Vorschläge für die Transkription exotischer Melodien. In C. Kaden & E. Stockmann (Eds., 1986), Tonart und Ethos - Aufsätze zur Musikethnologie und Musikpsychologie (pp. 112–150). Leipzig: Reclam.

Adams, C. R. (1976). Melodic contour typology. *Ethnomusicology, 20*(2), 179–215.

Adorno, T. W. (1962). Typen musikalischen Verhaltens. In Einleitung in die Musiksoziologie—zwölf theoretische Vorlesungen (pp. 13–31). Frankfurt a.M.: Suhrkamp.

Aiello, R. (1994). Can listening to music be experimentally studied? In R. Aiello & J. Sloboda (Eds.), *Musical perceptions* (pp. 273–282). Oxford: Oxford University Press.

Bigand, E. (2003). More about the musical expertise of musically untrained listeners. In G. Avanzini et al. (Eds.), The neurosciences and music II: From perception to performance. Annals of the New York Academy of Sciences 999 (pp. 304—312). New York: The New York Academy of Sciences.

Bigand, E., & Poulin-Charronat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*, 100–130.

Cook, N. (1987). The perception of large-scale tonal closure. *Music Perception, 5*(2), 197–206.

Cook, N. (1990). *Music, imagination, and culture*. Oxford: Oxford University Press.

Cook, N. (1994). Perception: A perspective from music theory. In R. Aiello & J. Sloboda (Eds.), *Musical perceptions* (pp. 64–95). Oxford: Oxford University Press.

Danielou, A. (1982). Einführung in die indische Musik (2nd ed.; Series: Taschenbücher zur Musikwissenschaft 36). Wilhelmshaven: Heinrichshofen.

Deliège, I., et al. (1996). Musical schemata in real-time listening to a piece of music. *Music perception, 14*(2), 117–160.

Dell'Antonio, A. (2004). Beyond structural listening: Postmodern modes of hearing. Berkeley: University of California Press (Ed. 2002).

Eerola, T., et al. (2002). Real-time predictions of melodies: Continuous predictability judgments and dynamic models. In Stevens, C., et al. (Eds.) Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney (pp. 473–476).

Eitan, Z., & Granot, R. Y. (2008). Growing oranges on Mozart's apple tree: "Inner form" and aesthetic judgment. *Music Perception, 25*(5), 397–417.

Huron, D. (2002). Listening styles and listening strategies. Talk presentation and handout. Society for Music Theory Conference, Columbus, Ohio, 1st Nov 2002.

Karbusicky, V. (1979). Systematische Musikwissenschaft. (Series: Uni-Taschenbücher 911). München: Wilhelm Fink.

Kessler, E. J., et al. (1984). Tonal schemata in the perception of music in Bali and the West. *Music Perception, 2*, 131–165.

Knösche, T. R., et al. (2005). Perception of phrase structure in music. *Human Brain Mapping, 24*(4), 259–273.

Krüger, F. (1926). *Komplexqualitäten, Gestalten und Gefühle*. München: C.H. Beck.

Krumhansl, C. (1997). Effects of perceptual organization and musical form on melodic expectancies. In M. Leman (Ed.), *Music, gestalt, and computing: Studies in cognitive and systematic musicology* (pp. 294–321)., Lectures notes in computer science1317 Berlin: Springer.

Kubik, G. (1973). Verstehen in afrikanischen Musikkulturen. In P. Faltin & H.-P. Reinecke (Eds.), *Musik und Verstehen: Aufsätze zur semiotischen Theorie, Ästhetik und Soziologie der musikalischen Rezeption* (pp. 171–188). Köln: Hans Gerig.

Levitin, D. J. (2009). The neural correlates of temporal structure in music. *Music and Medicine, 1*(1), 9–13.

Levitin, D. J., & Menon, V. (2005). The neural locus of temporal structure and expectancies in music: Evidence from functional imaging at 3 Tesla. *Music Perception, 22*(3), 563–575.

Mersmann, H. (1952). Musikhören. Frankfurt a. M.: Hans F. Menck.

Miller, G. A., & Gazzaniga, M. S. (1984). The cognitive sciences. In M. S. Gazzaniga (Ed.), *Handbook of cognitive neurosciences* (pp. 3–11). New York: Springer.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model*. Chicago: University of Chicago Press.

Narmour, E. (1992). *The analysis and cognition of melodic complexity: The implication–realization model*. Chicago: University of Chicago Press.

Neuhaus, C., et al. (2006). Effects of musical expertise and boundary markers on phrase perception in music. *Journal of Cognitive Neuroscience, 18*(3), 472–493.

Neuhaus, C., & Knösche, T. R. (2008). Processing of pitch and time sequences in music. *Neuroscience Letters, 441*(1), 11–15.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology, 112*, 713–719.

Patel, A. D. (2003). Language, music, syntax, and the brain. *Nature Neuroscience, 6*, 674–681.

Pearce, M. T., et al. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage, 50*(1), 302–313.

Picton, T. W., et al. (2001). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology, 37*, 127–152.

Rauhe, H., et al. (1975). *Hören und Verstehen: Theorie und Praxis handlungsorientierten Musikunterrichts*. München: Kösel.

Riemann, H. (1877). *Musikalische Syntaxis: Grundriß einer harmonischen Satzbildungslehre*. Leipzig: Breitkopf and Härtel.

Riemann, H. (1914/15). Ideen zu einer "Lehre von den Tonvorstellungen". In B. Dopheide (Ed., 1975), Musikhören (pp. 14–47). Darmstadt: Wissenschaftliche Buchgesellschaft.

Subotnik, R. R. (1988). Toward a deconstruction of structural listening: A critique of Schönberg, Adorno, and Stravinsky. In L. B. Meyer et al. (Eds.) Explorations in music, the arts, and ideas: Essays in honor of Leonard B. Meyer (pp. 87–122). Stuyvesant, New York: Pendragon Press.

Tillmann, B., & Bigand, E. (2004). The relative importance of local and global structures in music perception. *Journal of Aesthetics and Art Criticism, 62*(2), 211–222.

Van Dyke Bingham, W. (1910). Studies in melody. *Psychological Review, 50*, 1–88.

Von Ehrenfels, C. (1890). Ueber 'Gestaltqualitäten'. *Vierteljahrsschrift für wissenschaftliche Philosophie, 14*, 249–292.

Von Hippel, P. (2000). Review of the book Music in the moment by Jerrold Levinson [1997. Ithaca, NY: Cornell University Press]. Music Analysis 19 (1), 134–141.

West, R., et al. (1991). Musical structure and knowledge representation. In P. Howell et al. (Eds.) Representing musical structure. (Cognitive science series 5, pp. 1–30). London: Academic Press.

White, B. (1960). Recognition of distorted melodies. *American Journal of Psychology, 73*, 100–107.

Ziegenrücker, W. (1979). *Allgemeine Musiklehre*. Mainz, München: Goldmann Schott.

Part III
Applications

Virtual Room Acoustics

Michael Vorländer, Sönke Pelzer and Frank Wefers

1 Introduction

With the rapid development of computers, room acoustical simulation software was created and applied in sound field analysis in rooms. The processor speed, the memory space, and the convolution machines were sufficiently powerful finally in the beginning of the 1990s to allow room acoustical computer simulation on a standard personal computer. Since then, several improvements in the modelling algorithms, in binaural processing and in reproduction techniques were made.

Room acoustic auralization has been developed from simulation algorithms and binaural technology in a historic process of more than 20 years. Full-immersive virtual reality (VR) systems, such as CAVE-like environments, have been in use for more than 15 years. While computer graphics and video rendering are far developed with applications in film industry and computer games, high-quality audio rendering is still not on a comparable level. Watching a recent 3D movie production and comparing the quality of the visual and auditory representation can best illustrate this mismatch. While advanced projection systems deliver a good 3D vision, the auditory 3D impression lacks a realistic spatial sound impression, although rather complex surround sound systems are usually installed.

The concept of auralization was also applied to fields other than room acoustics since about the year 2000. The aim is now different, as not music and the quality of concert halls or other performance spaces are to be evaluated, but the perception of sound and noise. Thus, building acoustics, automotive acoustics, and machinery noise are areas of application. The task in all these applications is the evaluation of sound sources, transmission constructions, or products by listening instead of a numeric expression of the acoustic quality. Wave-based numerical acoustics such as the finite element method (FEM), the boundary element method (BEM), the finite time-domain difference technique (FDTD), or analytic models and any kind

M. Vorländer (✉) · S. Pelzer · F. Wefers
Institute of Technical Acoustics, RWTH Aachen University, Aachen, Germany
e-mail: mvo@akustik.rwth-aachen.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_9,

of structural acoustics transfer path method is suited as a basis for auralization. The link between simulation and auralization is the representation of the problem in the signal domain and the treatment of sound and vibration by signal processing.

Following the concepts of simulations in acoustics and vibration, we can describe the process of auralization by separation of the processes of sound generation and transmission into system blocks and description of these blocks with tools of system theory. Figure 1 illustrates the basic elements of sound generation, transmission and radiation [see also Vorländer (2008)].

One might ask why the problem cannot simply be treated by using a mono signal, an equalizer and a headphone. The need for a more complex reproduction technique with a spatial representation is given by the fact that human hearing extracts information about the sound event and the sound environment by segregation of acoustic objects due to common cues of spectral, temporal, and spatial attributes. This, for instance allows us to identify one speaker out of a cloud of diffuse speech (cocktail party effect). In situations of outdoor noise immission, the spectral, temporal, and spatial cues are extracted to judge the event as pleasant, annoying, informative, or neutral. As long as the specific acoustic (physical) and semantic content of the noise must be treated, restrictions in spectral, temporal, and spatial cues are not appropriate.

In room acoustics, the quality of the results must be very high. People are sensitive to the perception of music in all its aspects, temporal, spectral, and spatial. Therefore, the challenge in creating auralization in room acoustics is very high, and this applies to source recording, sound propagation (reverberation) rendering, and audio reproduction. While source recording and audio reproduction is discussed in other papers, this contribution is focused on simulation of the sound propagation in rooms.

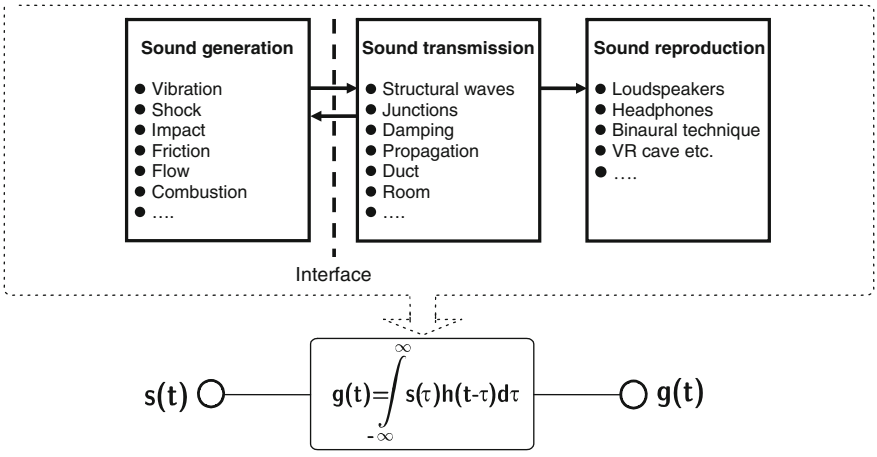


Fig. 1 Components of auralization and virtual acoustics: representation of sound and vibration sources, transmission, reproduction, and mapping on a task of signal processing

2 Room Acoustic Simulation

Today, simulated room acoustics are applied in various fields with great success. Their well-developed algorithms help to create realistic acoustics during architectural planning. Acoustic simulation tools are also used for designing sound reinforcement systems in churches, stadiums, train stations, and airport terminals.

2.1 Geometrical Acoustics: Ray Tracing and Image Sources

In geometrical acoustics the two basic models of geometrical sound propagation, ray tracing and image sources are applied. Often, however, the two philosophies are confused. It is important to highlight the differences in the physical meaning: Ray tracing describes a stochastic process of particle radiation and detection. Image sources are geometrically constructed sources which correspond to specular paths of sound rays. Often, image sources are constructed by using rays, beams, or cones, via a kind of ray tracing. Nevertheless, they remain to be "image source models". The fundamental difference between image sources and ray tracing is the way of calculation of contributions in impulse responses. Ray tracing only yields impulse response low-resolution data like envelopes in spectral and time domains. Image sources (via the classical method or via tracing rays, beams, cones, etc.) may be used for exact construction of amplitude and delay of reflections which narrow-band resolution depending on the filter specifications for wall reflection factors, for instance (Fig. 2).

2.2 Hybrid Models

Due to the contradictory advantages and disadvantages of ray tracing and image sources it was tried to combine the advantages in order to achieve high-precision results without spending too much complexity or computation time. Either ray tracing or radiosity algorithms were used to overcome the extremely high calculation time inherent in the image source model for simulation of the late part of the impulse response (adding a reverberation tail), or ray tracing was used to detect audible image sources in a kind of "forward audibility test". The idea behind is that a ray, beam, or cone detected by a receiver can be associated with an audible image source. The order, the indices, and the position of this image source can be reconstructed from the ray's history with storing the walls hit and the total free path. Hence the total travel time, the direction and the chain of image sources involved can be addressed to the image source. Almost all other algorithms used in commercial software are kind of dialects of the algorithms described above, and they differ in the way mixing of the specular with the scattered component is

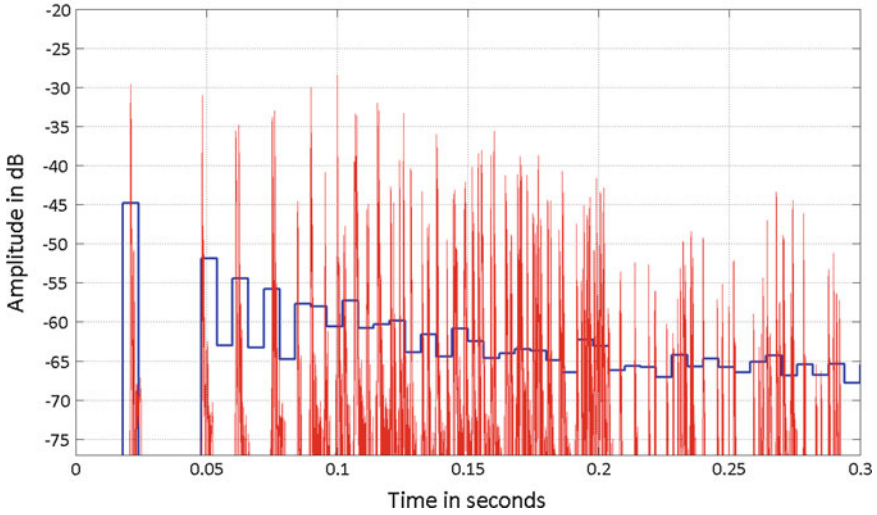


Fig. 2 Simulated room impulse response with comparison of reflections with precise time-resolution modeled by image sources (*red*), and time-quantized energy envelope modeled by ray tracing (*blue*)

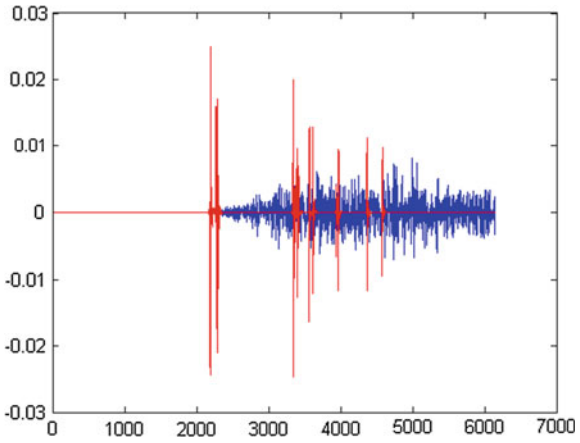


Fig. 3 Hybrid simulated room impulse response with early reflections by image sources (*red*) and late reflections, using the energy envelope predicted by ray tracing (*blue*). Scattered early reflections are handled by ray tracing instead of image sources

implemented. The specific choice of dialect depends on the type of results, particularly on the accuracy, spatial and temporal resolution (Fig. 3).

2.3 Room Acoustics Simulation at RWTH Aachen

At RWTH Aachen University, room acoustics simulation is in the focus since the mid 1980s, initially based on ray tracing and image source algorithms and later on

combinations of both approaches. Vorländer (1989) and independently van Maerke presented the basis for the cone, beam and pyramid tracing dialects, by showing that forward tracing is a very efficient method for finding audible image sources. Since then, the specular components of the room impulse response (RIR) were computable with high efficiency. The concepts of spatial subdivision were added for a quick processing of intersection tests which is the crucial subroutine in methods of Geometric Acoustics (GA). Then, during the 1990s it was shown that GA cannot solely be based on specular reflections (Vorländer 1995). The era of the implementation of scattering began with activities on the prediction, measurement, and standardization of scattering and diffusion coefficients of corrugated surfaces.

Progress in binaural technology enabled the incorporation of spatial attributes to room impulse responses. The key equation of the contribution of one room reflection, H_j, is given in frequency domain is

$$H_j = \frac{\mathrm{e}^{-\mathrm{j}kr_j}}{r_j} H_\mathrm{S} H_\mathrm{R} H_\mathrm{a} \prod_{i=1}^{n_j} R_i$$

where r_j is the reflection's travel distance, jkr_j the phase, $1/r_j$ the distance law of spherical waves, H_S the source directivity in source coordinates, H_a the low pass of air attenuation, R_i the reflection factors of the walls involved, and H_R the head-related transfer function of the sound incidence at a specified head orientation. The complete binaural room impulse response is composed of the direct sound and the sum of all reflections. This filter is appropriate for the convolution with anechoic signals to obtain audible results.

The stochastic part obtained from the ray tracing process must be post-processed in order to achieve an appropriate temporal resolution for audio sampling, but also to keep the spatial cues in a best possible way. This is implemented by using a shaped-noise technique with regard to spatial, temporal, and spectral attributes in the ray tracing results [see also (Schröder and Vorländer (2007)] (Fig. 4).

3 History of Acoustic Virtual Reality

3.1 VR Technology

In the early days of VR, head-mounted displays (HMDs) usually formed the heart of any VR-system in order to provide stereoscopic vision to the user. A HMD is a helmet-like display that features two small monitors positioned directly in front of the user's eyes to achieve stereopsis. Typically, a HMD is also equipped with earphones, where binaural synthesis has most often been used for the presentation of acoustic stimuli in the virtual 3-D space. However, due to fundamental problems such as wear comfort and user isolation from the real environment, today's HMDs are mainly used in low cost and mobile/portable VR-systems. Instead, especially in scientific and industrial applications, HMDs have been more and

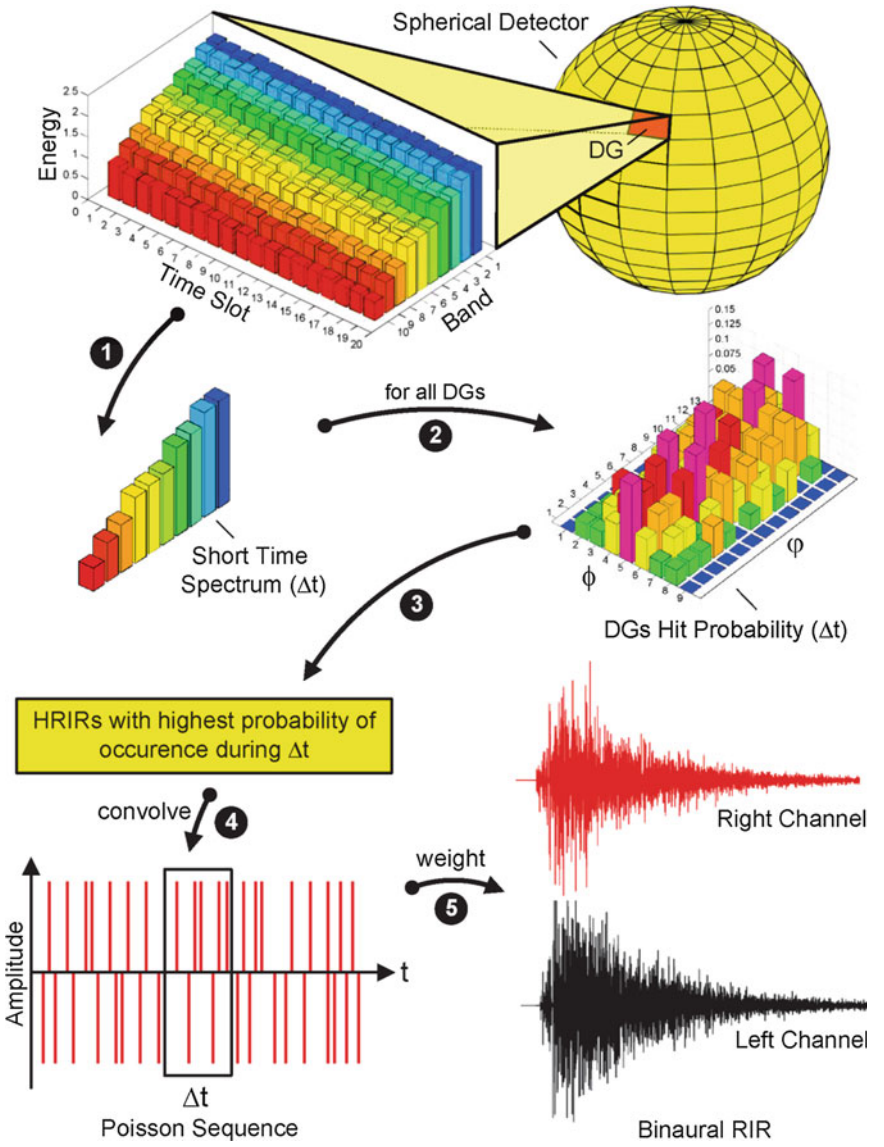


Fig. 4 Processing of ray tracing results sorted in directional groups (DG) to a binaural impulse response

more replaced by CAVE-like displays. These displays are room-mounted installations based on a combination of large projection screens that surround the user. Here, the stereoscopic vision is realized by means of light-weight polarized glasses that separate visual information from the stereoscopic projection. Since room-mounted VR-systems aim at an ergonomic, non-intrusive interaction as well as

co-located communication and collaboration between users, headphones should be avoided. As such, a sound reproduction based on loudspeakers is preferable. Head-mounted as well as room-mounted VR displays typically come along with a tracking system that captures the user's head position and orientation in real-time. This data is required for adapting the perspective of the visual scene to the user's current view-point and view direction. In addition, if binaural synthesis is applied for auralization, the user's position and orientation must be precisely known at any time in order to apply the correct pair of head related transfer functions (HRTF). A variety of tracking principles for VR-systems are in use, ranging from mechanical, acoustic (ultrasonic), and electro-magnetic techniques up to inertia and opto-electronical systems. Recently, a combination of two or more infrared (IR) cameras together with IR light reflecting markers that are attached to the stereo glasses, have become the most popular tracking systems. This type of tracking system is nearly non-intrusive, affordable and works with higher precision and lower latency in comparison to other technologies.

Real-time auralization systems have been investigated by many groups. Here to mention is the "EVE"-project at the TKK Helsinki University, Finland, and the "REVES" research project at INRIA, France. Recently, Open Source projects were launched such as "UNIVERSE". The aim of our VR activities is to create a reference platform for development and application of multimodal environments including high-quality acoustics. Such a system can be used for scientific research and testing as well as for development of complexity-reduced surround sound systems for professional audio or home entertainment. The group working on the acoustic VR-system is supported by the German Research Foundation, DFG, in a series of funded projects, where the Institute of Technical Acoustics, ITA, jointly worked with the Virtual Reality Group of RWTH Aachen University. The latter is the core group of a consortium of several institutions of our university and external partners covering the disciplines of computer science, architecture, civil engineering, mechanical engineering, electrical engineering, and information technology, psychology and medicine (Fig. 5).

In recent years, VR has proven its potential to provide an innovative human computer interface for applications areas such as architecture, product development, simulation science, or medicine. VR is characterized as a computer-generated scenario of objects. A user can interact with these objects in all three dimensions in real-time. Furthermore, multiple senses should be included to the interaction, i.e., besides the visual sense, the integration of other senses such as the auditory, the haptic/tactile, and the olfactory stimuli should be considered in order to achieve a more natural, intuitive interaction with the virtual world.

3.2 Audio Rendering

The final step in the real-time auralization chain is the convolution of the simulated impulse response with a dry excitation signal, usually speech, music, or ambient

Fig. 5 Users experiencing a virtual concert hall within the RWTH Aachen CAVE

sounds. Since room impulse responses are usually quite long, the convolution can become a computationally very intensive task. By now, powerful hardware and fast convolution algorithms exist that enable the realization of the entire audio rendering by means of high-quality FIR filtering. The mathematical background of convolution is well known and it can be easily implemented in time- or frequency-domain using MATLAB or similar tools. In contrast, convolution-based real-time audio rendering is an advanced problem in itself. Various requirements must be fulfilled, such as low latencies, rapid exchangeability of filters without audible artifacts and high signal-to-noise ratios. These can only be met by specialized algorithms. The state-of-the-art method is non-uniformly-partitioned convolution in the frequency-domain (Wefers and Vorländer 2012). It unites a high computational efficiency with low input-to-output latencies, by partitioning the filter impulse responses into a series of subfilters with increasing size. For a smooth exchange of filters cross-fading in the time-domain is commonly applied.

3.3 The Virtual Reality Center Aachen System

3.3.1 Background and Base Technology

Shortly after the establishment of the first VR developments at RWTH Aachen University, the activities in computer science were joined with those in acoustics. The advantage was that both groups had deep knowledge in their specific field so

that the competences could be combined with high synergy. The initial step was the integration of interactive VR technology (visual and haptic) with headphone-free audio reproduction. At that time the decision was made in favor of a stereo loudspeaker setup for an adaptive crosstalk cancellation. The first task was integrating head tracking and adaptive filters into a cross talk cancellation (CTC) system that turned out to be a flexible solution for various display environments (Lentz 2008).

In 2004 a CAVE-like five-sided surround-screen projection system was installed at RWTH Aachen University. Figure 6 shows an overview of the installation. It has a size of 3.6 × 2.7 × 2.7 m and can be reconfigured using a slide door and a movable wall. Stereoscopic images are produced by two images per screen with a resolution of 1,600 × 1,200 pixels each and are separated by polarized glasses. It uses several IR cameras for tracking several input devices and the user's head position/orientation. For the reproduction of acoustic signals, four loudspeakers are installed at the top of the CAVE. This setup was chosen over a simple stereo system in order to achieve a good binaural reproduction that is independent from the current user's orientation (see Figs. 7, 8).

3.4 The ViSTA Framework for Virtual Reality Software

At RWTH Aachen University, the VR Toolkit ViSTA has been under development for more than 10 years now in order to provide an open, flexible and efficient software platform for the realization of complex scientific and industrial applications. One of the key features of ViSTA comprises functionality for the creation

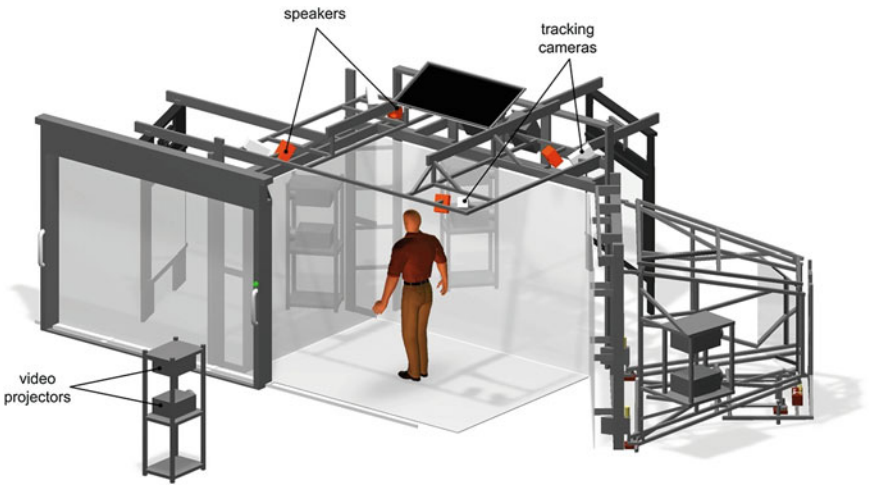


Fig. 6 The RWTH Aachen cave automated virtual environment (CAVE). The installation features five-sided passive stereoscopic vision (circular polarization) with optical IR tracking

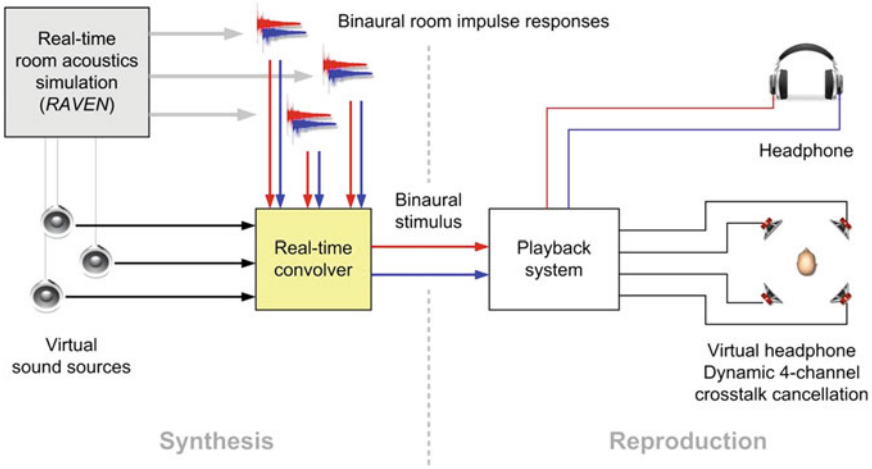


Fig. 7 Schematic block diagram of the VirKopf (Pelzer and Vorländer 2010) real-time auralization system

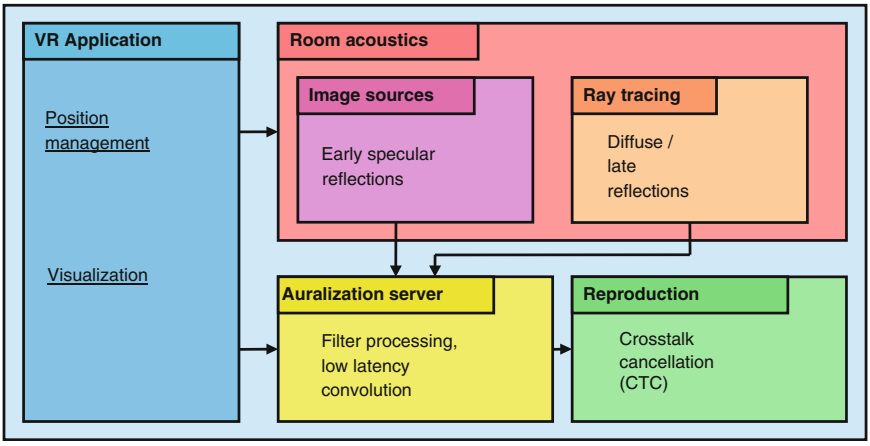


Fig. 8 Software components and their interaction

of multimodal interaction metaphors, including visual, haptic, and acoustic stimuli. For such elaborate, multimodal interfaces, flexible sharing of different types of data with low latency access is needed while maintaining a common temporal context. Therefore, ViSTA comes along with a high-performance device driver architecture. It provides a novel approach for history recording of input by means of a ring buffer concept that guarantees both, a low latency and a consistent temporal access to device samples at the same time.

In acoustical reproduction based on binaural synthesis and crosstalk cancellation, latency of the (optical) tracking system is especially critical. For this reason,

a compensation scheme has been developed for ViSTA that, based on current tracking samples, can predict the state of the human head position and orientation for the time of application.

3.5 Sketch-Based Interaction

Apart from the multimodal reproduction of a scene, it is also important that one can interact with the virtual environment in an intuitive way. An example for this is a virtual room acoustics laboratory, where the user can perform modifications of the scenery—e.g., by changing the material properties of a wall, creating a piano, or adjusting a reflector panel (see Fig. 9), and then directly perceive the impact on the acoustics. For this purpose, special interaction techniques are required that match the demands of immersive virtual environment, i.e., they are easy-to-use, use small and light-weight input devices and avoid disturbing graphical interface elements. Consequently, a sketch-based interface was developed for the interaction with architectural sceneries where the user can draw three-dimensional command symbols that are then recognized by a real-time symbol matching algorithm. Recognized symbols execute commands such as the creation of a window, and can also contain additional information such as the size and position of the window. The sketch-based interaction provides a considerable number of possible commands that are quickly executable by using an intuitive pen-like input device.

3.6 Sound Field Rendering

For real-time sound field rendering, the hybrid room acoustics simulation software room acoustics for virtual environments (RAVEN) was integrated into the VISTA

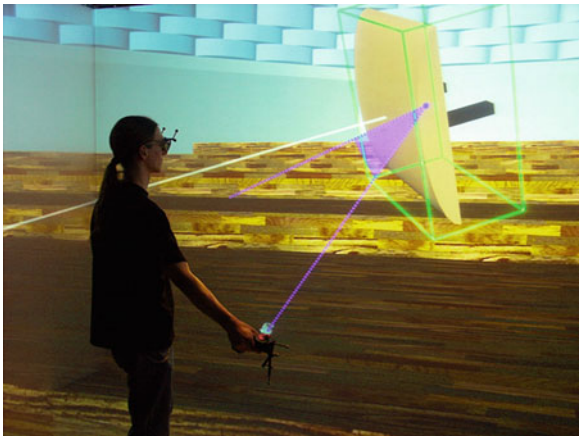


Fig. 9 Sketch-based interaction in a virtual concert hall. The user modifies the properties of an acoustic absorber element just by gestures with a pointing device in his *right hand*

frame-work as a network service. RAVEN combines a deterministic image source method with a stochastic ray-tracing algorithm in order to compute high quality room impulse responses on the basis of geometrical acoustics. However, basic methods of GA do not cover two important wave phenomena, that is sound transmission and sound diffraction. Thus, these methods fail to correctly simulate sound propagation from a hidden source to a receiver where the direct line of sight is blocked by other objects, e.g., an obstacle or a door to an adjacent room. Therefore, more sophisticated simulation techniques are required which reflect the real world experience. RAVEN therefore includes the simulation of the sound phenomena of sound transmission and sound diffraction (Pelzer and Vorländer 2010).

Sound transmission prediction tools are well established. They are based on statistical energy analysis and enable the calculation of energy transmission via sound and vibration transmission paths. The implementation of sound transmission in auralization software can be done using filters that are interpolated from spectra of transmission coefficients, with secondary sources radiating transmitted sound in adjacent volumes. In contrast, diffraction is—especially in real-time systems—often neglected or poorly modeled due to its analytical complexity. However, the lack of diffraction causes a significant error in most simulations. This becomes more evident by the example of a simple noise barrier that separates a sound source from a receiver. Here, a shadow zone grows clearer and sharper with increasing sound frequency. This zone results from a total cancellation of the incident wave by the diffraction wave which is radiated from the object's edges or perimeter to the receiver. Due to a matter of principle of GA, that is the linear propagation of sound rays, basic methods fail to detect any sound energy inside the shadow zone of such a barrier. Fortunately, analytical and stochastic diffraction models based on GA have been developed which maintain a smooth transition from the view zone to the shadow zone, meanwhile even in more complex scenarios (Fig. 10).

RAVEN imposes no constraints on scene interaction, meaning that not only sound sources and receivers can move freely in the virtual environment, but also the scene geometry can be manipulated by the user at runtime. This is achieved by using advanced modularized and flexible data structures that separate the simulation into single parallel processes that are then distributed and processed on a computing cluster (see below).

3.7 Diffraction

As mentioned above, RAVEN also accounts for the wave phenomenon of diffraction using a hybrid approach for the simulation of sound diffraction, which allows the simulation of higher-order edge diffraction. For this purpose, existing GA methods of edge diffraction have been adapted and optimized. The concept of secondary sound sources by Svensson was chosen for the image source method, as

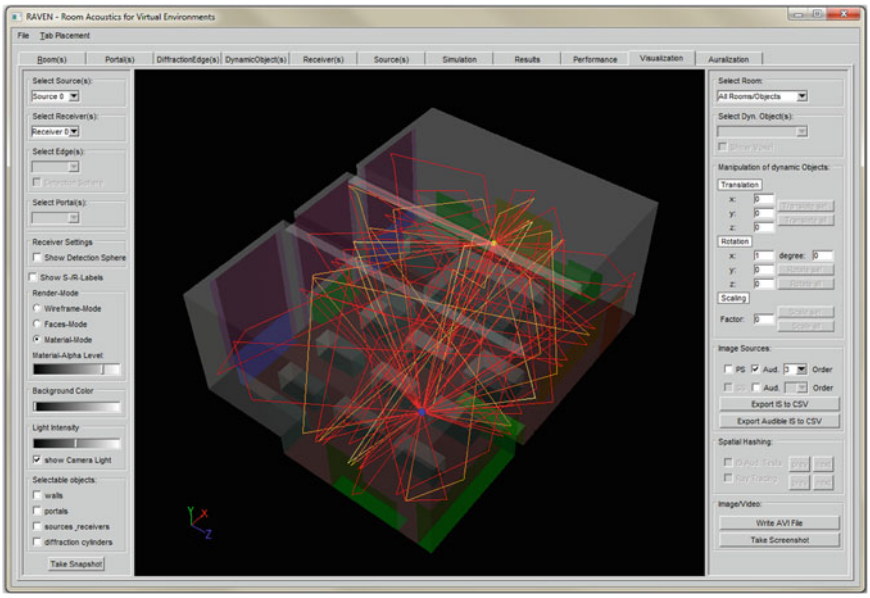
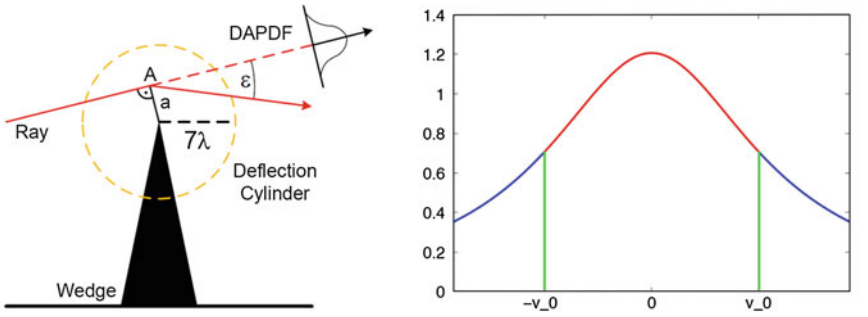


Fig. 10 Graphical user interface of the RAVEN simulation framework, here with a visualization of early reflections in a classroom

the method allows an exact analytic description of higher order diffraction of finite edges. Here, the demand for a very fast and accurate prediction conflicted with the problem of multiple diffracted reflections and the mirroring of the secondary sources, as the complexity of diffraction path searches and number of secondary sources rises exponentially with diffraction order. Therefore, two types of diffraction edges were introduced, static and dynamic edges. Static diffraction edges cannot be manipulated during the simulation, i.e., they cannot be moved or changed in size. This allows the pre-computation of visible edges and secondary sources, which are organized in efficient tree-graphs. By using these data structures diffraction paths up to a range from three to five can be taken into consideration for the online simulation (see Fig. 11a). For dynamic diffraction edges, i.e., edges that are fully scalable and moveable, this order must be reduced to at most order two due to the complexity of regenerating the graphs for higher order diffraction. However, it should be kept in mind that this affects only the actual process of manipulation. Once it has been modified, the object's state switches back into static mode.

The diffraction method by Stephenson, which is based on Heisenberg's uncertainty principle, was integrated in RAVEN's stochastic ray tracer. The core of this approach is the computation of a 2D-deflection-angle-probability-density - function (DAPDF) of energy particles when they pass an edge. This diffraction model fits perfectly to algorithms that model sound propagation as the dispersion of energy particles, such as stochastic ray tracing. Unfortunately, this approach is



(a) The edge's influence on the particle is described by the DAPDF.

(b) Deflection angle probability density function as function of v, $D_0 = 1$.

Fig. 11 Edge diffraction implementation in RAVEN using deflection angle probability density functions (DAPDF)

also computational demanding due to the underlying principle of energy dispersion for higher order diffraction that leads to a very large number of required energy particles. This problem was tackled by the introduction of cylindrical edge detectors, called deflection cylinders. A deflection cylinder counts impacting energy, which is then distributed to other detectors. This approach is not exact, but prevents the explosion of computing and delivers very good results in simple test scenarios. A detailed validation is still pending, though.

3.8 Geometry Manipulation

Another important design aspect for interactive room acoustics simulation is the creation of highly flexible algorithmic interaction interfaces that support a maximum degree of freedom in terms of user interactivity. While code adjustments for operations such as the exchange of material parameters and the manipulation of portal states were relatively easy to implement, the requirements of a modifiable geometry turned out to be an algorithmic challenge. After first test implementations it became apparent that RAVEN's acceleration algorithms based on binary space partitioning (BSP) do not meet the criteria of dynamically manipulatable geometry since any modification calls for a recalculation of at least large parts of the BSP trees. It was therefore decided to introduce two different modi operandi for scene objects: static and dynamic (similar to the states of diffraction edges). Static objects, such as walls, are not modifiable during the simulation and are therefore processable in a quick and efficient way. A dynamic object, for instance a reflector panel, is adjustable by a user at runtime (see Fig. 9), though it should be emphasized that there is no limitation on the object's shape in general, i.e., the whole room geometry can be defined as dynamic. For the unconditioned

modification of dynamic objects, RAVEN switches to a new approach in geometry processing—that of spatial hashing (SH). SH is a method in Computer Graphics, which is usually applied for collision tests with deformable objects. The concept of SH is based on the idea of subdividing the space by primitive volumes called voxels and map the infinite voxelized space to a finite set of one-dimensional hash indices, i.e., a hash table (HT), which are en/decoded by a hash function (see Fig. 12). The advantage of SH over other spatial data structures such as BSP-trees is that the insertion/deletion of m vertices into/from the HT takes only O(m) time. Thus, this method is perfectly qualified to efficiently handle modifications of a polygonal scenery in order to enable a real-time auralization of a dynamically-changing environment. However, a comprehensive performance analysis has shown that the principle of SH can never compete with the performance of the fast BSP tree on a single core computing unit. On the other hand, the approach significantly gains performance from any additional CPU core as the HT data structure is efficiently schedulable in parallel. On a state-of-the-art multicore CPU (four or more cores), the SH approach will therefore outclass the BSP-based method if a fully modifiable geometry is desired.

4 Audio Reproduction System

4.1 Sound Generation

The system supports several sound generation methods: In the simplest case, a virtual sound source plays back a single mono audio file, which can be looped, if necessary. This simple modeling is sufficient for transient sounds, such as background noise. However, interaction with virtual objects often results in a multitude

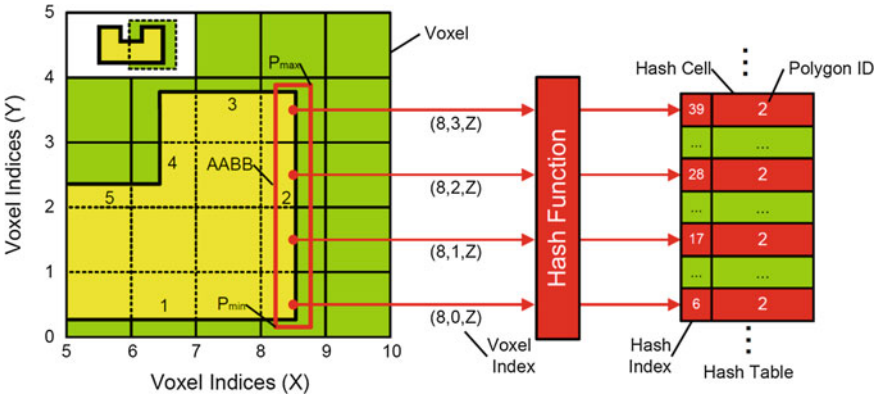


Fig. 12 Geometry data is organized using the acceleration technique 'spatial hashing' which allows fast updates on scene changes, such as moving a reflector panel or wall

of individual sounds and sound transitions. In order to adequately model these sounds, more advanced sound generation concepts are required. For this purpose, the system implements a sequencer, which allows to playback arbitrary sound samples on a virtual sound source. Medium synchronization between audio and video is ensured by using time codes. This technique applies for a wide range of objects, such as an electric sliding door or a virtual drum set, though it is not appropriate for objects whose sounds are driven by continuous parameters. An example is a virtual electric motor with freely adjustable revolution speed. For such applications the system offers real-time modal synthesis and post-processing filters can be added for high-quality sound authoring.

4.2 Dynamic Crosstalk Cancellation

Binaural playback strictly demands the ability to reproduce individual audio signals at each of the user's ears. Therefore, headphones are ideally suited for binaural playback. In contrast, sound emitted by a loudspeaker will—at least to a certain degree—always reach both ears of the listener. If left uncompensated, this crosstalk destroys the three-dimensional binaural cues. With knowledge of the sound propagation paths (speakers to each ear), a filter network can be designed that eliminates the crosstalk (see Fig. 13). This technology is known as crosstalk cancellation (CTC) and has been investigated for some decades now. Several filter design methods are known and different setups of loudspeakers are possible.

Since the user may freely move, the sound propagation paths change over time. Consequently, CTC filters must be adapted in order to keep the listener within the sweet spot and maintain a proper crosstalk cancellation. The position and orientation of the listener is obtained from the motion tracking with a frequency of 60 Hz. A threshold for translation (1° cm) and rotation (1°) is used to trigger the recalculation and update of the CTC filters. CTC filters prove to be stable only within certain angular ranges that depend on the orientation of the listener with respect to the loudspeaker setup. Only two loudspeakers are not sufficient to cover all possible user orientations within the CAVE. This problem is solved by combining multiple two-channel CTCs over a setup of four loudspeakers into a Dual-CTC algorithm. During runtime this method chooses the best speaker configuration by minimizing the compensation energy.

Using free-field HRTFs for compensating the sound propagation paths is valid only for anechoic conditions. The CAVE-like environment, however, is a confined space that is surrounded by acrylic glass walls. Reflections occur, which influence not only the crosstalk compensation, but also the binaural perception. Lentz (2008) investigated the impact of this issue on the localization performance for binaural auralization. A distortion of the perceived directions occurred, mainly in the elevation angle, though a combined audio-visual scenario reduced this mislocalization significantly.

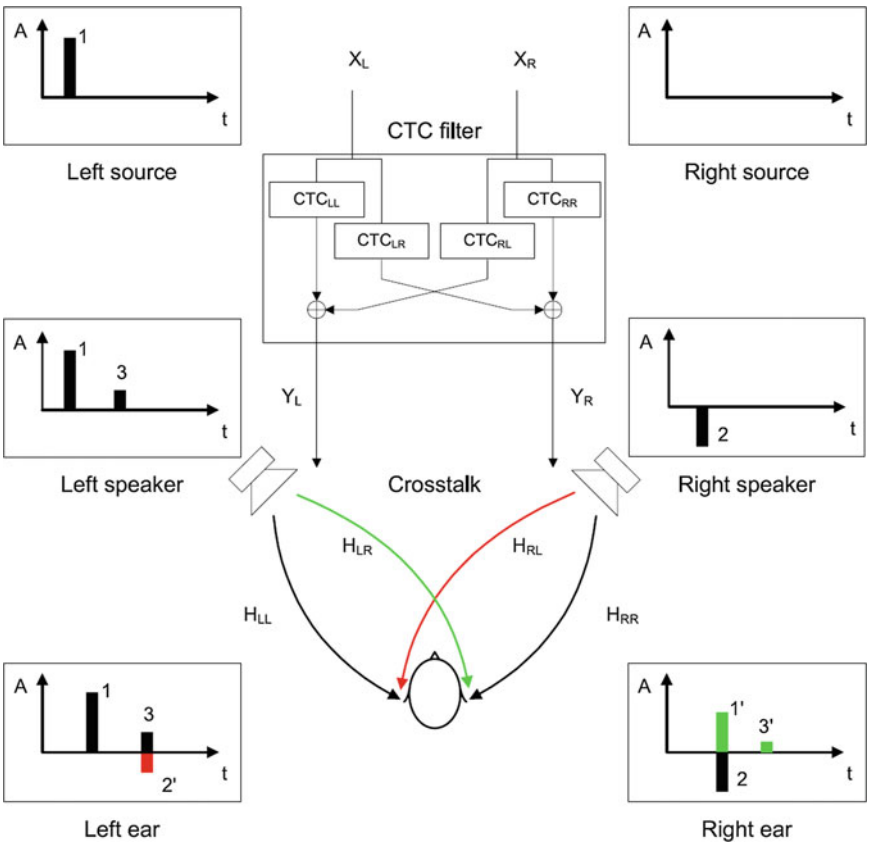


Fig. 13 Simplified overview of the crosstalk cancellation principle

4.3 Real-Time Convolution

The VR-system that is presented here performs real-time auralization on the basis of high-order FIR filters. The audio rendering itself is a complex and computationally intensive task, which is handled by a dedicated convolution engine. It filters the audio signal of each sound source with an independent binaural room impulse response (BRIR). For realistically sounding scenes, the signals of a high number of virtual sound source (50–100) must be convolved with the results of the room acoustics simulation. These auralization filters (BRIRs) typically consists of 20.000—200.000 filter coefficients. The latencies of the filtering must be very small (<20 ms) in order to reproduce the systems' reaction on user input (e.g., movement, rotation, actions) without audible delays. The overall system latency depends on the processing delay (input-to-output latency, but also additional delays for the exchange of filters). Moreover, the exchange of filters must not produce audible artifacts.

Time domain FIR filters (e.g., tapped delay-lines, TDLs) cannot be used for these extensive filter tasks. Their computational complexity scales linear with the filter length. The number of arithmetic operations becomes so large that even high-end computers or DSP cannot deliver the computation power for the requirements stated above. Specialized algorithms are needed.

Summarized under the term "fast convolution" several efficient methods for FIR filtering are known today. Many of them use the efficient fast fourier transform (FFT) to perform the convolution in the frequency-domain, where it can be implemented by simple multiplication of discrete Fourier spectra. FFT-based convolution is considered the most effective technique when the filters are long. Since the FFT is computed block-wise and not sample-wise, it introduces input-to-output latency equal to the block length of the audio stream. It is often misunderstood that one must use small FFTs in order to have small latencies. This is in general not true. One can for instance use one large 64 k-point FFT to process small stream blocks of 128 samples with an impulse response of 60,000 taps using the Overlap-Add scheme. In this case the filter is processed as a whole (single DFT spectrum), known as unpartitioned block convolution. One can show that it is more beneficial to split the filter into several smaller parts and convolve them individually, which is known as partitioned convolution. This keeps the delay low and allows for a much more efficient realization of the real-time filtering. Figure 14 shows an example of partitioned convolution with a non-uniform filter partitioning.

Splitting the filter impulse response into parts of equal length is called uniformly partitioned convolution. It can be easily implemented and delivers near to optimal computational efficiency for small filter (e.g., typically up to 1,024 taps). Therefore, it is the method of choice for free-field audio rendering, where only short HRIR filters (e.g., 100–300 taps) are used. For real-time room acoustics audio rendering the method is still not efficient enough. The most efficient of the currently known techniques is a non-uniformly partitioned convolution. Here the subfilter sizes vary and they are not fixed to a certain length as for the uniform

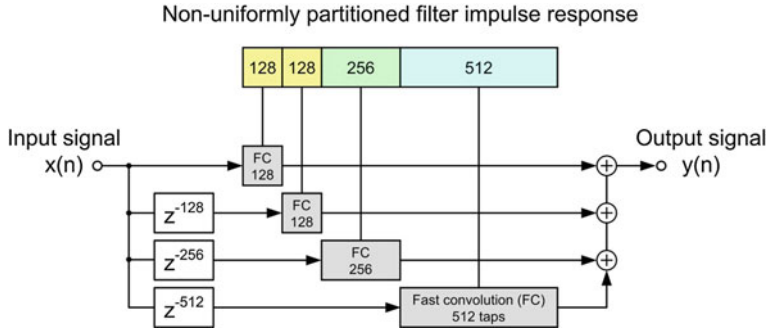


Fig. 14 An example for non-uniformly partitioned convolution. The filter impulse response is split into subfilters, which are realized using an individual fast convolution unit

case. At the beginning of filters it uses small block lengths to achieve short input-to-output latencies. Where affordable, it uses longer filter parts. These can be implemented with larger block length, which reduces the computational effort significantly. Table 1 illustrates the differences in computational complexity for several real-time FIR filtering techniques. The way the filter is partitioned is a key parameter for the algorithms, concerning the filter exchange, runtime stability, and computational efficiency. It is not trivial to find a partition which on the one hand minimize the computational effort, but on the other hand is actually realizable in a real-time system. Efficient optimization algorithms exist in order to maximize the convolution performance, but it is strictly necessary to consider aspects of realization as well in order to obtain practicable results.

The presented system uses a dedicated convolution engine called low-latency convolver (LLC). LLC has been developed at the Institute of Technical Acoustics for several years. It implements a parallelized non-uniformly partitioned fast convolution in the frequency-domain on multicore machines. A distinctive feature of LLC is the ability to allow an arbitrary impulse response partitioning, LLC uses a filter partitioning that is specifically optimized with respect to the available target hardware. The convolution algorithm is completely parallelized and can utilize the performance of current multi- and many-core systems. Since a hard switch of filters would result in signal discontinuities and thereby would cause unpleasant audible artifacts, time-domain cross-fading is applied to ensure a smooth transition.

5 Performance

5.1 Parallelization

Advanced real-time auralization concepts post high requirements on the computational performance. In order to meet these requirements, parallelization is extensively used in many components and on all architecture levels. On the most basic level, all arithmetically intensive computations are vectorized, using single-

Table 1 Comparison of the computational costs of several real-time FIR filtering methods

The VR-system that is presented here performs real-time auralization on the basis of high-filtering method	Computational cost (FLOPs/sample)	Speedup (Factor)
Time-domain FIR filter	129.999	1.00×
Unpartitioned block convolution (FFT-based, Overlap-Save)	29.108	4.46×
Uniformly-partitioned FFT convolution	4.142	31.39×
Non-uniformly partitioned FFT convolution with optimized filter partition	408	318.60×

The filter length is 65,000 taps and the streaming block length (latency) is 128 samples

instruction multiple-data (SIMD) instructions. Additionally, multi- and many-core computers are used to increase the performance and allow faster calculations. Ray tracing in particular can be efficiently parallelized using OpenMP. Furthermore, the crucial timing dependencies of the real-time convolution demands more advanced concepts. Therefore, special concepts such as flexible data structures are utilized for an efficient parallelization.

Optimizations for multicore machines allow realizing sceneries of medium complexity. The simulation of complex building environments exceeds the capabilities of a single PC with a limited maximum number of CPU cores. For the simulation of extensive sceneries with multiple coupled rooms, detailed geometries, and many sound sources, a cluster of multiple computers can be used to achieve the required computing performance. For this purpose, a cluster-capable version of RAVEN was developed, using MPI and the Viracocha-Framework. It allows distributing different subtasks of the computation, such as individual sound sources, frequency bands, or particle subsets to different cluster nodes. Specialized scheduling strategies distribute the computation evenly among all nodes. Furthermore, simulation tasks can be prioritized, guaranteeing that IS computations will not be delayed by prior, but less important RT calculations.

Since the top-level interface of RAVEN is unified, the underlying computation hardware is transparent and can be either a single computer or a computation cluster of varying size. This makes the approach very scalable so that the hardware can be chosen to match the complexity of the scenery. All in all, the optimized parallelization strategies make the room acoustics simulation fast enough for real-time processing (Fig. 15).

5.2 Real-Time Filtering Performance

Currently, a dedicated 2.4 GHz dual quad-core machine is used to realize the filtering. A RME Hammerfall series audio interface is used for sound input and output. Audio streaming is done using Steinberg's ASIO professional audio architecture, at 44.1 kHz with streaming buffersize (block length) of 512 samples. For BRIRs of 88,200 filter coefficients, LLC manages to filter the signals of more than 50 sound sources.

6 Future Work

An interesting idea for increasing the quality and speed of acoustic simulations at the same time has been introduced by frequency- and time-dependent room models (Pelzer and Vorländer 2010). This approach uses a set of models with graduated level of detail of the same scene geometry, where every single room model is optimized for a certain frequency range, which is important especially for correct

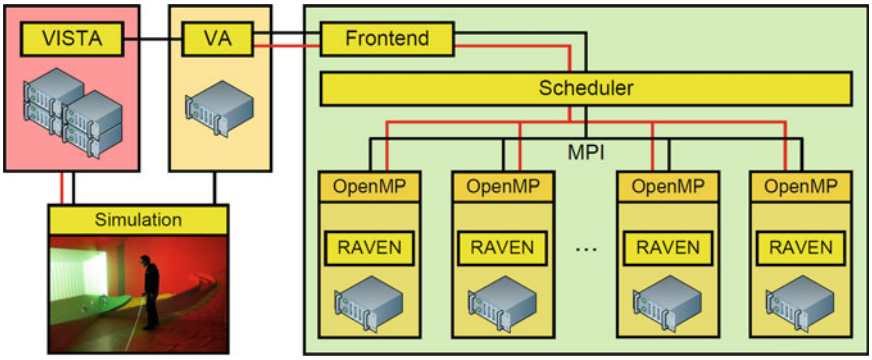


Fig. 15 Distributed room acoustics simulation system. Several cluster nodes are used to handle the large amount of room acoustics simulation tasks. All nodes communicate using MPI

reflection patterns at low frequencies. Additionally, with increasing time in the resulting impulse response, the level of detail can be decreased during the simulation, providing a total simulation speed-up of a factor of six when combined with the speed-up due to frequency-matched geometries. For even more complex scenes, the computation on graphic cards is very promising. This technique has also been successfully applied to auralization. For real-time GPU-based convolution, a highly optimized and promising solution was presented (Wefers and Berg 2010).

Room acoustic simulation by using geometrical acoustics is usually implemented with binaural receivers. Wave models such as FEM are easily applicable with binaural interfaces as well. This way, however, the signals are restricted to a specific set of HRTF, and a tedious task is to adapt the results to a proper reproduction system with very limited possibilities of listener individualization. With a more general interface such as spherical harmonics, room acoustic spatial data could be created in intermediate solutions. In post-processing this can lead to various binaural representations or to reproduction with Ambisonics. Then it must be discussed how standard routines in geometrical acoustics must be changed in order to implement multi-channel spherical microphone arrays. Furthermore, the corresponding output data can be multi-channel time signals or temporal SH coefficients or any other suitable spectral format. The amount of data and signal processing affects CPU time and memory. The discussion therefore is focused on feasibility and on consequences on the real-time performance on the one hand, and on the spatial quality of the room response, on the other.

For simulation of significant wave effects the next generation of room simulation models can be expected to include combination of wave acoustics and geometrical acoustics. Progress was made already in this direction but there remains lot of work to be done in order to provide a robust method and the necessary boundary conditions of surfaces. The latter is still an underestimated problem, as every simulation methods can only be sufficiently accurate with adequate input data.

Acknowledgments The authors would like to thank the German Research Foundation (DFG) for funding a series of projects. Torsten Kuhlen and his team at Virtual Reality Center Aachen (VRCA) are acknowledged for excellent and smooth collaboration.

References

Lentz, T. (2008). Binaural technology for virtual reality. *Ph.D. dissertation*, RWTH Aachen University.

Pelzer, S., & Vorländer, M. (2010). Frequency- and time-dependent geometry for real-time auralizations. In *20th International Congress on Acoustics (ICA)*, Sydney, Australia.

Schröder, D., & Vorländer, M. (2007). Hybrid method for room acoustic simulation in real-time. In *Proceedings of the 20th International Congress on Acoustics* (*ICA*), Spain.

Vorländer, M. (1989). Simulation of the transient and steady state sound propagation in rooms using a new combined sound particle—image source algorithm. *The Journal of the Acoustical Society of America, 86*, 172–178.

Vorländer, M. (1995). International round robin on room acoustical computer simulations. In: *Proceedings of 15th International Congress on Acoustics*, Trondheim, Norway.

Vorländer, M. (2008). *Auralization: Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Berlin: Springer.

Wefers, F., & Berg, J. (2010). High-performance real-time FIR-filtering using fast convolution on graphics hard-ware. In *Conference on Digital Audio Effects (DaFX)*.

Wefers, F., & Vorländer, M. (2012). Optimal filter partitions for non-uniformly partitioned convolution. *AES 45th International Conference on Time-Frequency Audio Processing*, Helsinki, Finland.

[Online]: www.vrca.rwth-aachen.de.

Further Reading

Schroeder, M. R., Atal, B. S.,& Bird, C. (1962). Digital computers in room acoustics. In *Proceedings of the 4th International. Congress on Acoustics*, Copenhagen, Denmark.

Vian, J., & van Maercke, D. (1986). Calculation of the room impulse response using a ray-tracing method. In *Proceedings of the ICA Symposium on Acoustics and Theatre Planning for the Performing Arts,* Vancouver, Canada.

Cruz-Neira, C., Sandin, D., DeFanti, T., Kenyon, R., & Hart, J. (1992). The CAVE: Audio visual experience automatic virtual environment. *Communications of the ACM, 35*, 64–72.

Dalenbäck, B.-I. (1995). A new model for room acoustic prediction and auralization. *Ph.D. dissertation*, Chalmers University, Gothenburg, Sweden.

Gardner, W. G. (1995). Efficient convolution without input-output delay. *Journal of the Audio Engineering Society (JAES), 43*, 127–136.

Egelmeers, G. P. M., & Sommen, P. (1996). A new method for efficient convolution in frequency domain by non-uniform partitioning for adaptive filtering. *IEEE Transactions on signal processing, 44*.

Stephenson, U. (1996). Quantized pryamidal beam tracing—a new algorithm for room acoustics and noise immission prognosis. *ACTA ACUSTICA united with ACUSTICA, 82*, 517–525.

Svensson, U. P., Fred, R. I., & Vanderkooy, J. (1999). An analytic secondary source model of edge diffraction impulse responses. *Journal of the Acoustical Society of America, 106*, 2331–2344.

Müller-Tomfelde, C, (2001). Time-varying filter in non-uniform block convolution. In *Proceedings of the Conference on Digital Audio Effects (DAFX-01)*.

Tsingos, N., Funkhouser, T., Ngan, A., & Carlbom, I. (2001). Modeling acoustics in virtual environments using the uniform theory of diffraction. *ACM Computer Graphics, SIGGRAPH'01 Proceedings* (545–552).

Lokki, T. (2002). Physically-based auralization—design, implementation, and evaluation. *Ph.D. dissertation*, Helsinki University of Technology.

García, G. (2002). Optimal filter partition for efficient convolution with short input/output delay. In *Proceedings of 113th AES convention*.

Hammershøi, D., & Møller, H. (2002). Methods for bin-aural recording and reproduction. *Acustica united with Acta Acustica, 88*, 303.

Teschner, M., Heidelberger, B., Müller, M., Pomeranets, D., & Gross, M. (2003). *Optimized Spatial Hashing for Collision Detection of Deformable Objects*. VMV '03.

Tsingos, N., Gallo, E., & Drettakis, G. (2004). Perceptual audio rendering of complex virtual environments. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*, *3*(23).

Funkhouser, T., Tsingos, N., Carlbom, I., Elko, G., Sondhi, M., West, J. E., et al. (2004). A beam tracing method for interactive architectural acoustics. *Journal of the Acoustical Society of America, 115*, 739–756.

Gerndt, A., Hentschel, B., Wolter, M., Kuhlen, T., & Bischof, C. (2004). Viracocha: An efficient parallelization framework for large-scale CFD post-processing in virtual environments. In *Proceedings of the 2004 ACM/IEEE Conference on Supercomputing*, Magdeburg, Germany.

Thaden, R. (2005). Auralisation in building acoustics, *Ph.D. dissertation*, RWTH Aachen University.

Lentz, T. (2005). Performance of spatial audio using dynamic cross-talk cancellation. In *119th AES Convention*, New York, NY, USA.

Cox, T. J., Dalenbäck, B.-I. L., Antonio, P. D., Embrechts, J. J., Jeon, J. Y., Mommertz, E., et al. (2006). A tutorial on scattering and diffusion coefficients for room acoustic surfaces. *Acta Acustica united with ACUSTICA, 92*, 1–15.

Schröder, D., & Lentz, T. (2006). Real-time processing of image sources using binary space partitioning. *Journal of the Audio Engineering Society (JAES), 54*(7/8), 604–619.

Raghuvanshi, N., & Lin, M. (2006). Interactive sound synthesis for large scale environments. In *Proceedings of the Symposium on Interactive 3D Graphics and Games*, Red-wood City, USA.

Lentz, T. (2006). Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments. *Journal of the Audio Engineering Society (JAES), 54*(4), 283–293.

Schröder, D., & Vorländer, M. (2007). Hybrid method for room acoustic simulation in real-time. In *Proceedings of the 20th International Congress on Acoustics (ICA)*, Madrid, Spain.

Stephenson, U. M., & Svensson, U. P. (2007). An improved energetic approach to diffraction based on the uncertainty principle. In *Proceedings of the 19th International Congress on Acoustics*, Madrid, Spain.

Schröder, D., Dross, P.,& Vorländer, M. (2007). A fast reverberation estimator for virtual environments. In *Proceedings of the 30th AES International Conference*, Saariselkä, Finland.

Assenmacher, I., & Kuhlen, T. (2008). The vista virtual reality toolkit. In *Proceedings of the IEEE VR SEARIS*, (pp. 23–26).

Schröder, D., & Assenmacher, I. (2008). Real-time auralization of modifiable rooms. In *2nd ASA-EAA joint conference Acoustics*, Paris, France.

Lundén, P. (2008). Universe acoustic simulation system: Interactive realtime room acoustic simulation in dynamic 3d environments. *The Journal of the Acoustical Society of America (JASA), 123*(5), 3937–3937.

Rausch, D., & Assenmacher, I. (2008). A sketch-based interface for architectural modification in virtual environments. In *5th Workshop VR/AR*, Magdeburg, Germany.

Noisternig, M., Katz, B., Siltanen, S., & Savioja, L. (2008). Framework for real-time auralization in architectural acoustics. *Acta Acustica United with Acustica, 94*(6), 1000–1015.

Rindel, J.-H. (2009). Auralisation of a symphony orchestra—the chain from musical instruments to the eardrums. In *EAA Symposium on Auralization*, Espoo, Finland.

Schröder, D., & Pohl, A. (2009). Real-time hybrid simulation method including edge diffraction. In *Proceedings of the EAA Auralization Symposium*, Espoo, Finland.

Wefers, F., & Schröder, D. (2009). Real-time auralization of coupled rooms. In *Proceedings of the EAA Auralization Symposium*, Espoo, Finland.

Tsingos, N. (2009). Using programmable graphics hardware for auralization. In *Proceedings of the EAA Symposium on Auralization*, Espoo, Finland.

Schröder, D., Svensson, P., & Vorländer, M. (2010). Open measurements of edge diffraction from a noise barrier scale model. In *Proceedings of the International Symposium on Room Acoustics (ISRA)*, Melbourne, Australia.

Schröder, D., Ryba, A., & Vorländer, M. (2010). Spatial data structures for dynamic acoustic virtual reality. In *Proceedings of the 20th International Congress on Acoustics (ICA)*, Sydney, Australia.

Schröder, D., Ryba, A., & Vorländer, M. (2010). Real-time auralization of dynamically changing environments. *Submitted to Acta Acustica united with Acustica.*

Dalenbäck, B.-I. (2010). Engineering principles and techniques in room acoustics prediction. In *Baltic-Nordic Acoustics Meeting*, Bergen, Norway.

Assenmacher, I., Rausch, D., & Kuhlen, T. (2010). On device driver architectures for virtual reality toolkits, *presence: Teleoperators and virtual environments*, *23*, 83–95.

The Wave Field Synthesis Lab at the HAW Hamburg

Wolfgang Fohl

1 Introduction

Wave field synthesis (WFS) is a technology to create a realistic acoustical impression of sound sources located in various places outside and—with certain limitations—inside the listening room. This is accomplished by driving a loudspeaker array with sound signals whose superposition creates the desired sound field.

The technology of wave field synthesis (WFS) has been developed and refined for more than two decades, for an overview see de Bruijn (2004). At present, there are systems commercially available that show a sufficient maturity and stability to employ them as lab equipment to create virtual-reality or augmented-reality environments for human interaction with virtual audio objects. In the year 2011 the department of computer science of the Hamburg University of applied sciences (HAW) equipped a lab room with a WFS system to create an augmented-reality audio (ARA) environment. In this article the experiences of the first half year of operation shall be reported. The following sections of this article start with an overview of WFS basics, then the installed system is described, followed by a description of additional lab systems that support the WFS system. The article closes with an overview of current and future projects in the WFS-ARA lab.

2 WFS Fundamentals

In this section a very short explanation of the WFS technology is given. A detailed overview of WFS fundamentals and applications is given in the thesis of Baalman (2008) and in the article of de Vries (2008).

W. Fohl (✉)
University of Applied Sciences, Hamburg, Germany
e-mail: fohl@informatik.haw-hamburg.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_10,

According to Huygens' Principle, each point of the wavefront can be considered the origin of an elementary wave. The superposition of all these elementary waves creates the original wave of the source (see Fig. 1).

For an ideal resynthesis of the sound field, the ensemble of secondary sources would have to be some sort of membrane where the displacement of each surface element can be controlled by an external device. In order to apply this principle to a real-world system, several modifications have to be applied.

Discrete actuators: Practically, the secondary sources are realized by an array of discrete loudspeakers instead of a continuous membrane. This will give rise to *spatial aliasing* effects, if the distance between the loudspeakers is too large compared to the sound wavelength. Spatial aliasing means, that there are audible artefacts in the sound, when moving along the loudspeaker array (see Fig. 2). Thus, a narrow loudspeaker spacing is desirable. The synthesized sound field for sufficiently low frequencies is shown in Fig. 3.

Limit to two dimensions: Instead of creating a closed 3D-surface around the listening room, the reconstruction is reduced to two dimensions in that the loudspeakers are lined up around a plane along the walls of the listening room.

Sources within the listening room: In extension of Huygens' law it is often desirable to simulate sources *within* the listening room (so-called *focused sources*). This can be accomplished by creating a *concave* sound field, where all wave fronts are directed towards a focal point in front of the loudspeaker array in the listening room. A listener in front of this focal point will perceive the sound source located in the focal point. If the listener moves between the focal point and the loudspeakers, the sound will be perceived as originating from the loudspeakers.

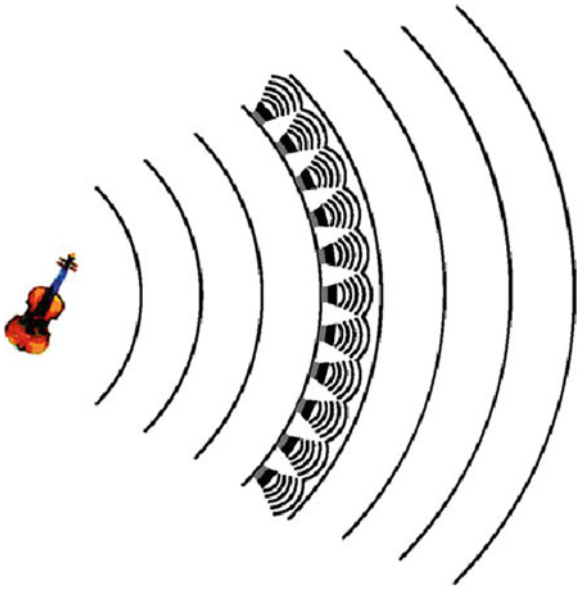


Fig. 1 Huygens' principle. The sound field of the primary source can alternatively be created by a set of secondary sources located along the wave front (Corteel and Caulkins 2004)

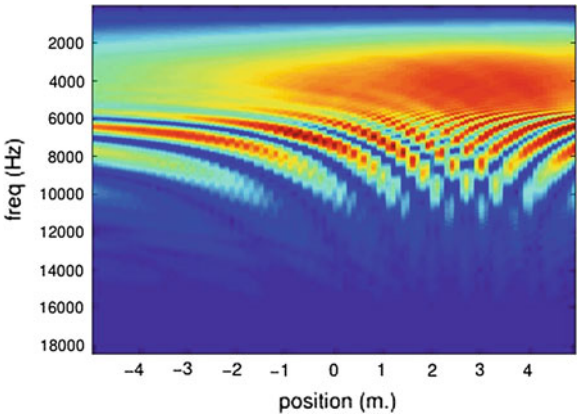


Fig. 2 Spatial aliasing. For frequencies above approx. 3 kHz, the frequency response in front of the speaker array depends strongly on the listening position *parallel* to the array. (Baalman 2008)

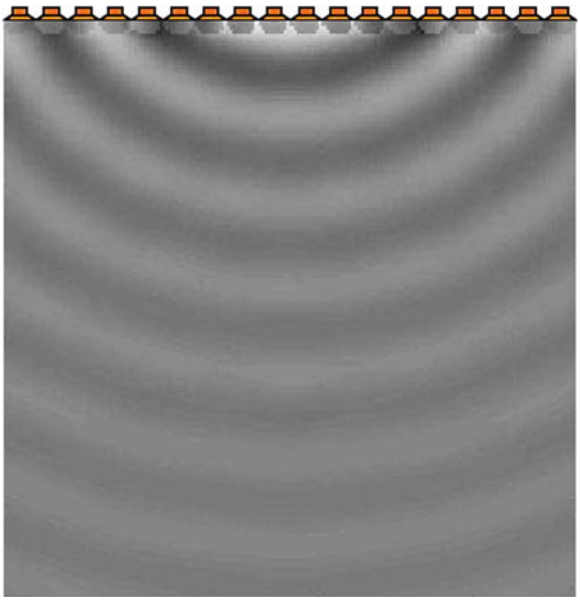


Fig. 3 Two-dimensional wave field of a point source (located *behind* the speaker array) (Bleda et al. 2003)

Figure 4 shows a focused source. Persons standing below the focal point will have the perception of a source positioned within the room.

Plane waves: In order to represent sources in indefinite distance, it is necessary to create plane waves. A practical application for plane wave sources is the simulation of reflections at the walls of a virtual room.

Directivity of the loudspeakers: In addition to the sound generated by the loudspeakers, there is also sound reflected by the walls, floor, and ceiling of the listening room, possibly creating unwanted interferences. Walls and ceilings will have to be equipped with curtains and carpets. Furthermore, the loudspeakers should be designed to have a wide radiation angle in horizontal direction, but a narrow angle in vertical direction.

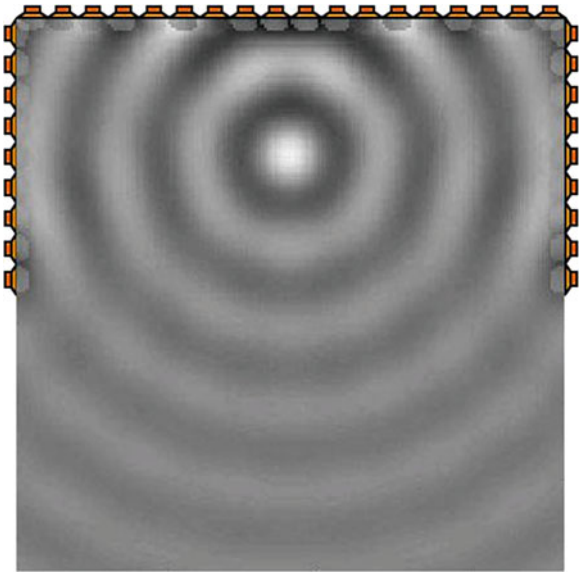


Fig. 4 Two-dimensional wave field of a focused source (located *in front* of the loudspeaker array) (Bleda et al. 2003)

Summary of source types:

- *Point source:* A point source is located behind the loudspeaker array.
- *Focused source:* This source is located in front of the loudspeaker array. It will only be perceived correctly for listeners on the far side of the listening room.
- *Plane waves:* Parallel wavefronts, mainly to simulate room reflections.

2.1 Typical WFS Scenarios

Mimic conventional surround stereo setups: To simulate a 5.1 speaker system, five point sources are positioned at the desired speaker positions. One benefit of the WFS system is that these virtual speakers can be positioned outside the listening room, thus enlarging the sweet spot.

Also Ambisonic speaker setups may be simulated this way, though the positions are limited to the plane of the WFS loudspeaker array.

Virtual stage, virtual concert room: Each performer is represented by a point source or a focused source located on the virtual stage. Plane waves (usually eight, for angles of incidence in steps of 45°) to simulate room reflections.

Virtual-reality/augmented-reality environment: virtual sound objects are represented by point or focused sources, source characteristics, as position, volume, and others, are controlled by interaction with the user(s).

3 Description of the WFS System at the HAW Hamburg

3.1 System Architecture

The WFS system consists of the loudspeaker modules and the rendering hardware and software. The whole system has been supplied by (Four Audio GmbH 2012), the rendering subsystem was provided by the team of Stephan Weinzierl at the TU Berlin (Department of audio communication 2012).

The lab room of the WFS system at the HAW Hamburg has a size of 6 × 7 m. Six loudspeaker modules are installed at the short sides and seven modules at the long sides of the room. Each module is 80 cm long and contains 8 channels, so the distance between two channels is 10 cm. The resulting number of channels is 208.

The driving signals for each of these 208 loudspeaker channels have to be calculated individually in realtime. These calculations are performed by a cluster of 3 PCs with an Ubuntu Studio operating system. The kernels of the PCs are modified for realtime operation. The user interface and the audio applications are executed on a frontend computer, an Apple MacPro with OS X operating system. The Mac and the PCs of the rendering cluster are interconnected by a local Ethernet network. Only the Mac has an internet access and can be interfaced to additional systems via WLAN.

Not only is the calculation of the audio signals for each of the 208 channels a time-critical task, the signals have also to be transported to the appropriate loudspeaker modules with virtually no delay. To that purpose, loudspeaker modules, rendering PCs, and frontend computer are interconnected by an DANTE audio network. This network solution has the benefit of fewer cables to be installed compared to a point-to-point connection of loudspeakers and MADI interfaces in the PCs. The DANTE audio network protocol is a layer on top of the TCP/IP protocol stack. It is a proprietary protocol developed by Audinate (Audinate Pty Ltd 2013) (Fig. 5).

3.2 Loudspeaker Modules

To obtain a high frequency limit for spatial aliasing, the spacing of the loudspeakers has to be as close as possible. The installed modules contain eight loudspeaker channels in a distance of 10 cm. Each module also contains two woofers, which are shared between four channels. To create a narrow vertical radiation angle, each channel consists of 3 tweeters aligned horizontally on top of each other. The desired radiation characteristic is obtained by beam forming. Beam forming is implemented by FIR filters for each of the three tweeters of a channel. These FIR filters are contained in the loudspeaker modules.

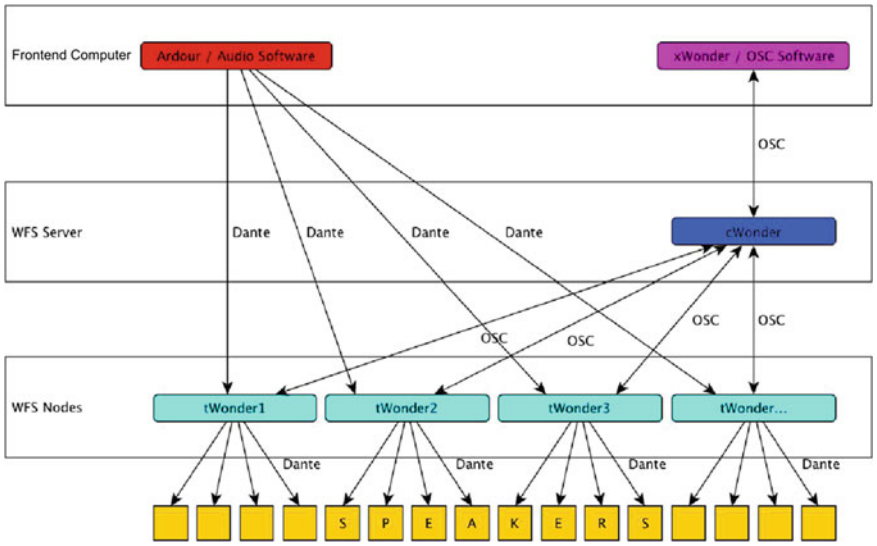


Fig. 5 Distribution of the WFS software components (FourAudio 2011)

An important requirement for the WFS loudspeaker modules is the maximum sound pressure level (SPL) per channel. For an ensemble of sources to be rendered in their correct loudness relations, it has to be considered, that the number of loudspeakers involved in rendering one source varies with the distance of the virtual source from the array: the farther away from the array, the more channels are involved, the closer to the array, the fewer channels are involved in rendering the source. An extreme situation is a point source located exactly at the position of one of the loudspeakers. I this case, this channel has to supply the complete power for this source. This case is crucial for the design of the installed audio power per channel. The system at the HAW has an installed audio power of 125 W per channel, resulting in a maximum SPL of 105 dB per channel.

In summary, a loudspeaker module contains the speakers, power supply, amplifiers, DANTE network interface, and the FIR filters for establishing the directivity and for compensation of the listening room acoustics. These filters are configurable via the network (Fig. 6).

3.3 Software

The rendering of the virtual sources, i.e., the calculation of the individual audio signals for each channel is performed by the software package Wonder (Baalman 2012). This software is developed at the TU Berlin in the department of audio communication. Wonder is open source software licensed under the GPL. It is a distributed software system designed to run on multiple Linux PCs forming a rendering cluster. The main components of the software are:

cWonder	Control and coordination of all other software components running on the cluster
xWonder	GUI for the management of WFS sources
tWonder	Programmable delay lines for each channel (time domain)
fWonder	FIR filters for frequency-domain signal manipulations

Fig. 6 Loudspeaker module. This is a predecessor model of the modules installed at the HAW Hamburg. Every channel consists of a *vertical* line of three tweeters (*bottom of the module*). Four channels share one woofer (*top of the module*) (Goertz et al. 2007)

The controller component (cWonder) transmits the audio signal of each WFS source to the WFS nodes. The WFS nodes are running multiple instances of the tWonder program to apply the appropriate time delays to the source signal and optionally apply frequency modifications by means of the fWonder software before transmitting it to the loudspeaker channels. The software in combination with the DANTE audio network is capable of handling 128 separate WFS sources.

The source properties like position, type (point or plane wave) can be controlled by the GUI provided by the xWonder component. Figure 7 shows a screenshot of the GUI.

The xWonder software component is running on the frontend computer and transmits the control information via Open Sound Control (OSC) messages. OSC is an audio control protocol transmitted via Ethernet connections.

Here are some examples for OSC messages to control WFS sources [from (Baalman 2008)]:

OSC Message	Action
/WONDER/source/position(id, x, y, z, t, dur)	Move source identified by *id* to position (*x, y, z*). Movement starts at time *t* and lasts *dur* seconds
/WONDER/source/angle(id, angle, t, dur)	Change angle of source (Only meaningful for plane wave sources)
/WONDER/source/type(id, type)	Change source type (point or plane wave)
/WONDER/source/mute(id, mute)	Mute source

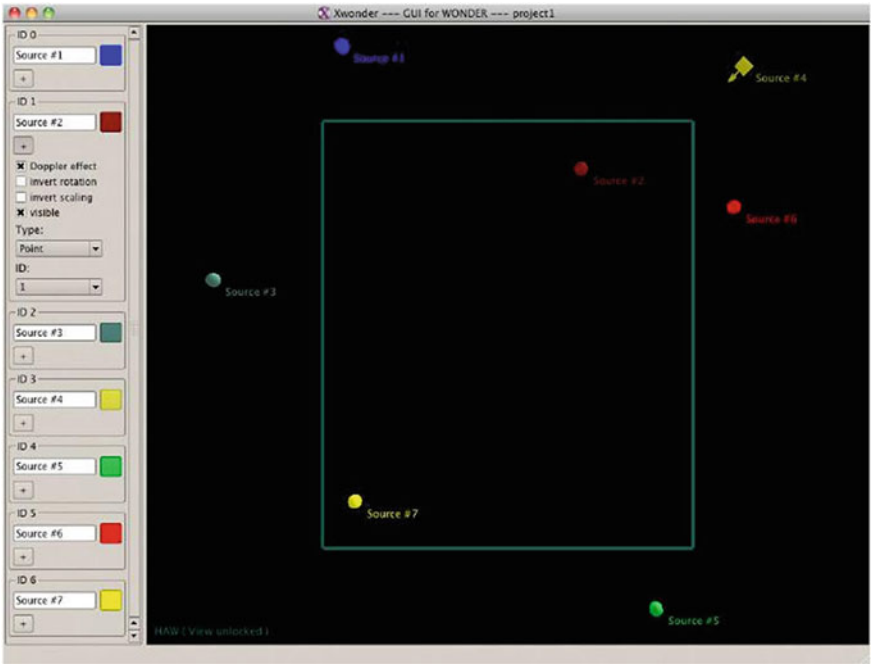


Fig. 7 Screenshot of the xWonder GUI

Every other software capable of sending OSC messages (like Csound, Super Collider and many others) may also send commands to the WFS system to control the sources. Since OSC is a simple text-based protocol, it is easy to write own programs that send OSC messages to control the system.

3.4 Audio Interfaces

As already stated, frontend and cluster computers, as well as loudspeaker modules are interconnected by the DANTE audio network. To the computers these network connections appear as a sound card with 128 inputs and 128 outputs. Each input or output represents one WFS source, so the number of I/O connections effectively limits the number of WFS sources.

The routing between the audio applications on the frontend computer and the DANTE-I/O ports is made by the software JACK, a low-latency audio routing server running on all computers of the WFS system. This offers a great flexibility of system configuration as alternative signal routings can easily be obtained. For example in the startup procedure, the standard audio routing is modified to test all of the 208 loudspeaker channels. The test program is a Super Collider script

running on the frontend computer. In the first stage, the first 104 audio outputs are routed to the first half of the 208 loudspeaker channels, and a short signal is played on each loudspeaker. Then the routing is switched, so that the audio outputs are routed to the second half of the loudspeakers. In this way, the correct function of each loudspeaker channel can be tested in reasonable time. Afterwards, the standard routing of the audio outputs to the WFS sources is established, and the WFS system is ready for operation. Jack is even capable of audio routing to remote computers via internet.

4 Auxiliary Lab Components

The WFS system is complemented by various other stand-alone systems located in the WFS lab.

4.1 Camera-Based Position Tracking System

The listening area of the WFS system is monitored by 6 infrared cameras. These cameras are calibrated to identify 3D-targets as shown in Fig. 8. From the evaluation of the six video streams of the cameras, the tracking system determines the location and orientation of multiple targets in realtime. These data is broadcasted via WLAN and is thus available for the frontend computer of the WFS system. In the next chapter an example project is presented that shows, how these data are utilized for a gesture control of WFS sources.

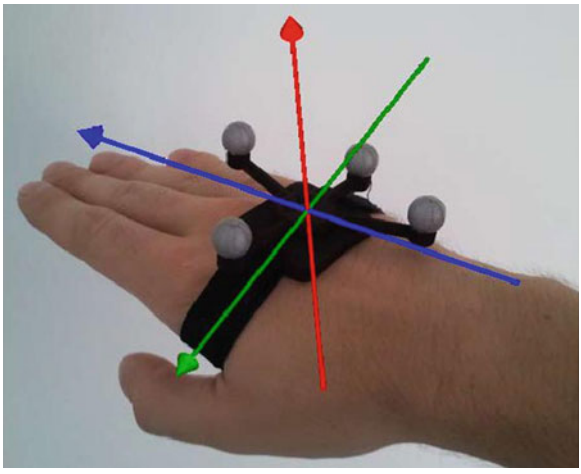


Fig. 8 A 3D-target applied to the hand for gesture control. *Blue x-axis, green y-axis, red z-axis* of the target coordinate system (Fohl and Nogalski 2013)

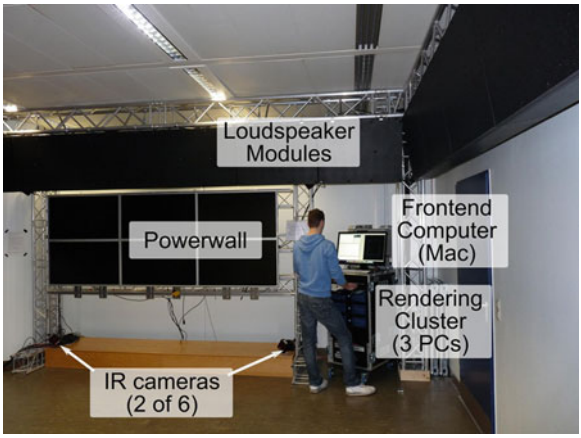


Fig. 9 View of the WFS lab

4.2 Video Power Wall

A video power wall consisting of a 2 × 3 monitor matrix is located at the front side of the listening area and allows the presentation of large-scale images and videos to add a visual component to the acoustical presentations (see (Fig. 9).

4.3 Sound Field Microphone

For spatial audio recordings, and for acoustic localization of users, the lab is equipped with a sound field microphone (TSL 2013). With this microphone it is possible to record a plane-wave-decomposition of the sound field at the microphone position. Based on these recordings, audio material for playback on the WFS system can be generated and positions of sound sources can be calculated.

5 Projects

5.1 Gesture Control of Sound Objects

The goal of this project is the control of sound objects (i.e.,WFS sources) by hand movements (Fohl and Nogalski 2013). A 3D-target (see Fig. 8) is fixed at the user's wrist and the tracking data is monitored by a software developed in the course of the project. The software processes the tracking data, applies the transformations between the different coordinate systems of the tracker and the WFS system, and detects the gestures for object control:

Select	Raise the hand above 45°
Deselect	Lower the hand below 45°
Move the source on a circular segment around the user	Turn the arm to the left or right
Get the source nearer/farther away	Bend/stretch the arm

When gestures are detected, the appropriate OSC messages are sent to the controller of the rendering cluster.

5.2 Distributed Concert Room/Virtual Conference Room

At the same time when the WFS system was installed at the HAW Hamburg, an identical system has been installed at the Hochschule für Musik und Theater (HfMT) in Hamburg. At the inauguration ceremony of this system, a distributed performance of John Cage's piece FIVE was presented: Three musicians were located at the HfMT, the remaining two were at the HAW. The music of the remote musicians was transferred by the JACK software via Ethernet to the other room and played by the WFS, so that in both rooms a complete rehearsal could be heard. This first proof-of-concept is going to be refined and extended.

The same setup can be used to create a virtual conference room. Here it is possible by means of the tracking system to have the audible locations of the remote participants synchronized with their real positions and movements in the remote room.

5.3 Source Visualization with Mobile Devices

The goal is an augmented-reality application for mobile (at the moment: Android-based) devices. The camera image is overlaid with a graphical representation of the sound source. The source then can be modified by the usual touch screen actions. A similar setup is described in (Delerue and Warusfel 2006).

5.4 Modifications of the Rendering Software

For interactive realtime applications, several aspects of the original Wonder software will have to be modified:

Latency: Audio Buffer sizes of all interconnections in the system have to be examined and minimized to reduce the latency from the current value of 110 ms to a value well below 20 ms. Small buffer sizes however bear the risk of audio dropouts, when audio samples cannot be delivered in time. The software will have

to be recompiled with reduced values for the buffer sizes to determine the minimum possible size without encountering audio dropouts.

Focused sources: When focused sources (sources in front of the loudspeakers) are rendered, a choice has to be made by the software, which loudspeakers to use for the rendering. This choice in consequence defines, in which part of the room the wave field of the source is correct. In the current version of the software the decision is only based on the source position: The loudspeakers with the smallest distance to the source are creating the sound field. In interactive augmented-reality applications, the choice of the rendering loudspeakers has to be based on the *user's* position to ensure that the wave field is correct at this part of the listening rooms. Special strategies will have to be developed for multiuser applications.

6 Summary

This article has showed the capabilities and principles of operation of the WFS lab at the HAW Hamburg. First projects have been presented and further project ideas have been shown.

After almost a year of operation, the system shows up to be stable and reliable. Since the rendering software is published under the GPL, is appears feasible to modify it in a way to support the special requirements arising from the interactive usage of the system. Very promising is the cooperation with the sibling system at the HfMT in Hamburg, the concept of a distributed concert room will be developed further in the near future.

Readers interested in project cooperation or in a demonstration of the system are kindly invited to contact the author.

Acknowledgments Thanks to Four Audio for the kind permission to use the data and figures of the technical manual of the system for this article.

References

Audinate Pty Ltd (2013). Audinate Homepage, http://www.audinate.com/index.php.

Baalman, M. A. (2008). On wave field synthesis and electro-acoustic music, with a particular focus on the reproduction of arbitrarily shaped sound sources, Ph.D. Thesis, TU Berlin.

Baalman, M. A. (2012). Wonder project page, http://sourceforge.net/projects/swonder/.

Bleda, S., López, J. J., Pueo, B. (2003). Software for the simulation, performance analysis and real-time implementation of wave field synthesis systems for 3d-audio. In *Proceedings of the 6th International conference on digital audio effects* (*DAFX-03*), London, UK.

Corteel, E., & Caulkins, T. (2004). *Sound scene creation and manipulation using wave field synthesis*. Paris: IRCAM.

de Bruijn, W. (2004). Application of wave field synthesis in videoconferencing, Ph.D. Thesis, Technische Universiteit, Delft.

de Vries, D. (2008). Wave field synthesis: history, state-of-the-art and future. In Second international symposium on universal communication.

Delerue, O., & Warusfel, O. (2006). Mixage mobile. In *IHM '06: Proceedings of the 18th International Conference of the Association Francophone d'Interaction Homme-Machine.*

Department of audio communication (2012). Department homepage, http://www.ak.tu-Nogalski, synthese-Anlage berlin.de/menue/fachgebiet_audiokommunikation/.

Fohl, W., and Nogalski, M. (2013). *A gesture control interface for a wave field synthesis system*. Submitted to the international conference on new interfaces for musical expression (NIME'13) Daejeon/Seoul, Korea Republic.

Four Audio GmbH (2012) Four Audio Homepage, http://www.fouraudio.com/.

Four Audio (2011). *Anleitung zur Wellenfeldsynthese-Anlage*, .

Goertz, A., Makarski, M., Moldrzyk, C., & Weinzierl, S. (2007). *Entwicklung eines achtkanaligen Lautsprechermoduls für die Wellenfeldsynthese*.

TSL Professional Products Ltd. (2013). Soundfield homepage, http://www.soundfield.com/.

The *μ*·cos*m* Project: An Introspective Platform to Study Intelligent Agents in the Context of Music Ensemble Improvisation

Jonas Braasch

1 Introduction

Automated music agents have a long tradition in Artificial Intelligence (AI) research. Starting first as composition tools (Cope 1987; Friberg 1991; Widmer 1994; Jacob 1996), computers are meanwhile sufficiently fast to allow these systems to improvise music with others in real time. Typically music composition/improvisation systems use a symbolic language, most commonly in form of the Musical Instrument Digital Interface (MIDI) format. Successful systems such as Lewis' *Voyager* system (Lewis 2000) and Pachet's *Continuator* (Pachet 2004) use MIDI data to interact with an individual performer whose sound is converted to MIDI using an audio-to-MIDI converter. The research described in this paper stems from a larger project with the goal of developing a *Creative Artificially-Intuitive and Reasoning Agent* Caira. Instead of using the simple audio-to-MIDI converter, the agent uses standard techniques of Computational Auditory Scene Analysis (CASA), including pitch perception, tracking of rhythmical structures, and timbre and texture recognition (see Fig. 1). The CASA approach allows Caira to extract further parameters related to sonic textures and gestures in addition to traditional music parameters such as duration, pitch, and volume. This multi-level architecture enables Caira to process sound using bottom-up processes simulating intuitive listening and music performance skills as well as top-down processes in the form of logic-based reasoning. The low-level stages are characterized by a Hidden Markov Model (HMM) to recognize musical gestures and an evolutionary algorithm to create new material from memorized sound events. The evolutionary algorithm presents audio material processed from the input sound which the agent trains itself on during a given session, or from audio material that has been learned by the agent in a prior live session. The material is analyzed using the HMM

J. Braasch (✉)
Director, Center for Cognition, Communication, and Culture,
Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA
e-mail: braasj@rpi.edu

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_11,

machine listening tools and CASA modules, restructured through the evolutionary algorithms and then presented in the context of what is being played live by the other musicians.

The logic-based reasoning system has been designed for Caira so she can "understand" basic concepts of music and use a hypothesis-driven approach to perform with other musicians (see top-down processes in Fig. 1). The benefits of including a logic-based reasoning system are many. Firstly, we hope to see this multi-level approach lead to a more natural system response by trading off several techniques, thus making the underlying processes less transparent to the human musicians while not lessening the overall responsiveness of the system. Secondly, we would like the agent to be able to create new forms of music with the specific goal that the agent be able to develop her own concepts by expanding and breaking rules and monitoring the outcome of these paradigm changes. Thirdly, we want to document the performance of the system, which is not easy to do, when the agent simulates intuitive listening in the context of Free Music. By adding a logic-based reasoning system, we can now assess communication between the agent and the human musicians by comparing the internal states of the agent and the human musicians. In our project, foot switches are used to record the internal states of the human participants.

This paper focuses on the third goal for our logic-based reasoning stage. In particular, I will describe a self-exploratory approach to test the performance of Caira within a trio ensemble. In my μ·cos*m* approach (pronounced: microcosm), I control two independent musical instruments, the second one using a foot-operated interface, to probe the inter-ensemble communication skills of Caira. The

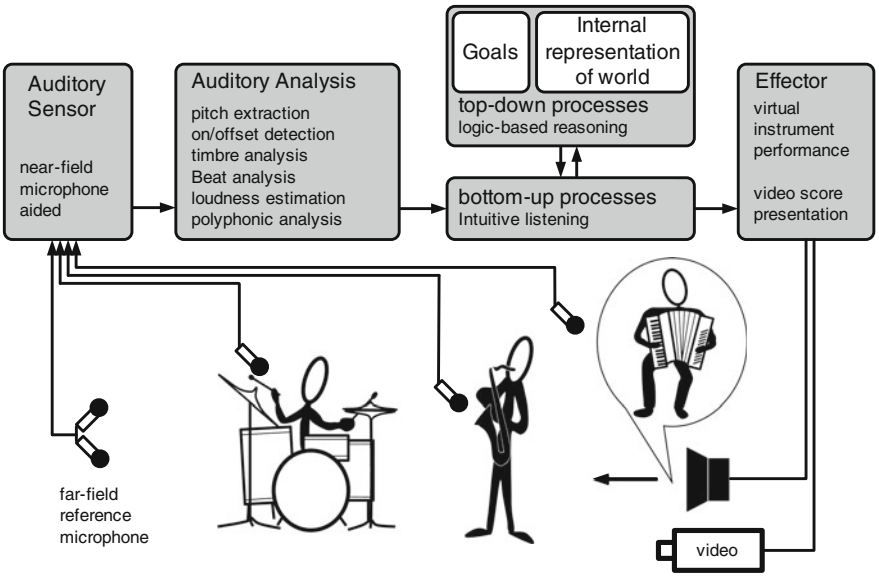


Fig. 1 Schematic of the creative artificially-intuitive and reasoning agent Caira

approach, which will be described in further detail below, is inspired by experimental ethnomusicology methods practiced by Arom (Arom 1967) and others. A more detailed description of the Caira lower and higher level architecture and her ability to operate using the fundamental concepts of music ensemble interaction will precede this discussion.

2 Gestalt-Based Improvisation Model Based on Intuitive Listening

The artificially-intuitive listening and music performance processes of Caira are simulated using the *Freely Improvising, Learning and Transforming Evolutionary Recombination system* (FILTER) (Van Nort et al. 2009, 2010, 2012). The FILTER system uses a Hidden Markov Model (HMM) for sonic gesture recognition, and it utilizes Genetic Algorithms (GA) for the creation of sonic material. In the first step, the system extracts spectral and temporal sound features on a continuous basis and tracks onsets and offsets from a filtered version of the signal. The analyzed cues are processed through a set of parallel Hidden Markov Model (HMM) -based gesture recognizers. The recognizer determines a vector of probabilities in relation to a dictionary of reference gestures. The vector analysis is used to determine parameters related to maximum likelihood and confidence, and the data is then used to set the crossover, fitness, mutation, and evolution rate of the genetic algorithm, which acts on the parameter output space (Van Nort et al. 2009).

3 Logic-Based Reasoning Driven World Model

In order to better understand the relationship between bottom-up and top-down mechanisms of creativity, a knowledge-based top-down model complements the bottom-up stages that were described in the previous two sections. Caira's knowledge-based system is described using first-order logic notation [for a detailed description of Caira's ontology see (Braasch et al. 2011a, b)]. For example Caira knows that every musician has an associated time-varying dynamic level in seven ascending values from *tacit* to *ff*. The agent possesses some fundamental knowledge of music structure recognition based on jazz music practice. It knows what a solo is and understands that musicians take turns in playing solos, while being accompanied by the remaining ensemble. The agent also has a set of beliefs. For example it can be instructed to believe that every soloist should perform exactly one solo per piece.

One of the key analysis parameters for Caira is the estimation of the tension arc, which describes the current perceived tension of an improvisation. In this context, the term 'arc' is derived from common practice of gradually increasing the tension until the climax of a performance part is reached and then gradually decreasing

tension to end it. While tension often has the shape of an arc over time, it can also follow other trajectories. It is noteworthy that we are not focusing here on tonal tension curves that are typically only a few bars long (i.e., demonstrating low tension whenever the tonal structure is resolved and the tonic appears). Instead, we are interested in longer structures, describing a parameter relates to *Emotional Force* (McAdams et al. 2002) as well.

Using the individual microphone signals, the agent tracks the running loudness of each musical instrument using the Dynamic Loudness Model of (Chalupper and Fastl 2002). The Dynamic Loudness Model is based on a fairly complex simulation of the auditory periphery that includes the simulation of auditory filters and masking effects. Additionally, the psychoacoustic parameters of roughness and sharpness are calculated according to (Danial and Weber 1997; Zwicker and Fastl 1999). In its current implementation, Caira estimates tension arcs for each musician from simulated psychophysical parameters. Based on these perceptual parameters and through its logic capabilities, the system recognizes different configurations for various patterns. For example it realizes that one of the musicians is performing an accompanied solo, by noticing that the performer is louder and has a denser texture than the remaining performers. The system can also notice that the tension arc is reaching a climax when all musicians perform denser ensemble textures. Caira takes action by either adapting her music performance to the analysis results, or by presenting a dynamic visual score. Caira can, for example, suggest that a performer should end his or her solo, because it is becoming too long or it can encourage another musician to take more initiative. It can guide endings and help an ensemble to fuse its sounds together.

4 Tension Arc Calculation

In a previous study, we decided to calculate the tension arcs T from a combination of loudness L and roughness data R (Braasch et al. 2011):

$$T = L^4 + a \cdot L^3,$$

with an adjusting factor a. In a further study, we also suggested including *information rate* [e.g., as defined by Dubnov (2003), Dubnov et al. (2006)] as an additional parameter for the tension arc calculation (Braasch et al. 2012). A real-time capable solution was developed to measure the rate and range of notes per 2-second time interval. To achieve this, pitch is measured using the YIN algorithm (de Cheveigné and Kawahara 2002a, b) and converted to MIDI note numbers. Next, the number of notes is counted within a 2 s interval, ignoring the repetition of identical notes. The standard deviation of the note sequence is then determined from the list of MIDI note numbers. Finally, the information rate is determined from the product of *number of notes* and *standard deviation of MIDI note numbers*. Practically, we measured values between 0 and 100. The tension curve is calculated using the following equation:

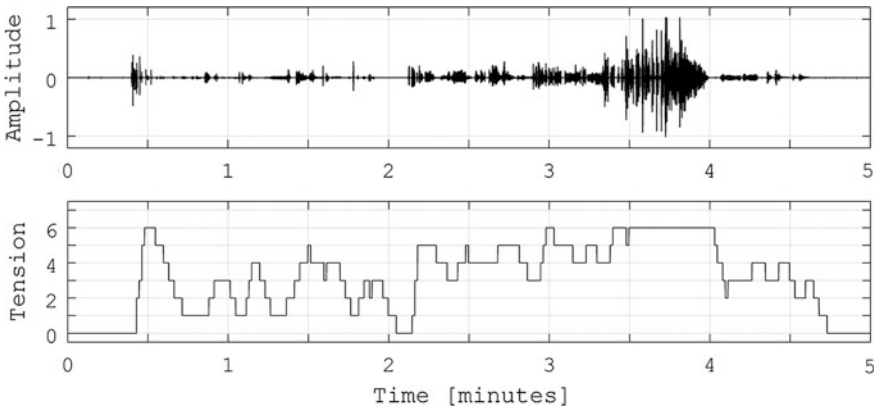


Fig. 2 Tension arc calculation for a soprano saxophone sample. *Top* Waveform of the saxophone, recorded with a closely positioned microphone. *Bottom* Calculated tension arc curve

$$T = \frac{1}{a+b}(a \cdot L + b \cdot ((1-q) \cdot R + q \cdot I)),$$

with the Information Rate I, Loudness L, and Roughness R. Note that all parameters, *L, R, and I* are normalized between 0 and 1 and the exponential relationships between the input parameters and T are also factored into these variables. The parameter q is the quality factor from the YIN pitch algorithm. A value of one indicates a very tonal signal with a strong strength of pitch, while a value of zero indicates a noisy signal without defined pitch. The parameter is used to trade off roughness and information rate between tonal and noise-like signals. The parameters a and b are used to adjust the balance of loudness and the other input parameters for individual instruments. All tension curves are scaled integer values between zero and seven. Figure 2 shows an example of how a tension curve is estimated from the instruments' sound pressure signal.

5 Ensemble State Calculations

A Bayesian model was used to find an a posteriori estimation of the most likely ensemble state from the obtained tension curves. The ensemble states describe the instantaneous relationships between the musicians of an ensemble using methods in jazz ensemble practice. To keep the interaction sufficiently simple, we define six Ensemble States E for a trio shown in the schematic in Fig. 3:

1. Solo A: Performer A performs a solo part
2. Solo B: Performer B performs a solo part
3. Solo C (Caira): Caira performs a solo part

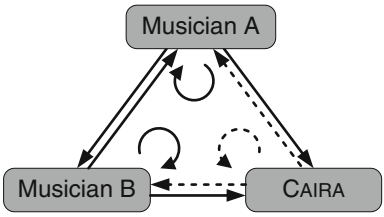


Fig. 3 Schematic communication scheme for a free music trio performance. Each musician has to establish individual communication channels to all other musicians and also observe his/herself. *Dashed lines* symbolize the agent's machine listening channels

4. Low-Tension Tutti: All ensemble members perform a tutti part with low tension
5. High-Tension Tutti: All ensemble members perform a tutti part with high tension
6. End: All musicians come to an end.

The Ensemble States are determined using a logic-based reasoning approach published in (Braasch et al. 2011a), the practical rules that were derived in this study are given in Fig. 4. We cannot assume that each of the six states is performed equally long in time, but by using a Bayesian approach we can improve the Ensemble State estimation by recording how often each state occurs as a percentage over the whole training duration. To this purpose, the human performers use a foot pedal to update the Ensemble State. In addition, we can compare the states with instrumentally measured parameters. To see the general approach, let us focus on the analysis of the time-variant tension curves of Musicians *A* and *B*. We define seven discrete levels of Tension *T*. Curves will be computed for each participating musician and for Caira, so we have 3 tension curves: (T_a(t), T_b(t), T_c(t)). We can compute how often each tension level combination is observed for a given ensemble state:

$$p(E|T_{a,b}) = \frac{p(T_{a,b}|E)p(E)}{p(T_{a,b})}.$$

The parameter $T_{a,b}$ is the observed combined Tension *T* for Musicians *A* and *B*. The Tension Curve T_c is not part of the analysis, since the intelligent agent Caira will observe the other two musicians to predict the current Ensemble State *E*. We have 49 discrete values for $T_{a,b}$ (7·7 Tension State combinations). The term $p(T_{a,b}|E)$ is the likelihood that the joint Tension Curve $T_{a,b}$ is observed for a given Ensemble State *E*. The term $p(E)$ is the probability that State *E* occurs independently of the tension curve status, and $p(T_{a,b})$ is the probability that the joint Tension Curve $T_{a,b}$ occurs independently of the ensemble state. Using the Equation given above we can compute the posterior estimate for each possible Ensemble

Ensemble States	Musician A	Musician B	Caira C
1 Solo A	T_A+1> T_B	T_B–1< T_A	T_C–1< T_A *
2 Solo B	T_A –1< T_B	T_B +1> T_A	T_C–1< T_B*
3 Solo C	0< T_A <5	0< T_B <5	Decision needed
4 Low Tension Tutti	0< T_A <5	0< T_B <5	Decision needed
5 High Tension Tutti	T_A >4	T_B >4	T_C>4*
6 Ending	T_A ==0	T_B ==0	T_C=0*

Fig. 4 Ensemble State calculations based on logic-based reasoning. The variables T_A, T_B, and T_C represent the tension curves of Musicians A, B, and Caira. The *asterisks* denote that Caira does not have to follow the suggestions by the other two musicians but can also respond by using a different tension curve level

State E_1–E_7 for any Tension Curve pair $T_{a,b.}$. An Ensemble State curve will be discussed further below (see also Fig. 5).

6 The *μ*·cos*m* Trio

Since my key interest became to study our intelligent music system Caira in the context of ensemble work, I needed to find an ensemble work with. Our general ensemble to study the Caira project is a trio named *Triple Point* consisting of Pauline Oliveros (V-Accordion, an physical modeling synthesizer with an accordion-type user interface), Doug Van Nort (*Granular-feedback Expanded Instrument System*, GREIS, a laptop-based system to generate electronic textural and gestural sounds), and myself on the soprano saxophone. While Triple Point turned our to be an effective way to test the two systems Caira and FILTER, I was keen to possess an alternative, flexible tool, so that I could test the ensemble capabilities of Caira whenever I wanted, without having to gather other human musicians for a test.

In Western music tradition, the most common way to simulate an ensemble is to use a keyboard instrument, for example the piano, to simulate one instrument with the right hand and another with the left. However, since Caira is working with actual sound textures and her capabilities go beyond symbolic, note-based music representation, a more flexible musical instrument was needed to complete the *μ*·cos*m* trio. As a solution, I started to use an Arturia Moog Modular V Synthesizer, a piece of software that simulates the Moog Synthesizer. The instrument is flexible in its sound generation, can produce sound textures as well as traditional note-based material and has a recognizable sound characteristic. The latter is important for

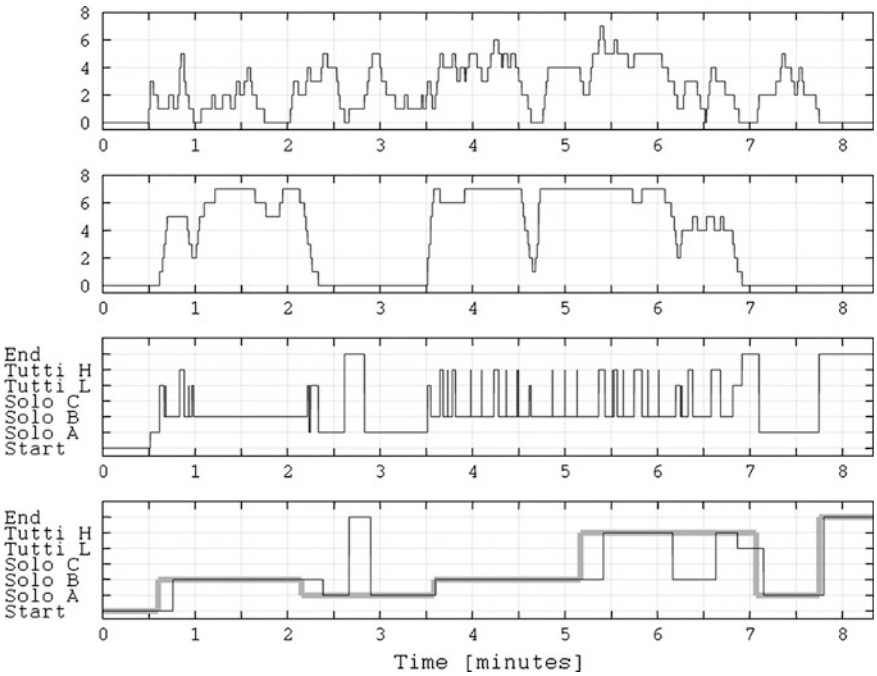


Fig. 5 Ensemble state example for a μ·cos*m* trio session. *Top graph* Tension curve for the saxophone (with variables $a = 1.2$ and $b = 0.6$); *2nd graph from Top* tension curve for the moog synthesizer ($a = 1.2$, $b = 0.4$); *3rd graph from top* caira's short term ensemble state estimations; bottom graph: Caira's final (long-term) ensemble state estimations (*solid think black line*) versus human (my own) ensemble state estimations (*solid thick gray line*)

archival purposes, so the listener can distinguish between the three musical streams produced by Caira, the synthesizer, and the saxophone.

A foot-controller was the user interface, so I could operate the synthesizer simultaneously with my saxophone. The foot controller consists of a MIDI foot controller (Behringer, FCB 1010), which can be used to play notes as well as control MIDI data and has two pedals for continuous data control. In addition, a trackball interface (Logitech T-BC21) was also operated by foot, which enabled me to repatch and control every aspect of the synthesizer in real time. The trackball buttons were handled with the other, right foot using a modified foot-switch from a dictation machine (Grundig, 526A foot control) that was equipped with electronics from a USB mouse.

My new experimental trio now consists of soprano saxophone, the Arturia Moog Synthesizer, and the agent Caira, who currently performs based on V-Accordion recordings by Pauline Oliveros. Since I was not interested in using the μ·cos*m* approach as a substitute for our *Triple Point* work, but rather to have this trio for additional tests, the tradeoff between ensemble realism and flexibility is acceptable.

7 Examining the Self-Exploratory Approach

One questions the legitimacy of self-evaluating an intelligent music improvising agent. While the benefit of an introspective approach is clearly that one can prototype and evaluate the system within a very short cycle, most agreed-on psychophysical test paradigms build on testing a larger number of human subjects. The reasons for this are manifold. First, we typically want to obtain data that is valid across a large population. We may, for example, is the measurement and standardization of the absolute threshold of hearing curves, where we want to know what the average threshold of hearing curves are so that we may use these data as the reference for hearing screening tests. For the current standard for normal equal-loudness-level contours [ISO 226:2003, see also Suzuki et al. (2003)], over a hundred normal-hearing subjects were tested in twelve studies. In contrast, *Experimental Music* is not a mainstream culture, where one would seek for data across a large population. The genre is practiced by a very small group of people, who believe their music contrasts with popular culture. Instead of understanding the ability of a large portion of the population, we are more interested in the general possibility of a particular ability, even if it can only be demonstrated through one subject or a handful of subjects.

The second problem is to avoid skewed data. In order to collect unbiased data, we often prefer to use a double-blind test paradigm, for example by setting up a computer-controlled psychophysical experiment. The advantage of this approach is that it can be suitable for introspective research, where the experimenter is serving as his or her own test subject. However for an introspective test paradigm, this technique only works on a subset of research questions that do not affect preference or background knowledge. For example, if I am testing my own system, I am not in a neutral position to provide an unbiased preference rating. Neither will I be able to carry out a Turing test, since I have a priori knowledge that I will be performing music with an intelligent machine and not a human person. I should note that I generally question the importance of the Turing Test in this context, because I am primarily interested in communicating with an intelligent being who inspires me, and puts me in a creative environment. I care less about whether this inspiration is human and more about finding a source that inspires me to do things that I have not thought of doing before. A non-human performance partner might even be beneficial because in *Experimental Music* the goal is to stretch boundaries.

To my opinion, the demonstration of intelligent communication is the key for a successful agent implementation. In the subsequent paragraphs I will lay out how communication with an agent can be examined using a test paradigm that is open to introspective methods. In the Caira project, Caira's ability to communicate can be tested by determining the degree of agreement of the identified ensemble states between Caira and the human performer(s) over time. In the trio scenario, we will have three ensemble state curves over time, one for each performer. In the μ·cos*m* project, of course, we have only two curves one for Caira and the other one for myself, because I perform two of the ensemble roles, saxophone and

synthesizer, simultaneously. Now we can look at events and see if they match in general. We can also determine which of the states are more often confused with each other and investigate the temporal mismatch between the agent and human performers when entering a new state. Figure 6 shows an example of a session that I carried out with Caira. The two top graphs show the estimated tensions curves for this session for the saxophone and the Moog Synthesizer. Based on these tension curves, Caira estimated the short-term ensemble states (3rd graph from top) as well as the final (long-term) ensemble states (bottom graph, solid thin black line). The bottom graph also shows my own ensemble state estimates for comparison purposes. I should emphasize that it is crucial that both participants do not know of each other's scores. In most cases, the two ensemble states (Caira's and my own) are in agreement with a few noticeable exceptions. At approximately 2:40 min, Caira believes the session has ended, because both observed instruments remain silent for a brief period, while I do not intend to terminate the piece at this point. In the time segment between 5 and 7 minutes, Caira fluctuates between the *Solo B* and the *High-Tension Tutti* states, where my own judgment indicates a solid *High-Tension Tutti* block. Please also note that the judgment of Caira is typically trailing my own, since the agent requires time to analyze my intentions.

It would be fair to argue that demonstrating successful communication does not guarantee an excellent music performance. However, communication is the key for a successful performance, and lack of communication is usually a good indicator of failed ensemble work. It is also important to keep in mind that we need to start somewhere with the goal in mind to develop adequate methods for the quality judgment of an experimental music session or any other type of music. Further, it should be pointed out that other, more traditional research fields also accept indirect measures. In the field of sound source segregation, for example, we show

Fig. 6 *Left Photo* Technical set-up of the μ·cos*m* trio. *The left computer* hosts the arturia moog synthesizer, *the right computer* the caira agent. *Right Photo* Close-up of the foot controller in operation

the improvement of removing an unwanted sound source in terms of the signal-to-noise ratio (SNR), which does not account of how the desired signal is distorted. For example, if speech is the desired signal, it can be harder to understand even though the SNR improvement predicts a positive outcome, because the masking-signal removal process introduces distortions that degrade important speech cues [e.g., compare Roman et al. (2006)]. The reason why we often do not measure perceptual improvement directly is that we do not have auditory models in place to execute an automated assessment in lieu of a psychophysical experiment.

Interestingly, the small data challenge that we face using an introspective approach is a familiar problem in musicology, where we often tend to understand historical developments through the statements of a few witnesses from that time. While, for the most popular composers, for example, Mozart, we have enough records from his contemporaries to understand how his work was received during his lifetime, in the case of less known composers the margin of error is much larger because we can only draw from the comments from a very few time witnesses. I encountered this problem during my work on Georg Joseph Vogler (Braasch 2004). The approach is valid, however, because it is still the best available method to understand past developments given our lack of a time machine, which would enable us to interview a larger sample population about their preference of a given musical event or composer's work. While musicologists are trained as neutral observers who assess the situation without interfering with it, this paradigm has often not proven useful in ethnomusicological research. Without interference, one could write a very convincing theory, but methods to test if the theory is correct are very limited. The need for better techniques has then lead to establishing the field of experimental ethnomusicology. Speaking in practical terms, the traditional ethnomusicologist would visit a region, listen and record the music, interviewing native people who are familiar with the music tradition under investigation. Based on the recorded data the expert then develops a model and publishes it. In contrast, the experimental ethnomusicologist takes his or her theory back to the studied community to test it. An example of the latter approach is the work of Arom (Arom 1967; Arom et al. 1997, 2007). Arom studied the music of a local African tribe. After he had developed a model of their music practice, he played along with the tribal musicians on his French horn. Arom updated and refined his model using an iterative process, based on the local musicians' feedback.

The evaluation process for the μ·cos*m* project can be seen as an extension of the experimental ethnomusicology approach. Here, the agent Caira takes the role of the tribe that is being investigated, and is probed by the experimenter to examine her response within a given theory. A major difference from the experimental ethnomusicology case is that we know the underlying theory–because we developed it. In the Caira project, we are merely testing whether Caira responds according to the developed theory. In the future we hope the agent evolves its own theories through concept building; then our approach will close the gap to experimental ethnomusicology.

Finally, I would also like to discuss the experimental character of the music used in this research. In experimental music we try to stretch the boundaries of

what is currently possible and acceptable in music. So, do I expect that my documented communication with Caira is generalizable to other human musicians? Probably not right away, but by demonstrating that the agent is able to communicate using a specific protocol, other musicians could learn how to perform with the music agent by learning the rules. This is in line with traditional forms of music where musicians must also learn the specific conventions of the genre to be able to play along. Further, our object of observation, the human mind, is a time-variant system unlike the physical world, where laws are assumed to be time-invariant. An example is the switch in general judgment of a sixth being a dissonant interval in the Middle Ages [e.g., Gustav Jacobsthal 1873, c. 642], while it is being considered as a consonant interval today. In the culture of AI, frequently the question comes up to whether a demo is staged or not. While I argue that this is not the case for the Caira project, I definitely had to adjust my performance so my sonic gestures were understandable by the agent. It is normal practice in music, however, that one should be able to clearly articulate musical form and adjust one's style to be recognizable to the other musicians, whether human or synthetic characters. A classic example is the practice of ending patterns, so the ensemble can easily pick up cues by the soloist to perform a rehearsed ending ad hoc. If done well, it will sound freely improvised, although a lot of work has gone into making this happen.

In closing, I would like to encourage the reader to visit our Caira project webpage, where a selection of recordings with Caira can be found demonstrating the agent's current state of musicianship: http://www.jonasbraasch.com/Caira.html.

Acknowledgments This material is based upon work supported by the National Science Foundation under Grant No. 1002851. The real-time implementation of the Caira system was written in Max/MSP utilizing various custom externals and abstractions as well as the FTM, Gabor and MnM packages from IRCAM, externals from CNMAT and Tristan Jehan's toolboxes (also using their loudness and roughness algorithms for a single-machine, stand-alone version of Caira).

References

Arom, S. (1967). The use of play-back techniques in the study of oral polyphonies. *Ethnomusicology, 20*(3), 483–519.

Arom, S., Léothaud, G., & Voisin, F. (1997). Experimental ethnomusicology: An interactive approach to the study of musical scales. In I. Deliège & J. Sloboda (Eds.), *Perception and cognition of music* (pp. 3–30). Hove: Taylor and Francis.

Arom, S., Fernando, P.N. & Marandola, P.F. (2007). An *innovative method for the study of african musical scales: Cognitive and technical aspects*: *Proceedings of the 4th Sound and Music Computing Conference SMC'07*, *Greece* (pp. pp. 107–116) July 11–13 2007.

Braasch, J. (2004). *Über die Verbreitung der Durchschlagzunge durch Georg Joseph Vogler [On the popularization of free reeds by Georg Joseph Vogler],* in: Orgelregister mit Durchschlagzungen. Geschichte, Konstruktion und akustische Eigenschaften [Organ stops with free reeds. History, construction and acoustical properties] (pp. 29–64), Berlin.

Braasch, J., Bringsjord, S., Kuebler, C., Oliveros, P., Parks, A. & Van Nort, D. (2011). *Caira—A creative artificially-intuitive and reasoning agent as conductor of telematic music improvisations*: *Proceedingsof 131th Audio Engineering Society Convention*, October 20–23, 2011, New York, NY, Paper Number 8546.

Braasch, J., Peters, N., Van Nort, D., Oliveros, P. & Chafe, C. (2011). A *spatial display for telematic music performances*. In Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera & H. Kato (Eds.), *Principles and applications of spatial hearing*: *Proceedings of the 1st International Workshop on IWPASH* (pp. 436–451). Singapore: World Scientific Pub Co Inc, ISBN: 9814313874.

Braasch, J., Van Nort, D., Oliveros, P., Bringsjord, S., Sundar Govindarajulu, N., Kuebler, C. & Parks, A. (2012). *A creative artificially-intuitive and reasoning agent in the context of live music improvisation*.: Music, Mind, and Invention Workshop: Creativity at the Intersection of Music and Computation, March 30–31 2012, The College of New Jersey, URL: http://www.tcnj.edu/ mmi/proceedings.html, last accessed: August 10, 2012.

Chalupper, J., & Fastl, H. (2002). Dynamic loudness model (DLM) for normal and hearing-impaired listeners. *Acta Acustica united with Acustica, 88*, 378–386.

de Cheveigné, A. & Kawahara, H. (2002a). *Matlab toolbox: YIN, a fundamental frequency estimator for speech and music*. URL: http://audition.ens.fr/adc/sw/, last accessed, January 10 2012.

de Cheveigné, A., & Kawahara, H. (2002b). Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111*, 1917–1930.

Cope, D. (1987). An expert system for computer-assisted composition. *Computer Music Journal, 11*(4), 30–46.

Daniel, P., & Weber R. (1997). Psychoacoustical roughness: Implementation of an optimized model. *Acustica*, *83*, 113–123

Dubnov, S. (2003). *Non-gaussian source-filter and independent components generalizations of spectral flatness measure*: *Proceedings of the International Conference on Independent Components Analysis (ICA2003)*, *Porto* (pp. 143–148).

Dubnov, S., McAdams, S., & Reynolds, R. (2006). Structural and affective aspects of music from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology, 57*(11), 1526–1536.

Ellis, D. P. W. (1996). *Prediction-driven computational auditory scene analysis*. Massachusetts Institute of Technology: Doctoral Dissertation.

Friberg, A. (1991). Generative rules for music performance: A formal description of a rule system. *Computer Music Journal, 15*(2), 56–71.

Jacob, B. (1996). Algorithmic composition as a model of creativity. *Organised Sound, 1*(3), 157–165.

Jacobsthal G. (1873). *Die Anfänge des mehrstimmigen Gesanges im Mittelalter [The beginnings of polyphonic chant in the Middle Ages]* (vol 41 pp. 641–646). Leipzig: Allgemeine Musikalische Zeitung.

Lewis, G. E. (2000). Too many notes: Computers, complexity and culture in voyager. *Leonardo Music Journal, 10*, 33–39.

McAdams, S., Smith, B.K., Vieillard, S., Bigand, E., & Reynolds, R. (2002). Real-time perception of a contemporary musical work in a live concert setting. In C. Stevens, D. Burnham, G. McPherson, E. Schubert & J. Renwick (Eds.), *Proceedings of the 7th International Conference on Music Perception and Cognition*, Sydney, Australia

Van Nort, D., Braasch, J. & Oliveros, P. (2009). *A system for musical improvisation combining sonic gesture recognition and genetic algorithms*: *Proceedings of the SMC 2009-6th Sound and Music Computing Conference*, *Portugal* (pp. 131–136) July 23–25 2009.

Van Nort, D., Oliveros, P. & Braasch, J. (2010). *Developing systems for improvisation based on listening*: *Proceedings of the 2010 International Computer Music Conference (ICMC 2010)*, *New York*, June 1–5 2010.

Van Nort, D., Braasch, J. & Oliveros, P. (2012). *Mapping to musical actions in the FILTER system*: *The 12st International Conference on New Interfaces for Musical Expression (NIME)*, May 21–23, Ann Arbor, Michigan.

Pachet, F. (2004). Beyond the cybernetic jam fantasy: The continuator. *IEEE Computer Graphics and Applications, 24*(1), 31–35.

Roman, N., Srinivasan, S., & Wang, D. L. (2006). Binaural segregation in multisource reverberant environments. *Journal of the Acoustical Society of America, 120*, 4040–4051.

Suzuki, Y., Mellert, V., Richter U., Møller, H., Nielsen, L., Hellman, R., Ashihara, K., Ozawa, K. & Takeshima, H. (2003). *Precise and full-range determination of two-dimensional equal loudness contours*, Technical Report, Tohoku University, Sendai.

Widmer, G. (1994). *The synergy of music theory and AI: Learning multi-level expressive interpretation*, Technical Report OEFAI-94-06, Austrian Research Institute for Artificial Intelligence.

Zwicker, E., & Fastl, H. (1999). Psychoacoustics: Facts and models, 2nd edn. Berlin: Springer.

Acoustical Measurements on Experimental Violins in the Hanneforth Collection

Robert Mores

1 Some Basics of Violin Acoustics

There has been extended research on violin acoustics over the last century comprising several thousands of individual studies on string motion, specific components, body acoustics and radiation. The interested reader might begin with a comprehensive overview (Fletcher and Rossing 1998; Rossing 2010) or with Hutchins' fine selection of research papers (1997) before going into depth studies of violin physics (Cremer 1981). Therefore, this section will not add to existing knowledge but rather provide a tutorial for the first time reader. From the broad knowledge those basic principles are outlined, which luthiers are likely to be fully aware of, and, at the same time, which are key to understanding constructive change and its impact to sound.

For the purpose of amplifying string motion and of radiating sound, most stringed instruments employ resonating air, vibrating plates, and a body.

1.1 Vibrating Plates

The frequencies of vibration modes on a plate are strongly destined by length L, width W and strength (height) h of the plate. In the case of using wood for the plate, the two direction-dependant sound velocities c_l and c_w are essential for the potential modes (Fletcher and Rossing 1997, p. 89):

$$f_{mn} = 0.453 \cdot h \cdot \left[c_l \cdot \left(\frac{m+1}{L} \right)^2 + c_w \cdot \left(\frac{n+1}{W} \right)^2 \right] \quad (1)$$

R. Mores (✉)
Faculty of Design, Media & Information, University of Applied Sciences, Finkenau 35, 22081 Hamburg, Germany
e-mail: Robert.Mores@haw-hamburg.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_12,

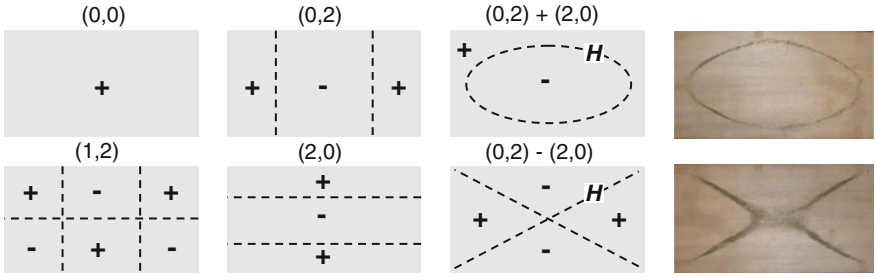


Fig. 1 Modes in hinged or free plates [mode (0,0) only in hinged plates], *left* the abstract notation for prominent modes including the superposition ring mode and X-mode, *right* Chladni patterns of superpositioned modes established on a 3 mm thick plate of Sitka spruce with a length-width-ratio of about 1.9

First of all there are modes, or standing waves, of order m possible across the length of the plate or of order n possible across the width of the plate. The respective frequencies, however, are always constituted by both sound velocities. The mutual interdependence can be explained by material strain properties: stretching or compressing the material into one direction will cause stretching or compression into other directions, expressed by six Poisson modules (General Technical Report 1999) that co-determine sound velocity, together with Young's modulus.

Obviously, the plate strength h is a forceful parameter. Variations of a few tenths of a millimetre on a plate of typically a few millimetres will change frequencies of all modes by some ten percent. Disconcertingly, the strength of top plates usually varies strongly, not only from violin to violin, but sometimes even in a range of 2–5 mm for the same violin top plate. Therefore, the given analytical approach will fail to predict mode frequencies.

Among the prominent modes and important to luthiers are the (2,0)- and (0,2)-modes[1] and their superpositions, see Fig. 1. The superpositions are called ring mode and X-mode and are only possible, if the frequencies of the underlying (2,0)- and (0,2)-modes are close to each other. In plates of homogenous material, identical frequencies would be given for a square plate. In wooden plates, the direction-dependant sound velocities require a rectangular shape of the plate to allow frequencies of different modes to approach each other. For Sitka spruce, the material properties suggest a length–width-ratio of $L/W = 1.9$. This ratio can be considered as a determining factor for the evolution of string instruments. Violin makers routinely listen to the ring mode and X-mode while they hand craft a violin top or back. They usually hold the plate at nodal lines, indicated by the holding point H in Fig. 1.

Superposition of modes does not mean that modes would coexist independently in one and the same system and can simply be added. Vibrations are rather

[1] Notation: a (x,y)-mode has x nodes across direction x and y nodes across direction y.

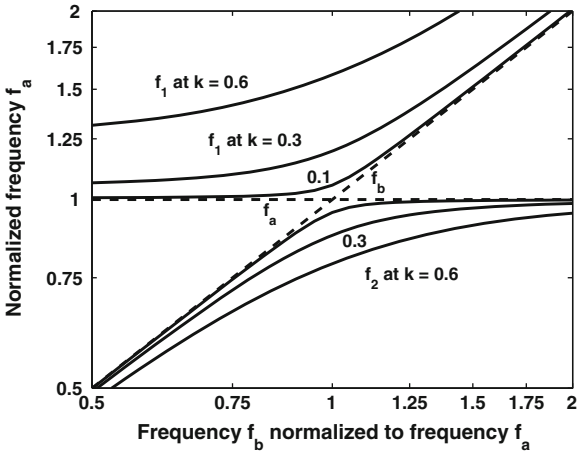


Fig. 2 When two systems with resonance frequencies f_a and f_b are coupled to each other, the resulting resonance frequencies f_1 and f_2 will deviate depending on how strongly the systems are coupled to each other, expressed by factor k

mutually coupled within a physical system. When two resonating systems with frequencies f_a and f_b are coupled to each other, the respective resonance curves[2] will not simply be added, eventually leading to an excessive increase. Vibrations will rather interact mutually and the resonance frequencies are likely to even shift. The coupled system will resonate on frequencies f_1 and f_2, see Fig. 2. These frequencies are derived from solving

$$\left(f^2 - f_a^2\right) \cdot \left(f^2 - f_b^2\right) = f^4 \cdot k^2 \tag{2}$$

for systems coupled by factor k. Such coupling is given in a plate, and the Poisson modules qualitatively translate to k.

For the practical example of Fig. 1, the spruce plate ($L = 39.4$ cm, $W = 21$ cm, $h = 3$ mm) developed the X-mode at $f_2 = 117$ Hz, the ring mode at $f_1 = 128$ Hz and the (2,0)-mode was measured at $f_b = 120$ Hz. The effect of these coupled resonance frequencies is also reported for violin tops by Hutchins (1981) and Molin et al. (1988), referencing to mode 2 and mode 5 for the X-mode and the ring mode, respectively. These modes are reported to be favourable in violin acoustics and luthiers usually report a beautiful ringing of the "tap-tone" for well tuned plates. However, frequencies of these modes are not as adjacent for violin plates as they are in the free plate example above. Frequencies are even recommended to be spaced by one octave (Hutchins 1994) to achieve a full and rich sound. Jansson et al. (1988) reports the crafting process from a plate to a violin top and its accompanying tuning steps. In conclusion, plate tuning is an essential skill for the successful violin maker.

One more aspect is relevant to be prepared for the investigation of Hanneforth's collection: plate resonance frequencies are shifted upwards, when plates are

[2] The resonance curve describes spectral properties of the resonator: position on a frequency scale, and width of the curve as a reciprocal quality measure.

bended. Bending is usually not practiced in violin making, but is quite common for guitar backs. Jahnel (1981) reports an increased intensity for frequencies above 500 Hz in guitars.

1.2 Vibrating Body

For the completed violin body, there are finally several hundred modes that constitute the so-called resonance profile. More than 30 modes below 1,300 Hz can be identified by their nature and assigned to different classes (Marshall, 1985). For the purpose of this investigation we roughly consider (1) coupling between the top and the back plate, (2) two-dimensional body modes, and (3) one-dimensional body modes.

Top and back of the violin are brought together with their individual set of modes. Clearly, the two plates are coupled to each other by the ribs and the trapped air. In the same way, the frequencies of the (2,0)- and (0,2)-mode are suggested to be brought together to form ringing superpositions for the unattached plate, the frequencies of these superpositions are suggested to be brought together for the attached top and back plates to achieve a full and rich sound in a completed instrument (Hutchins 1994). Again, the tuning process is of importance, not only for individual components but also for mutual coupling between components.

The rigid body as such with its given height can be considered as a plate as well, supporting two-dimensional modes. The waist enhances mobility and facilitates torsional modes. Gough (2010) has investigated the principles of what he calls the strongly radiating "signature" modes in the range of 400–550 Hz. Essential to these modes are the breathing and the bending of the violin corpus, see Fig. 3 on the left side. Again, as the frequencies of these two modes approach each other there will be superpositions following the coupling paradigm. The resulting pair of body modes is often referred to as *B1* modes, see Fig. 3 on the right side. And the frequency of these superposition modes "will never meet" as Gough

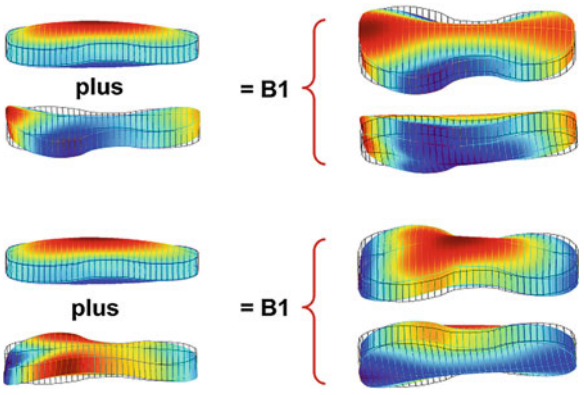


Fig. 3 Superpositions of breathing mode and bending mode in a violin body. *Left top* breathing added to bending; *left bottom* breathing added to bending of opposite phase; *right side* resulting body modes (two view angles on the body)

emphasizes this behaviour as a physicist, confirming the argument above. This example demonstrates that coincidence of mode frequencies and the combination of modes results in stronger vibration since two vibrations add together for a larger displacement. The nature of mode coupling, however, will not transform two peaks into one stronger peak, but into two stronger peaks.

While the breathing mode is unlikely to vary with rib strength, the bending mode is very likely to depend on body stiffness contributed by rib strength. Therefore, without changing the plates, the tuning of frequencies can be achieved by the rib strength. Depending on the tuning, either superposition mode will be of slightly higher frequency (often referred to as *B1+*, equivalent to f_1 in Fig. 2), and the other superposition mode will be of slightly lower frequency (often referred to as *B1−*, equivalent to f_2 in Fig. 2). Usually, these two modes are likely to be positioned somewhat above the plate modes.

Yet another prominent body mode is of lower frequency but not as strongly radiating as the *B1* modes. This body mode is of type (1,2), see Fig. 1, and is often denoted as *CBR* mode. Gough's models support mutual interaction not only between breathing and bending of the body, but likewise between *CRB* and bending modes.

An example of a one-dimensional body mode is the bending of the entire instrument, say the violin corpus on one side and the neck with pegbox and scroll on the other side vibrating against each other. There are usually different modes due to the different directions of displacement. These modes are usually located at about 200 Hz in terms of frequency (Marshall 1985, p. 703), usually denoted as the first body mode *C1*. Another noticeable mode results from bending of the fingerboard against the body, with the neck serving as spring. Many of these modes are visualized by means of an early finite element model of the violin (Knott 1987).

1.3 Resonating Air

Vibration of the air trapped in the violin body is usually referred to as the Helmholtz mode *A0* and cavity modes of higher order, *A1*, *A2*, and so on. The frequency of *A0* is defined by the volume V of the cavity, the area A of the *f*-holes for neckless cavities, the speed of sound c, (Bissinger 1992a, Eq. 2)[3]

$$f_0 = \frac{c}{2 \cdot \pi} \sqrt{\frac{A}{\left(h + a \cdot \sqrt{A}\right) \cdot V}} \tag{3}$$

[3] For the neckless cavity, the length L of the neck corresponds to the strength of the plate. Therefore, strength h is used in this equation. Equation 3 already implies an end correction term and holds for circular apertures. For *f*-holes the frequency can increase by as much as 60 %.

where a is determined to $a = 0.959$. Frequency f_0 is located between 260 and 290 Hz for most of the violins. *A0* is of particular interest, since it amplifies tones around C4/C#4 and composition relies on the existence of such strong and full tone on a violin in terms of dramaturgy. Another important aspect of Helmholtz resonance is its contribution to a full tone. Even when notes of a pitch higher than f_0 are played, the bow-string-friction will always contain enough noise to stimulate *A0*, effectively enhancing tone colour across the entire scale.

Investigations reveal little impact of the shape of the *f*-holes to cavity frequencies. The impact of variations of *f*-hole positions is negligible for *A0* and lower for *A1* than for modes *A2*, *A4* and *A5* (Bissinger 1992a, p. 15). Frequencies for *A1*, *A2*, *A4* and *A5* are roughly located at 500, 1,100, 1,300 and 1,600 Hz, respectively, and vary only little with variations of cavity height (Bissinger 1992b, p. 20).

1.4 Combinations of Modes

Apart from the mode superposition already outlined for the plate and for the body modes, there are several mode combinations that are relevant for sound, projection, and playability.

Helmholtz mode *A0* and body mode *B0*—Hutchins denotes the bending of the whole instrument to mode *B0*, and reports the frequency of *B0* to be sometimes higher, sometimes lower than the frequency f_0 of *A0*, but close to each other in most instruments (1985). She also reports several hundred successful adjustments of *B0* frequency by altering stiffness of fingerboard and mass of chinrest and pegs. Coincidence of frequencies will again result in what has led to the "beautiful ringing" of plates. Here, Helmholtz resonance is enhanced by a tuning instance most violin makers are aware of.

Cavity mode *A1* and body mode *B1*—The spacing between the frequency of the cavity mode *A1* and the frequency of the body mode *B1* (that is usually split into two peaks, see above) relates to what musicians and owners report in terms of tone quality and playing qualities: a spacing of over 100 Hz results in a harsh sound and the violin is practically unplayable, a spacing of 70–80 Hz delivers a bright edgy tone to skilled soloists, 50–70 Hz bring about a fine tone and projection for soloists and concertmasters, 40–50 Hz result in a good tone and projection, 20–30 Hz gentle tone for chamber music and easy playing, below 20 Hz a gentle sound with little power (Hutchins and Benade 1997, p. 663).

Other cavity and body modes are also present with the violin. There are more pairs of coinciding modes that potentially improve sound quality when properly tuned. Some of these even imply tuning of the tailpiece (Hutchins and Voskuil 1993; Zopf 2000). These other options seem to be less familiar among luthiers.

1.5 Sound Post and Bridge

Sound post and bridge are very relevant to sound. The sound post's main task is to introduce asymmetric impedances at the bridge feet, but it has also a secondary effect of strongly enhancing high frequency components. This can be explained by the small (few mm) space between the node introduced by the sound post and the exiting bridge. Even very tiny alterations will strongly influence the high frequency components embodied in the timbre. Such tiny alterations will—at the same time—influence the general plate modes only marginally (Saldner et al. 1996) which can be explained by the fact that the sound post position is located on modal nodes for most of the plate modes. This is well visible in the holographic interferograms of Jansson et al. (1994). In conclusion, the sound post will change only little for plate, body and cavity modes below 1,500 Hz. The experimental constructions of the violins in the Hanneforth collection, however, will change a lot of what has been outlined so far in terms of plate, body, and cavity.

Similarly, the bridge has a main task—transforming string motion to rocking as well as to up-down movements, depending on frequency (Reinecke 1973)—but also secondary tasks of filtering high frequency noise (above 6 kHz) and directly radiating sound (in the range of several kHz). Sound adjustments on bridges will likewise bring no change to the fundamental plate and body motion. Therefore, the bridge as well is not of interest in this investigation, the experimental violins do not focus on these two components.

2 Measurements and Violin References

2.1 Measurements

Three types of measurement are used in this investigation: (1) input admittance at the bridge, (2) radiated sound in two dimensions, (3) sample of radiated sound at one point under reverberant condition.

For input admittance measurements, the impulse response is measured by means of an accelerometer sensor[4] attached to the bridge, while the impulse is generated by an impulse hammer.[5] Force direction and direction of displacement measurement coincide with the plane of the bridge and strictly follow the direction of rocking motion. The impulse introduced by the hammer is measured as well, so that from the impulse and its response[6] the transfer function can be derived in the

[4] Kistler sensor 8778A500.

[5] DYTRAN impulse hammer 5800SL, 2 g.

[6] Measurements are normalized to each other and calibrated by use of a reference mass.

frequency domain. This transfer function represents the mechanical admittance[7] *Y*, or, in other words the mobility of the bridge. Peaks in admittance plots represent high mobility of the investigated structure, which not always directly translates to sound radiation. This is a familiar measurement method widely used for structure borne sound analysis and in engineering.

The two dimensional measurement of radiated sound is done according to a set-up of Schleske (2002) on the transversal plane around the violin, at the height of the bridge and at a resolution of 10°. Room characteristics are averaged across directions by repeating all measurements while the entire setup including violin and exciter is rotated stepwise within the room. The result is a direction-dependant sound level for that plane. Illustrations in this chapter use only the total sound level averaged across all directions.

For measuring sound radiated from instruments of the Hanneforth collection, only a single microphone was used, perpendicular to the violin top, above the bridge at a 50 cm distance, in a reverberant room ($T_{60} \approx 2$ s). The exceptional Stroh violin was recorded with a microphone in front of the horn.

2.2 A Fine Reference: The 1712 "Schreiber" Stradivari

Figure 4 shows the bridge mobility as well as the averaged sound level for the 1712 "Schreiber" Stradivari.

The bridge admittance Y illustrates all the principles outlined so far[8]: (1) the bending of the violin, denoted *C1* at 200 Hz, (2) the Helmholtz mode *A0* at 265 Hz, (3) the range of signature modes *CBR* and *B1* at 400–530 Hz with (4) the cavity mode *A1* in between, (5) a spacing of some 50 Hz between *B1*- and *A1*, (6) plate modes from 700 to 2,000 Hz (Moral and Jansson 1982, p. 332).

Observations on comparing structural admittance in relation to radiated sound are: (1) the mobility does not translate to radiation one by one, which is well known, (2) the above mentioned coincidence of *A0* and *B0* modes is observable in form of a double peak in the radiation plot, but not in the mobility plot, (3) CBR does not radiate as prominent as it appears in the mobility plot (Rossing 2010, p. 222).

The "Schreiber" Stradivari is considered as an exceptionally good example of a soloist violin, and it might appear wise to use its mobility plot as reference together with the outlined principles above as supported by research:

Principle (a)—*A0* must be in a certain frequency range to meet the musicians' expectations in the context of existing compositions. *A0* must also be sufficiently strong, violin makers usually strive for high levels here. Bissinger (2008) would even distinguish among poor and good instruments along the *A0* level.

[7] Mechanical admittance *Y* is the mechanical impedance *Z* inversed.

[8] There exist several styles of denoting modes, refer to Rossing for an overview (2010, p. 222).

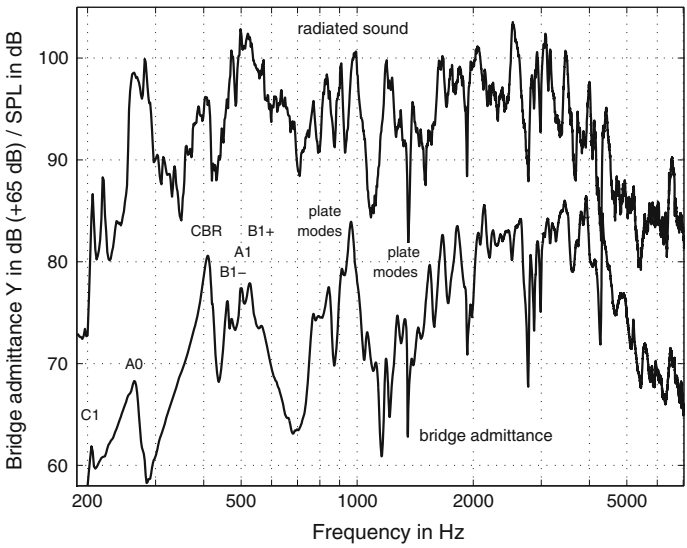


Fig. 4 Bridge admittance *Y* and sound pressure level SPL for the 1712 "Schreiber" Stradivari

Principle (b)—The range of body modes/signature modes appears to be different from violin to violin. However, in this fine instrument the target of a wide range of resonating frequencies is well identifiable.

Principle (c)—The sound level across the entire spectrum is balanced in a way as to contain both body and brilliance. From their field studies on valued instruments Meinel (1937 and 1939) and Dünnwald (1982, 1984 and 1991) identified desirable proportions of energy in defined bands. The Dünnwald findings suggest high levels in a band "fullness of sound" from 190 to 650 Hz as well as in a band "brilliance and clarity" from 1,300 to 4,200 Hz and lower levels in the band in between these two bands and above 4,200 Hz. The Stradivari represents this and its mobility plot might serve as reference.

Principle (d)—Most of the discussions on mobility or resonance focus on prominent peaks, promising an appreciated sound level. There is little said about spectral zeros, or dips. Most of the violins, however, even student level instruments reveal a strong dip in the area of 700–800 Hz. Such a dip is reported to be desirable and seems to be of importance for perceiving a good quality tone.[9] This emphasizes but also particularizes issue (c).

This set of principles would certainly be extended by issues of brilliance and timbre and some more when high quality instruments are to be evaluated. For the experimental violins discussed here this set will basically cover the issues.

[9] Personal communication with luthier Martin Schleske.

2.3 A Moderate Reference: A 1900 Markneukirchen Student Level Violin

The violin investigated is an ordinary student level violin built around 1900 in Markneukirchen, and it is not part of the collection. It is a fine piece of workmanship, very easy to play, nicely merging in chamber music, but it has practically no projection and a gentle rather than a full sound.

The reason for this choice is to relate ordinary violin making and the too perfect Stradivari reference so as to understand the scope of changes that "making copies" would bring about, without the idea of undertaking experiments. A scale from moderate to perfect seems to be a reasonable starting point to further relate and adequately discuss the changes experimental violins bring about.

Figure 5 illustrates the bridge mobility of both violins, the fine and the moderate reference. At a first glance the general character seems to match. A closer look at the outlined principles might explain the perceived sound while playing the moderate instrument: (a) The Helmholtz resonance *A0* is perfectly located, something that is practically always given, even with the cheapest factory violins. However, the level of *A0* is almost 6 dB lower than that of the Stradivari. (b) The signature modes are located as expected and have an acceptable level. However, the modes are spaced closely to each other and the tuning result of the violin maker delivers a small band for the signature modes, about half the width of the fine reference. This fact together with the poor *A0* mode explains the thin sound. And the narrow spacing also explains the easy playing. (c) The moderate violin is short of some 5 dB on average in the "brilliance and clarity" band when compared to the fine reference. However, the violin is still balanced, as the "fullness of sound" band is weak as well. (d) The characteristic dip is given, however, at the somewhat low frequency of 620 Hz.

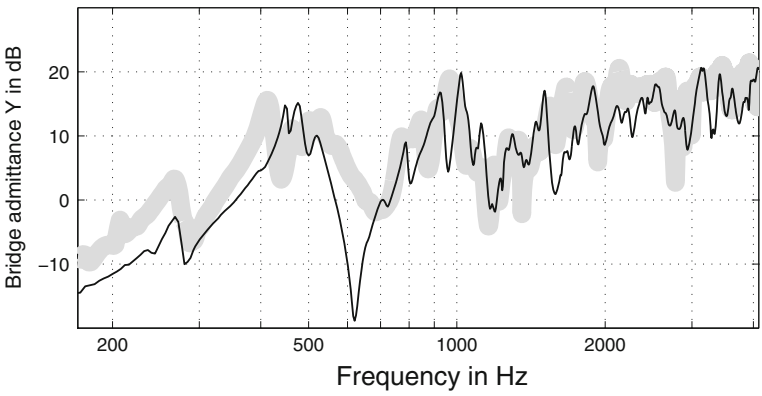


Fig. 5 Bridge admittance *Y* of the Markneukirchen student level violin (*thin, black*) in relation to the 1712 "Schreiber" Stradivari violin (*thick, gray*)

In conclusion, the structural sound analysis explains quite well what is perceived by a musician and therefore, translates to the value of an instrument.

3 Results for the Experimental Violins

3.1 1930 Bechstein/Moor Violin with Recessed Top Seam

Heinrich Bechstein (1865-?) and Emmanuel Moor (1863–1931) built this violin (serial number 30) about 1930 in Germany (MKG inventory number 2012.56). The choice of wood follows conventional preference for spruce and maple and the general proportions of the violin are conventional, but the construction of the top differs. The top seam is recessed effectively forming a fold in the middle of the top. As a consequence, the top is straight lengthwise and has a strong arch in lateral direction, see Fig. 6. Bechstein and Moor expected to achieve more tension in the back with this construction and to increase loudness. They patented the construction (P 517226, 1931) and also mandolins and guitars have been built following this principle.

It is not obvious how this construction would increase tension in the back, but it is clear the construction adds stiffness to the plate and improves stability. The deep fold in the middle can be compared with an increased plate strength h, but only in effect for one direction. Therefore, following Eq. 1, the (0,2)-mode will be strongly increased in frequency while the (2,0)-mode might remain unaffected. The frequencies of the two modes will never coincide to form the desired ring mode and X-mode.

Figure 7 again shows the bridge mobility referenced against the Stradivari violin. (a) Helmholtz mode *A0* is strongly diminished, by more than 8 dB. (b) The signature modes are still in place but strongly diminished. This is to be expected, because the back plate still contributes to the breathing mode of the body even though the top will not help much. And the bending component in the *B1* pair of modes is not primarily defined by the plane, but by the rib strength that contributes

Fig. 6 1930 Bechstein/Moor violin with recessed top seam

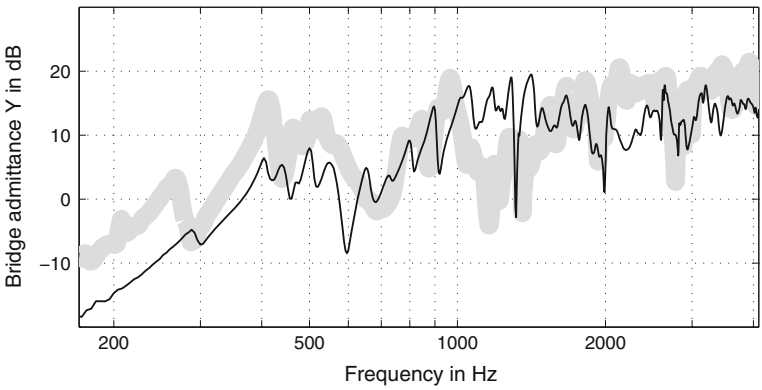


Fig. 7 Bridge admittance *Y* of the Bechstein/Moor violin (*thin, black*) in relation to the 1712 "Schreiber" Stradivari violin (*thick, gray*)

to body rigidness. Conventional rib strength will deliver a useful bending mode frequency, and the range of signature modes is fine here. However, the level is some 9 dB below the reference. (c) The "brilliance" band is 5–10 dB too low. (d) The dip, if any, is at the very low frequency of 600 Hz.

Another observation is that plate modes generally located at 800–1,000 Hz are now shifted towards 1,000–1,500 Hz, fully in agreement with Eq. 1 and the analogy between stiffness and plate strength. These plate modes unfortunately contribute to the level in a frequency band that should normally stay behind the "fullness" and the "brilliance" bands, but here it is the strongest band. This fact additionally opposes principle (c).

In conclusion, the violin will deliver neither a desirable tone nor an adequate projection. Categorical listening will conclude on a violin without any doubt, and even the properties of wood and other typical violin attributes will be clearly audible. However, the sound quality must be rated poor from a conservative point of view.

3.2 1820 Chanot Violin with Strings Knot to a Bridge on Top

François Chanot (1788–1825) designed this violin (MKG inventory number 2012.47), which was built around 1820 in Paris, see Fig. 8. From his engineering perspective the corner blocks would hamper mobility and therefore he shaped the violins like guitars. A patent for this was granted in 1818 and the violin has been preferred against a Stradivarius violin in blind tests in those days. The tonal quality is reported to have been diminished before long. This can be explained by the unusually high tension at the top that comes from mounting the strings to a terminating bridge at the top causing strain and deformation.

Fig. 8 1820 Chanot violin with strings knot to a bridge on *top*

Figure 9 illustrates the bridge mobility of the Chanot violin referenced against the Stradivari. Again, the general structure is similar due to common measures such as size or air volume. In terms of the outlined principles, there is one surprise: (a) *A0* is almost 4 dB stronger than in the Stradivari. This is surprising, because it is known from contemporary violin makers that it is very hard to ever come close to the old masterpieces. Outperforming the old masterpieces would only be possible with an increase of tension along the instrument above a limit of long-term stability. (b) Signature modes are fine in terms of frequency, level and spread. (c) The "brilliance" band is some 5–10 dB weaker than the reference, slightly out of balance when related to the level of the signature modes and the strong *A0*. (d) The dip is located at the rather low frequency of 560 Hz.

However, there is another problem and that is why this violin asks for a more detailed discussion than other experimental violins. Attaching the strings to the top rather than to the end block brings about some change to plate and body vibrations. A vibrating string will pull at its ends twice for every turn, or, period. Therefore, the top will be excited via the terminating bridge, where the string is attached to. This will have several effects. Helmholtz resonance is likely to be stronger as the

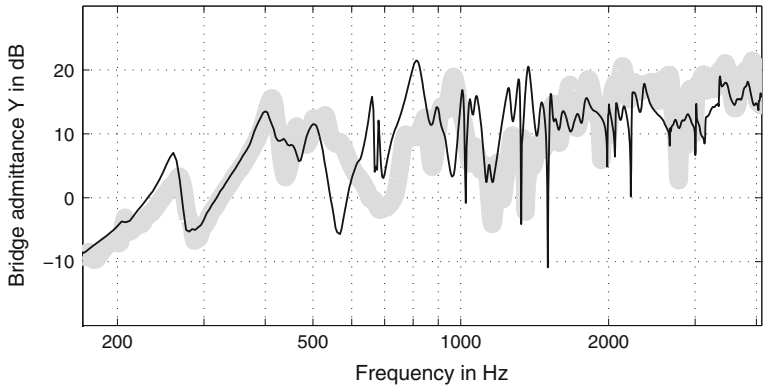


Fig. 9 Bridge admittance *Y* of the 1820 Chanot violin (*thin, black*) in relation to the 1712 "Schreiber" Stradivari violin (*thick, gray*)

string pulls at a most effective point for a breathing mode as confirmed by measurement. Given the two excitation points—conventional bridge *and* terminating bridge—and given the fixed distance between these two points, some plate modes are expected to be strongly amplified while others suffer from cancellation, depending on the phase.[10] Such zeros can be seen in the plot at 1,330 and 1,500 Hz. Similar to the situation of two loudspeakers in a room playing a broadband mono sound, the two excitation points here will generate some kind of periodical spectrum. For instance, the peaks at 404, 812, 1,248, and 1,620 Hz follow a rather strict integer arithmetical series with multiples of 405 Hz. Likewise the series of peaks at 504, 1,008, 1,524, and 2,004 Hz reveal regularity. It is not primarily the plate or the body structure that defines the mobility but it is the structure of the generating process that dominates the system while employing two excitation points.

In this context, the superelevations at 816 and at 1,376 Hz can be explained. Here the twin excitation principle matches plate modes. Imagine the plate longitudinally subdivided into three segments with two nodes in between. Contraction of a string—when a string's displacement is maximum—will push the bridge and therefore the middle segment down and it will pull the terminating bridge and therefore the bottom segment up. This works together for perfect amplification and is supported by Robert's FEM simulations (1986, page 10). Among 14 simulated modes of a hinged violin top there are just these two modes with two longitudinal nodes: a (1,2)-mode at 852 Hz and the (2,2)-mode at 1,399 Hz, perfectly confirming the observation.

And yet another observation is relevant to perceived sound quality. While some frequencies are favoured due to the excitation principle, many others are systematically suppressed. The mobility plot of the Chanot reveals significantly fewer peaks at all than the plot of other violins. This brings about some kind of singularity that is even reinforced by the periodicity of peaks. Here the "sizzle of a jagged frequency response" of a violin is missing, as Curtin would say (Rossing 2010, p. 215).

3.3 1836 Howell Violin with Bottle-Shaped Body

Thomas Howell, trading with musical instruments in Bristol, was concerned about the comfort of musicians when playing strings. The construction of the violin which was built in 1836 in Bristol (MKG inventory number 2012.74) shows two targeted measures. The body is shortened and the neck likewise lengthened to simplify the playing in high registers, and secondly, the body is widened and formed concavely rather than convexly at its lower end, to facilitate comfortable

[10] Systems theory claims that sampling a system twice will cause the spectrum to be weighted by a sin(x)/x function, causing periodicity and zeros.

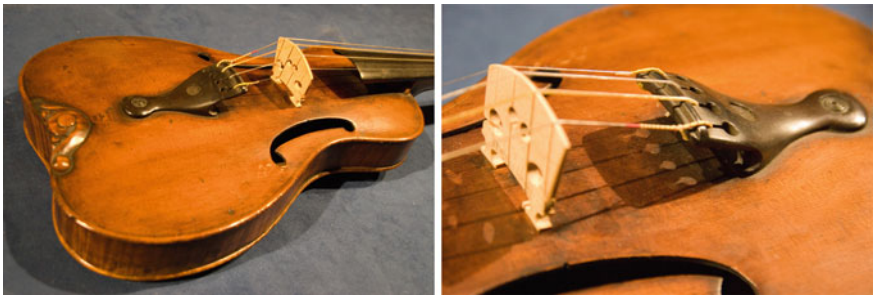


Fig. 10 1836 Howell violin with *bottle-shaped* body and strings attached to a terminating bridge

holding of the instrument, see Fig. 10. Comfort was also the idea behind introducing the terminating bridge on the top, to clear the space where the chin rests on the violin. His inventions where granted a patent in 1835 (Newton 1836). As there are constructive similarities with the Chanot violin, measurements might reveal similar effects to sound, too.

Figure 11 references the bridge mobility of the Howell violin against the Stradivari. (a) The Helmholtz mode *A0* is quite strong, 2 dB above the Stradivari. This is similar to the Chanot violin, caused by the same logic of strings pulling at the top at a most efficient point. The *A0* frequency is 308 Hz and hits note D#4 in an unexpected way. This increased frequency is due to a smaller air volume, see Eq. 3. Cavity modes should shift upwards due to reduced length of the body. (b) Signature modes are in place and occasionally even stronger than in the Stradivari. Both, breathing and bending are still supported by this construction and tuning of the bending mode by alterations of rib strength should still be possible. (c) The "brilliance" band is some 5–10 dB weaker than the reference, like in the Chanot violin, hampering the balance when related to the level of the occasionally strong

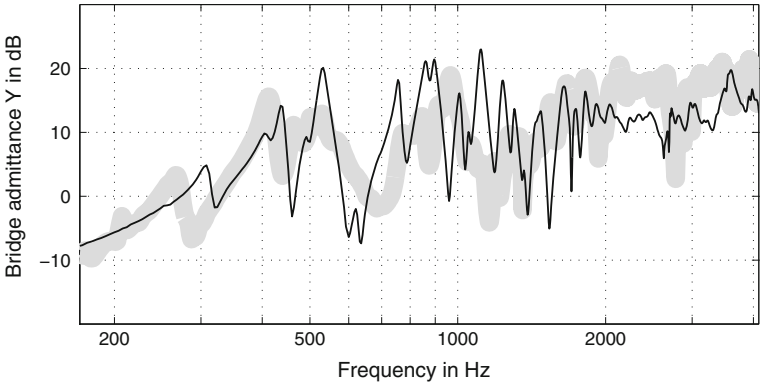


Fig. 11 Bridge admittance *Y* of the 1836 Howell violin with *bottle-shaped* body (*thin, black*) in relation to the 1712 "Schreiber" Stradivari violin (*thick, gray*)

signature modes and to *A0*. (d) The dip is strong, −30 dB, at the rather low frequency of 600 Hz.

Like in the Chanot violin, the principle of the twin excitation rules the system rather than plate or body structure. While in the Chanot violin the proportions and the positions of the two excitation points suggested a three-segment system with two nodal lines in between, the Howell with its much shorter body suggests a two-segment system with one nodal line in between. Consider that the bridge is positioned clearly closer to the top end of the body than to the bottom end. In the Chanot the twin excitation resulted in a prominent plate mode at 812 Hz, here the same principle causes a prominent plate mode at the somewhat lower 756 Hz. Periodicity caused by the twin excitation principle can also be identified in this violin, however, not as clearly as in the Chanot violin.

In summary, similarities can be found between Chanot and Howell. This violin might be comfortable to play and embody enough energy in the low bands, but it is not as balanced as an ordinary student level instrument of conventional construction.

3.4 20th Century Philomele

A contour of an arrow is characteristic to this folksy instrument, probably developed by zither makers around Munich and named after the Greek myth figure Philomele, who was turned into a nightingale by Zeus. The body of this violin is rather flat and usually comes with frets. This fine example here comes without frets and was built in the first half of the 20th century (MKG inventory number 2012.64) (Fig. 12).

Figure 13 represents the bridge motion of the Philomele. (a) *A0* is more than 5 dB below reference and its frequency rather high at 292 Hz. The size and the flat plates result in a slightly smaller volume, therefore increasing *A0*, see Eq. 3. (b) There is just one signature mode in the expected range, although there should be no reason why the breathing and bending should not be possible with the given structure.

Fig. 12 20th century Philomele

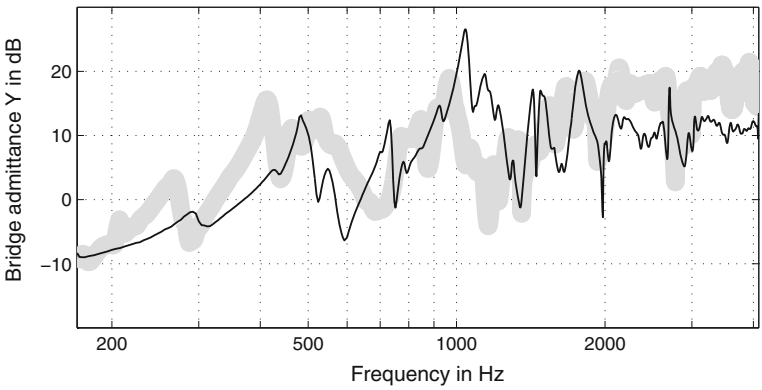


Fig. 13 Bridge admittance *Y* of the Philomele (*thin, black*) in relation to the 1712 "Schreiber" Stradivari violin (*thick, gray*)

A tuning which would arrange B1 modes, A1 and CBR modes in a wider range did obviously not happen. (c) The "brilliance" band is roughly 8 dB lower than the reference. This might seem balanced in relation with the weak "fullness of sound" band; but the violin is not soft in general, as a prominent plate mode at 1,044 Hz exceeds the reference by some 7 dB. This dominant mode can be explained by the relatively flat arching and is located in the band that should modestly stay behind the "fullness" and "brilliance" bands. Whereas in the Stradivari violin plate modes around 1,000 Hz and the "brilliance" band are on the same level, the violin embodies a 15 dB difference. (d) The characteristic dip between body and plate modes is somewhat low in frequency at 600 Hz, constricting the signature mode range.

In conclusion, the violin will be loud enough, but have a very thin tone. Its strength in the middle register might well be associated with Philomele singing as a nightingale. Nevertheless, the violin has little chance to project its tone.

3.5 Zoller Bottle-Shaped Violin

Julius Zoller (1893–?) was engineer at Telefunken and developed his bottle-shaped violin from an engineering point of view. Apart from the unconventional shape, there are several more modifications: the f-holes are replaced by holes in the ribs (five on each side), plate strength strongly diminishes from the middle to the ribs, plates are coated with polished metal on the inside and an additional resonating string is placed below the bridge, tuned to C4. Zoller constructed 60 violins over a course of 5 years before concluding on this design, which was then manufactured in a larger lot of 50–60 pieces a month. He believed that his design would facilitate mass production without degrading tone quality. The violin here, Fig. 14, was

Fig. 14 Zoller *bottle-shaped* violin

rebuilt by Framus[11] probably in the early seventies (MKG inventory number 2012.67).

The bridge admittance measured on this violin reveals the following: (a) Mode *A0* is right in place at 260 Hz, well supported by the resonating string that has been tuned to C4 at 261 Hz. The level is 2.5 dB below the Stradivari. (b) Signature modes are distinct with peaks of 20 dB and more, ranging well across the range of 370–620 Hz. The level is some 5 dB higher than in the Stradivari reference and can be explained with the rather thin body, that would ease bending and the rather spacious plate that would support breathing. However, a more careful tuning process could have brought the modes together to mutually couple with each other. (c) Whereas the "fullness of sound" band is well supported, the "brilliance" band suffers some 5–10 dB with respect to the reference. While this might still be rated as balanced some dominant plate modes foster energy in the wrong band. (d) The dip between body and plate modes is well located between 700 and 800 Hz, but the dips within the range of signature modes are likewise strong (Fig. 15).

In conclusion, the Zoller violin has the potential to develop a full and balanced sound, but it should be tuned in a better way. Although never of a quality for soloists, the value of this cheap violin is high.

[11] Unproven information of the collector.

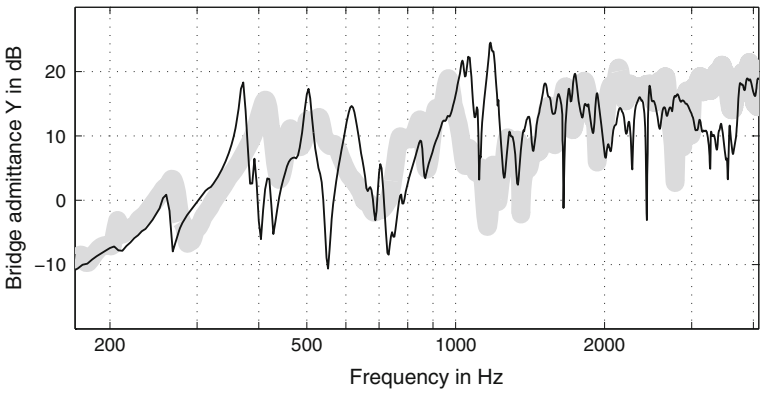


Fig. 15 Bridge admittance *Y* of the Zoller *bottle-shaped* violin built by Simmons in 1928 (*thin, black*) in relation to the 1712 "Schreiber" Stradivari violin (*thick, gray*)

3.6 1893 Stelzner Violin with Top and Back Under Tension

Dr. Alfred Stelzner (1852–1906), educated in music, maths, and physics, believed to have solved the problem of assuredly building quality instruments, while others seemed to achieve a good quality for their copies only by chance and in rare cases. He introduced the idea of attaching the top and back to a rib of parabolically shaped height. Given the highest point in the middle of the corpus, the plates are under tension after attachment. He preferred parabolic and elliptic shapes instead circular shapes throughout his construction believing that this would improve resonances. More than 300 instruments were built in Stelzner's factory and the investigated violin (serial number 180) was built in 1893 (MKG inventory number 2012.89), see Fig. 16. From the seven violins investigated here, the Stelzner violin is closest to conventional construction.

The bridge admittance of the Stelzner violin discussed: (a) A0 is situated well but with a level of more than 6 dB lower than the fine reference. (b) Signature modes are well developed across the acceptable range between 440 and 540 Hz,

Fig. 16 1893 Stelzner violin with *top* and *back* under tension

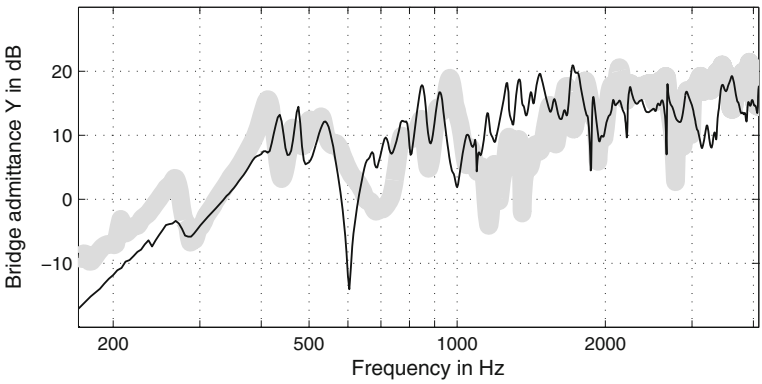


Fig. 17 Bridge admittance *Y* of the 1893 Stelzner violin (*thin, black*) in relation to the 1712 "Schreiber" Stradivari violin (*thick, gray*)

and the level is comparable with that of the reference. (c) The "brilliance" band is less than 5 dB below reference and is balanced when related to the "fullness" band. Plate modes below 1,000 Hz are likewise close to the reference, but plate modes above 1,000 Hz are almost 10 dB stronger than they should. These modes are likely to hamper the idea of a violin sound that should be dominated by a full body, low frequencies, and brilliance. (d) The characteristic dip between body modes and plate modes is placed at the low frequency 600 Hz (Fig. 17).

The Stelzner violin is close to conventional violins not only by construction but also in terms of the structural response. Air, body, and plate modes are as characteristic as they can be for a violin, see Fig. 5 for comparison with the moderate reference. Only the slight exaggeration of plate modes above 1,000 Hz is unfamiliar and can be explained by the tension in the plates created with the attachment on the curved ribs.

This constructive measure of plates under tension, however, brings yet another change to the acoustics: on regarding the admittance curve, the level difference between peaks and dips is generally less than 10 dB, while the fine and the moderate violins both embody level differences of more than 10 dB. A similarly narrow corridor of levels between peaks and dips is observed for the Bechstein/Moor violin with the recessed seam on top. These two violins have in common that constructive measures add to stiffness in the plate. A physical point of view would suggest that the narrow corridor comes from damping. This is not investigated here, as the material properties of the wood in these instruments have not been measured. However, the construction allows an explanation as well. X-mode and ring mode are developed in plates where the waves across the two directions are tuned in frequency, effectively coinciding. Plates out of tune will not develop such clear peaks, which can be observed during the process of tuning a plate and generating the Chladni patterns.[12] This is also true for the higher orders of modes.

[12] As has been done with the Sitka spruce plate shown in Fig. 1.

Here, both violins have a similar problem: Stiffness (Bechstein/Moor) in longitudinal direction will ask for a likewise change of the general proportions, the violin body should be longer. Tension (Stelzner) will also ask for a longer body for the same reason as we establish an increased sound velocity in longitudinal direction. Tension raises the additional question of how to tune the unattached tension-less plate.

In terms of sound perception, the level corridor between peaks and dips deserves a brief discussion. We identify a narrow corridor of 5–10 dB for Stelzner and Bechstein/Moor, and a wider corridor of 10–15 dB for the fine and moderate reference. Violins introducing some kind of singularity, either by the twin excitation principle (Chanot and Howell) or by unusual large and flat plates (Zoller) reveal a 20 dB corridor. Mathews and Kohut (1973) investigated perceived sound quality while varying this parameter on an electronic violin in a range between 0 and 30 dB. He concluded that the corridor should be around 10 dB. Such preference of moderate but not too strong peaks is understood today as Weinreich (1997) explains it: the effect of directional tone colour depends on the peaks but also the amplitude modulation contained[13] in a vibrato depends on the peaks.

In conclusion, the Stelzner violin is a moderate violin. Its construction principle will add some power in an unwanted frequency band, and will obstruct the plate tuning process.

3.7 1928 Savart Trapezium-Shaped Violin

In the early 19th century, the physician Félix Savart (1791–1841) studied the physics of the violin, following the findings of Chladni. The trapezium-shaped body is a result of his studies on glass and metal plates. The violin at the MKG is a replica built by Cyril D. Simmons in Wembley in 1928 (MKG inventory number 2012.34). Simmons used beech wood instead of maple for the back plate. Savart intensified his studies and sacrificed several valued Stradivari and Guarneri violins ceded by Vuillaume. This violin has not been measured.

3.8 1920 Stroh Violin

Johannes Matthias Augustus Stroh (1828–1901?) was a watchmaker born in Frankfurt. He was inspired by the London International Exhibition and moved to England. A fruitful working period with many technical innovations and patents followed. His idea to combine a violin with the horn known from Edison's

[13] A vibrato will cause the harmonics to cruise along the peaks, effectively transforming frequency modulation into amplitude modulation.

Fig. 18 1910 Stroh violin, signed "Stroviols Trade Mark registered … Patent No. 112548 …"

phonograph was patented in 1899. In principle, the bridge motion directly drives a membrane, already leveraged by the proportions of the yoke. A compression chamber drives the horn, effectively amplifying the signal again. Stroh's violin was built until 1942 by his son and became quite popular. It was played in Jazz bands in the Golden Twenties and by gipsy musicians, today it is still widely used in folk music on the Balkan and in Far East and replicas are still built. The investigated example is original and was built in London in ca. 1910 (MKG inventory number 2012.90) (Fig. 18).

Replacing the wood resonator by a horn introduces significant changes. All of the above outlined principles of body and plate resonances are replaced by the principles of a compression wave developing in a horn. The pole-zero plot of the impedance at the end of a horn defines the frequencies of amplification. The Stroh's violin has a conical horn, and frequencies are here periodically spaced like in the cylindrical tube. In singly closed cylindrical tubes of length l the general periodical pattern follows the $f = (2n - 1)^* c_a/4\ l$ regime as compared to the $f = n^* c_a/2\ l$ regime in doubly open pipes, given the speed of sound c_a and taking natural numbers n. For conical horns, the regime falls in between these two, depending on the ratio of diameters at the ends (Fletcher and Rossing 1998, 210–213).

The bridge admittance plot of the violin reveals that the first four peaks directly relate to the conical horn. The frequencies 272, 448, 676, and 964 Hz follow the outlined principles.[14] Note that the row of frequencies does not strictly follow an arithmetic progression. Especially the first peak steps out. This fully complies with the findings of Ayers et al. (1985) quoted by Fletcher and Rossing in that the first peak is strongly increased in frequency in relation to the other peaks, when the mouth is wider than the throat in singly closed conical horns. And the other peaks also do not follow a strict arithmetic progression for all singly closed conical horns. The peak frequencies measured here match with the findings of Ayers.

[14] Calculation without correction term for estuary. The horn is about 35 cm long. The relation of diameters is about 4.

Surprisingly, the air mode *A0* and the first signature mode both perfectly fall together with the fine reference. Clearly, the horn was tuned to match such parameters known from conventional violins. The rest of the plot does of course not match the wooden reference. The brief discussion from the conservative perspective: (a) *A0* matches in frequency, as discussed, and the level is only 2.5 dB below reference. (b) There is only one mode in the range of signature modes, however, it is wisely positioned between 400 and 500 Hz and comes with a level of 4 dB above the strongest mode in the Stradivari. (c) The "brilliance" band is about 8 dB below reference. This is similar to Bechstein/Moor, Chanot, Howell, and Zoller. This violin, however, differs in that the band in between "fullness" and "brilliance" is not strong, but it is rather moderate. This sound is appreciated and this might be one of the reasons why Stroh's violin is so popular. (d) A characteristic dip between signature and plate modes cannot be expected. One of the dips between horn peaks falls at 550 Hz (Fig. 19).

When listening to Stroh's violin, one will clearly hear a violin. This is due to the string bowing mechanism and the related characteristic of string excitation. At the same time, a metal horn can be heard. Some of these violins sound like a conventional violin played back on an old phonograph. So the signature of the horn can well be heard categorically and the violin played by a bow can be heard categorically, simultaneously.

Another important sound attribute is the violin's loudness. The bridge motion is amplified thoughtfully taking advantage of the impedances given at each step of conversion. Secondly, the horn has a narrow beam, while the violin has a dipole and multi-pole character. Both effects work together for a "four-fold sound intensity", as Stroh stated.

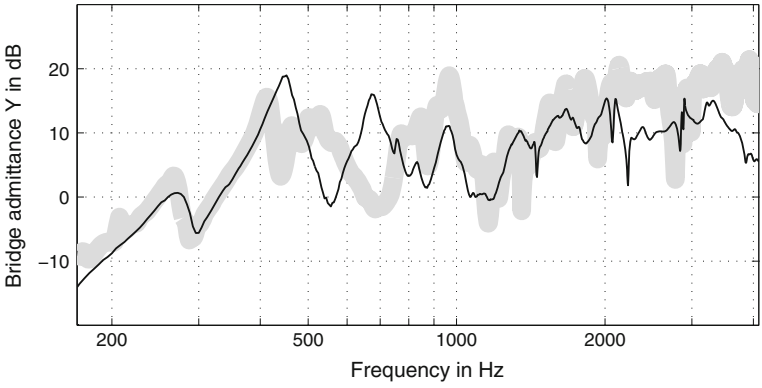


Fig. 19 Bridge admittance *Y* of the 1910 Stroh violin (*thin, black*) in relation to the 1712 "Schreiber" Stradivari violin (*thick, gray*)

3.9 Sound Radiation in Comparison with Bridge Mobility

Bridge mobility translates to radiation rather directly, however, not one by one. This can be seen in Fig. 4. Peaks in the mobility plot will be found in the SPL plot, but often at unexpected levels or bandwidths. One main reason for differences is the directional tone colour, as beams become narrower with increased frequency. For analyses in frequency bands above 2 kHz statistical approaches are common. Other causes are differences in impedance matching and phase relations.

Figure 20 compares the SPL level differences in low frequency bands, as discussed before and well related to construction. More precisely, the bars in the figure represent the level of *A0* and the level of the signature modes, both related to the level of the plate modes in the range of 1 kHz, for all violins.

First of all, the fine reference is outstanding as *A0* and signature range both radiate strongly in relation to the band around 1 kHz. Instruments made by Luthiers are moderate in this respect, see the student instrument, Howell's violin and Philomele, while the truly experimental constructions perform poorly, see Bechstein, Zoller, and Stelzner. Secondly, the idea that *A0* radiates somewhat weaker than the signature modes is familiar for conventional violins, compare the two reference violins against all others.

For the Chanot and the Howell violin one expects a confirmation of the above finding that the twin excitation principle is of advantage for *A0*. In fact, Howell's violin is the only one with an *A0* level above the level of signature modes. And for the Chanot violin the *A0* level reaches that of signature modes, uncommon with conventional violins. The finding is confirmed. One could argue that Bechstein and Stelzner also achieve comparable levels for *A0* and signature modes, but equality is caused by poor signature modes, as discussed above, and is not caused by any

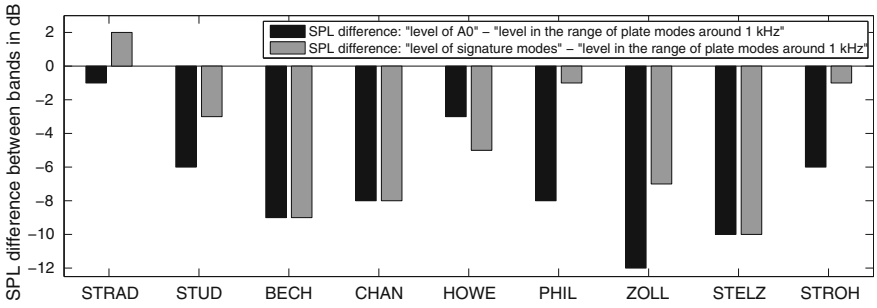


Fig. 20 SPL difference in dB between low frequency bands for the 1712 "Schreiber" Stradivari (STRAD), 1900 Markneukirchen student level violin (STUD), 1930 Bechstein/Moor violin (BECH), 1820 Chanot violin (CHAN), 1836 Howell violin (HOWE), 20th century Philomele (PHIL), Zoller violin (ZOLL), 1893 Stelzner violin (STELZ), 1910 Stroh violin (STROH); *black* level of *A0* minus level of plate modes around 1 kHz, *grey* level of signature modes minus level of plate modes around 1 kHz

improvement on *A0*. Finally, it is the one prominent plate mode in the Chanot violin that obstructs the given advantage on *A0*.

The Zoller violin has the lowest *A0* level. This is partially due to the fact, that measurements are done perpendicular to the top, while there are no *f*-holes on the top and only some small holes in the ribs.

Music students who played the different instruments reported that they liked the Philomele best. This seems surprising against the background of bridge admittance measurement which concluded on a thin sound. However, in terms of radiation the Philomele appears to be well balanced. The level relations between bands are almost as good as for the student level instrument and definitely better than for all experimental violins. Therefore, the students' report is in agreement with Dünnwald's discussions on bands. In this respect also the Stroh violin seems to be balanced which might be one of the reasons why its sound is appreciated and it is still rebuilt and widely used.

4 Summary

We investigated seven experimental violins acoustically and referenced the results against a fine Stradivari violin and against a moderate student level instrument. From the perspective of conventional violin making and when guided by the principle ideas of plate and body tuning, all the constructive changes and their impacts to acoustics can be explained. Most of these changes result in a poorer balance of the instrument or in thin sound. There are two main conclusions when considering old masterpieces and at the same time these experimental violins from the 19th and 20th centuries: the old masters understood very well the acoustical outcome for each step of the crafting and tuning process, and that is why the performance maximum 300 years ago brought forth the most admired instruments. Secondly, the engineering perspective brought some fresh but disruptive innovations, most of which ignored the existing knowledge and failed to revolutionize violin making. A rare exception is the Stroh violin, as it effectively constitutes a new musical instrument and as it is engineered well enough to produce an appreciable sound.

References

Ayers, R. D., Eliason, L. J., & Mahgerefteh, D. (1985). The conical bore in musical acoustics. *American Journal of Physics, 53*, 58–537.

Bissinger, G. (1992a). Effect of f-hole shape, area, and position on violin cavity modes below 2 kHz. *Journal of Catgut Acoustical Society, 2*(2), 12–17.

Bissinger, G. (1992b). Effect of violin cavity volume (height) changes on the cavity modes below 2 kHz. *Journal of Catgut Acoustical Society, 2*(2), 18–21.

Bissinger, G. (2008). Structural acoustics of good and bad violins. *Journal of the Acoustical Society of America, 124*(3), 1764–1773.
Cremer, L. (1981). *Physik der Geige (Engl. The physics of the violin)*, S. Hirzel Verlag Stuttgart.
Dünnwald, H. (1982). Messung von Geigenfrequenzgängen. *Acustica, 51*, 282.
Dünnwald, H. (1984). *Akustische Messungen an zahlreichen Violinen und Ableitung objektiver Kriterien für deren klanglichen Eigenschaften*. PhD thesis, RWTH Aachen.
Dünnwald, H. (1991). Deduction of objective quality parameters on old and new violins. *Journal of Catgut Acoustical Society, 1*(7), 1–5.
Fletcher, N. H., & Rossing, T. D. (1998). *The physics of musical instruments*. New York: Springer.
General Technical Report FPL–GTR–113. (1999). *Wood handbook. Wood as an engineering material* (Ed.) United States Department of Agriculture, Madison, Wisconsin.
Gough, C. (2010). A finite element approach towards understanding violin structural modes. *Journal of the Acoustical Society of America, 127*(3), 1791.
Hutchins, C. M. (1981). The acoustics of violin plates. *Scientific American, 245*, 170–186.
Hutchins, C. M. (1985). Effects of an air-body coupling on the tone and paying qualities of violins. *Journal of Catgut Acoustical Society, 44*, 12–15.
Hutchins, C. M. (1994). A measurable result of bi-tri octave plate tuning. *Journal of the Acoustical Society of America, 95*(5), 2913.
Hutchins, C. M. & Benade, V. (1997). Research Papers in Violin Acoustics 1975–1993, Vol. I + II, *ASA*, *Woodbury*.
Hutchins, C. M., & Voskuil, D. (1993). Mode tuning for the violin maker. *Journal of Catgut Acoustical Society, 2*(4), 5–9.
Jahnel, F. (1981). *Die Gitarre und ihr Bau (Engl. The guitar and ist manufacturing)*. Fachbuchreihe das Muskinstrument Bd.25, Verlag Das Musikinstrument, Frankfurt a. M.
Jansson, E. V., Molin, N.-E., & Saldner, H. O. (1994). On eigenmodes of the violin–Electronic holography and admittance measurements. *Journal of the Acoustical Society of America, 95*(2), 1100–1105.
Jansson, E. V., Moral, J. A., & Niewczyk, J. (1988). Experiments with free violin plates. *Journal of Catgut Acoustical Society*, ***1***(2), 2–6.
Knott, G.A. (1987). *A modal analysis of the violin using MSC/NASTRAN and PATRAN*. MSc thesis, Naval Postgraduate School, Monterey, CA.
Marshall, K. D. (1985). Modal analysis of a violin. *Journal of the Acoustical Society of America, 77*(2), 695–709.
Mathews, M. V., & Kohut, J. (1973). Electronic Simulation of violin resonances. *Journal of the Acoustical Society of America, 53*, 1620–1626.
Meinel, H. (1937). Über Frequenzkurven von Geigen. *Akustische Zeitschrift, 2*, 22–33 and 62–71.
Meinel, H. (1939). Akustische Eigenschaften klanglich hervorragender Geigen. *Akustische Zeitschrift, 4*, 89.
Molin, N.-E., Lindgren, L.-E., & Jansson, E. V. (1988). Parameters of violin plates and their influence on the plate modes. *Journal of the Acoustical Society of America, 83*(1), 281–290.
Moral, J. A., & Jansson, E. V. (1982). Eigenmodes, input admittance, and the function of the violin. *Acustica, 50*(5), 329–337.
Newton, W. (1836). *Repertory of patent inventions* (Vol. 8, p. 171). United Kingdom: Sherwood Gilbert and Piper.
Reinicke, W. (1973). *Die Übertragungseigenschaften des Streichinstrumentenstegs*. PhD thesis, Technical University of Berlin.
Rossing, T. D. (2010). *The science of string instruments*. New York: Springer.
Saldner, H. O., Molin, N.-E., & Jansson, E. V. (1996). Vibration modes of the violin forced via the bridge and action of the soundpost. *Journal of the Acoustical Society of America, 100*(2), 1168–1177.
Schleske, M. (2002). Empirical tools in contemporary violin making: part II: psychoacoustic analysis and use of acoustical tools. *Catgut Acoustical Society Journal 4* (5) (Series II).

Weinreich, G. (1997). Directional tone color. *Journal of the Acoustical Society of America, 101*(4), 2338–2346.

Zopf, S. R. (2000). *Untersuchung neuer und historischer akustisch- optischer Meßmethoden im Geigenbau*. Master‘s thesis, Universität Wien.

Fourier-Time-Transformation (FTT), Analysis of Sound and Auditory Perception

Albrecht Schneider and Robert Mores

1 Introduction

In the present chapter, we will reexamine the so-called Fourier/time transformation (FTT) that has been proposed by Ernst Terhardt (1985, 1992, 1998) as a tool for analysis and representation of audio signals such as speech and music. The main reason for suggesting such an approach was that Terhardt (1985) saw a different interpretation of the Fourier transform (as is widely used for spectrum analysis), on the one hand, and a need to develop a transform suited to perform time/frequency analysis comparable to that of the mammalian auditory system, on the other. Hence the aim of the FTT is to provide a time-to-frequency transformation equivalent to parameters in auditory processing as well as a "natural" approach to signal analysis (cf. Terhardt 1985, 1998, 78–97). In order to assess the possibilities the FTT approach might offer in regard to signal analysis, some other methods relevant for musical acoustics and psychoacoustics such as the short-time Fourier transform (STFT), autoregressive spectral modeling (AR) and Wavelet transform (WT) are presented in a brief survey, and are illustrated by some examples. Different approaches to time/frequency analysis are also viewed as to their power with respect to the so-called uncertainty product $\Delta t\ \Delta f$.

Over the past decades, there has been a broad range of research directed at understanding the functional anatomy and physiology of the auditory system (for summaries of research, see Oertel et al. 2002, Pickles 2008, Winer and Schreiner 2011). Since about 1980, computational models of the auditory system have been issued that were progressively taking neurophysiological data and results from

A. Schneider (✉)
Institute of Musicology, University of Hamburg, Neue Rabenstr. 13, D-20354 Hamburg, Germany
e-mail: aschneid@uni-hamburg.de

R. Mores
University of Applied Sciences, DMI Faculty, Finkenau 35, D-22081 Hamburg, Germany
e-mail: mores@mt.haw-hamburg.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_13,

behavioral studies into account (for an overview, see de Cheveigné 2005, Meddis et al. 2010). By including elements representing hair cell transduction and neural activity patterns in the auditory nerve (AN) as well as in some of the relays along the subsequent neural pathway, complexity of the models as well as realism in performance has been increased by far (see, e.g., Meddis and O'Mard 1997, 2006). While most current models are based in the time domain, there are some operating in the frequency domain. Traditionally, analysis in the time domain has been concerned with signal periodicity detection and estimation of 'pitch' from the repetition frequency of the envelope (f0). Analysis in the frequency domain typically has been done with the spectrum comprising a fundamental frequency f_1 and higher harmonics $n \times f_1$ in view. For both approaches that have been pursued in auditory research for more than 150 years now (see de Boer 1976; de Cheveigné 2005), there are reasons at hand referring to the structure of audio signals (that can be represented both in the time and in the frequency domain) as well as with the functional anatomy and physiology of the mammalian auditory system. Considering only the first stages of auditory processing, and allowing for a rather schematic view, there is (1) transfer of waves from the environment through the ear channel to the tympanon. Then there is (2) a mechanical transmission line from the tympanon by means of the ossicles to the oval window where the pattern of vibration is transferred into (3) the cochlear fluid system in which a travelling wave with a relatively steep maximum for individual frequencies corresponding to sine tones is observed. Hence it has been concluded that a complex harmonic wave is decomposed in the fluid channel such that several maxima representing single partials or groups thereof will be observed. The cochlear partition with (4) the basilar membrane (BM) as well as structures combined with the BM are regarded as a filter bank of k channels capable to decompose a complex signal into partials or groups thereof. (5) Inner hair cells (IHC) effect mechanoelectrical transduction so that the output of each of the BM channels is coded into a train of neural spikes that are (6) represented in fibers of the AN. Modeling transmission of audio signals from the pinna to the stapes (a mechanical system with impedances and admittances) and within the fluid ducts of the cochlea (a hydromechanical system that incorporates nonlinearities; see Nobili and Mammano 1999) as well as the transduction mechanism on the IHC and AN level is quite complex since every element in the transmission chain as well as their interaction must be adequately covered, that is, as close as possible to empirical data from (mostly, animal) experiments and behavioral studies (cf. Meddis and Lopez-Poveda 2010).

In regard to such a complex transmission line that may incorporate also relays of the auditory pathway such as the cochlear nucleus (CN) or models for processing at even higher levels (the superior olivary complex and the inferior colliculus), restricting an analysis to peripheral filtering processes as effected in the cochlea (as is done in this chapter) may seem odd. The point, however, is that initial analysis on the BM and IHC level seems decisive since it can be shown that distinctive features of complex sounds such as salient or ambiguous pitch structure, harmonic or inharmonic spectrum (leading to percepts classified as consonant or dissonant),

and also phenomena such as combination and difference tones are derived from peripheral processing (for examples, see Schneider and Frieler 2009). In the case of the peripheral processing lacking sufficient precision (consequent to, for example, inappropriate design of BM filters), feature extraction at this stage of processing and also on higher levels of the auditory pathway can be significantly hampered.

2 Uncertainty Relation and Time/Frequency Resolution

The uncertainty relation known from quantum mechanics states that a particle can be defined exactly either as to its impulse p or to its place x. Since exact definition of the impulse precludes exact definition of the space (in regard to wavelength), a situation where both have to be taken into account leads to the product of place and impulse such that $\Delta x\, \Delta p \geq \hbar/2$ ($\hbar = h/2\pi$ with $h =$ Planck's constant). This basic equation became known as the uncertainty relation and has been adapted, with necessary modifications, into various fields of science such as communication theory and acoustics (Gabor 1946). According to Gabor (1946), for signals a limit for the product of time resolution and frequency resolution exists like

$$\Delta f \Delta t = 1/2 \tag{1}$$

This minimum is restricted to very few 'ideal cases' (see below) so that for real signals such as sound of a certain duration and bandwidth values above 0.5 will apply. In a general formulation, the uncertainty relation for acoustic phenomena such as impulses (cf. Meyer and Guicking 1974, 92ff.) can be given as

$$\Delta t \Delta f \geq 1 \tag{2}$$

As can be demonstrated by calculation, the lower limit of $\Delta t\ \Delta f = 1$ can be achieved for a Gaussian impulse while for almost every other pulse type $\Delta t\ \Delta f > 1$ applies.

Taking two extremes, a Dirac-δ (with a duration approaching zero and an impulse height approaching infinity) and a sine wave of an arbitrary frequency f_i lasting from $-\infty < t < \infty$, the impulse is defined exactly as to time t (ms), and the sine wave as to frequency f (Hz), in a two-dimensional time–frequency space. "Real-world" signals such as produced by musical instruments including the human voice are neither as short in duration as a Dirac-δ, nor infinite in duration as the undamped sine wave repeating itself at the same frequency. Of course, in regard to spectral bandwidth, the Dirac impulse and the sine tone of a given frequency also represent two extremes. In music as well as in other audio signals such as human speech or birdsong, the situation typically is that a number of complex sounds each comprising n harmonic or inharmonic partials occur at a certain time, and have disappeared due to damping forces after a duration of, in most cases, a few hundred milliseconds or perhaps several seconds. Hence we are

dealing with sequences of complex sounds such as melodies, or with several such sequences played or sung more or less in parallel (in regard to tracks of fundamental frequencies) as well as more or less synchronous (as regards onsets of tones/notes) as in homophonic and polyphonic music.

In this respect, conventional western staff notation constitutes an acceptable approximation to a two-dimensional time/frequency representation with the ordinate *y* giving frequency on a log scale, and the abscissa *x* time on a linear scale (cf. Rossing 1982, 134–135). One can therefore substitute staff notation with semilogarithmic graph paper to yield a similar (but more precise) notation for monophonic or polyphonic music (for an example of a Bach chorale with four voices, see Schneider 2001). It has to be noted, in this context, that western staff notation in regard to 'pitch' information represents the fundamental frequency f_1 (as is obvious from definitions such as standard pitch $A_4 = 440$ Hz or "middle c" $[C_4] = 261.6$ Hz in equal temperament). Whether the tone notated on staff as C_4 is a pure (sine) tone or a complex tone cannot be gained from Western staff notation, which does not include spectral information. However, it is implied from $A_4 = 440$ Hz that any complex tone played to render this note audible should comprise a fundamental frequency f_1 at 440 Hz (though, at least in perception, a 'pitch' corresponding to 440 Hz could be realized also with an envelope repetition frequency f0 = 440 Hz while the fundamental of the spectrum is weak or even missing).

Of course, one could further substitute staff notation with a melogram or spectrogram (sonogram) as a two-dimensional representation of sound and music in a time/frequency space. We will do this with a musical example offered recently by Florian Messner (2011) who, together with another singer, recorded a phrase noted down in staff notation by Franchino Gafori (Franchinus Gaffurius, 1451–1521), in his *Practica musicae* (Milan 1496). Gafori (Lib. III, cap. 14: de falso contrapuncto) gave us this piece of two-part music then still in practice in the Lombardic in vigils and in the mass for the dead because he thought it defied all rules of counterpoint (…*ab omni modulationis ratione seiunctus est*). What in fact singers were performing was vocal music where two voices go in parallel with dissonant intervals (seconds, fourths) between them. Singing styles as well as instrumental music organized as a diaphonia with two voices forming narrow intervals were or even still are in use in the Balkans (notably in areas of Bosnia and Herzegowina, Croatia, Albania, Bulgaria). Since two notes sung in parallel at the interval of a minor or a major second will have fundamental frequencies so close as to fall into one 'critical band' (CB), they cannot be separated by the auditory filter bank, and thus a sensation of roughness from the interaction of fundamental frequencies as well as from other partials in their respective CBs will result. In Bulgarian diaphonic singing, one finds two (female) voices approaching each other as close as ca. 45–80 cents (cf. Schneider et al. 2009), that is, from about a quarter tone to a chromatic semitone.

For the Lombardic *contrapunctus falsus* as performed by two male singers, the spectrogram shown in Fig. 1 results.

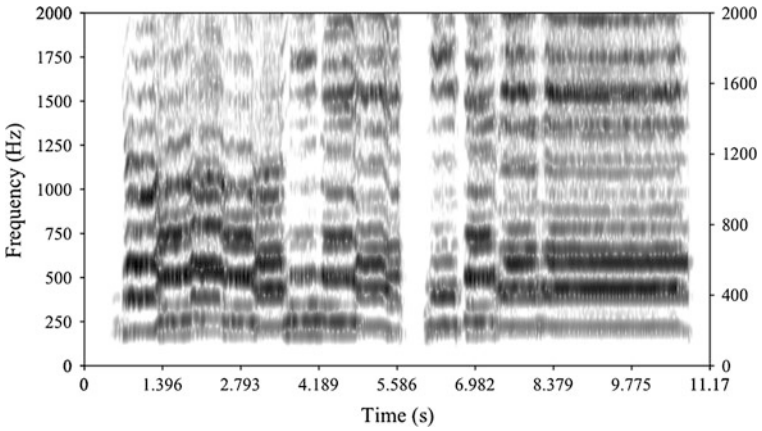


Fig. 1 Lombardic diaphony, two male singers, spectrogram 0–2 kHz

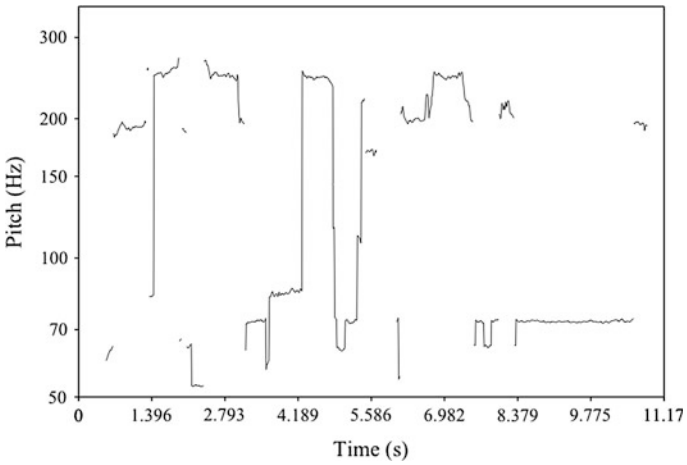


Fig. 2 Pitch (f0) tracking for lombardic diaphonia, autocorrelation method

Though the spectrogram has been calculated in the frequency domain with a rather high resolution as to time and frequency,[1] the trajectories of the fundamental frequencies for the two voices will be difficult to recognize. Also kind of a melogram representing the pitches (calculated in the time domain with a special autocorrelation algorithm, Boersma 1993) will give only some rough idea as to the movement of the voices (see Fig. 2).

[1] Settings for the analysis performed with the Praat software (Boersma and Weenink 2011) were a time window of 30 ms with a Gaussian weighting, a time step of 2 ms from one frame to the next, an analysis bandwidth of 2 kHz and a frequency step of 2 Hz. The sound sample of 11.17 s was processed in 5253 (overlapping) frames.

It is possible to find the fundamental frequencies for the two male voices even for narrow intervals with a standard frequency analysis based on FFT, provided the window of analysis is long enough to ensure that relevant components can be separated.

Applying a Discrete Fourier Transform (DFT, cf. DeFatta et al. 1988, 238ff.) to a digital signal $x(n)$ with a period of T, the frequency resolution Δf depends on the sampling rate F_s and the transform length (often also called 'frame' or 'window') of size N. The discrete frequencies f_k for a spectrum $X(k)$ of the signal can be calculated as

$$f_k = k\,(F_s/N) \quad \text{where } k = 0, 1, 2, 3, \ldots, N-1 \text{ is the frequency index.} \tag{3}$$

The frequency resolution hence depends on the ratio Fs/N and can also be expressed as

$$\Delta f = 1/T = F_s/N \tag{4}$$

It is obvious from Eq. (3) that basic relations defined for analogue band pass filters hold likewise in the digital domain. For a narrow-band filter (cf. Küpfmüller 1968, 71f.), the response time τ is defined as

$$\tau = 2\pi/\Delta\omega = 1/\Delta f (\text{for } \omega = 2\pi f) \tag{5}$$

Hence the response time and bandwidth of the filter are in reciprocal relation. For any frequency resolution Δf designed for the filter, a corresponding response time τ can be calculated; since τ in this respect defines Δt of the filter (taken as an ideal, non-dispersive band pass; cf. Meyer and Guicking 1974, 92ff., 346ff.), the product $\Delta f\,\Delta t \geq 1$ applies equivalent to Eq. (1).[2] The uncertainty relation that, as a general principle, needs to be adopted for specific areas, underlies also digital sampling and frequency analysis (Eqs. 2, 3) where a signal $x(n)$ of period T sampled at F_s can be determined in regard to its spectrum $X(k)$ the better the longer the transform size N is chosen. This, however means that good frequency resolution Δt can be achieved only at the cost of rather poor time resolution Δt.

With respect to our example, the Lombardic diaphonia, the sample rate of 44100 per second will require a window size or transform length of at least $2^{12} = 4096$ to ensure a frequency resolution $\Delta f \sim 10.77$ Hz. As can be easily checked, the exact value for Δf is 10.7666 Hz; Δt is determined by the transform of length $N = 4096$ samples $= 92.8798$ ms. If we leave out windowing and other effects, the product of time and frequency achieved in FFT-based analysis indeed would be unity.[3] For the analysis of the sound example, FFT windows of 2^{12}, 2^{13} and 2^{14} samples were employed together with a spectral peak estimation algorithm. Frequency readings were confined to full frequency values (e.g., 195, 222 Hz) averaged over the

[2] A formal proof can be given on the basis of the Cauchy-Bunjakowski-Schwarz inequality (cf. Meyer and Guicking 1974, 95, 108; Papoulis 1962, 63).

[3] Applying no specific windowing function means a rectangular window is chosen for which the so-called Equivalent Noise Bandwidth (ENBW [Bins], see DeFatta et al. 1988, 262ff.) is 1.0.

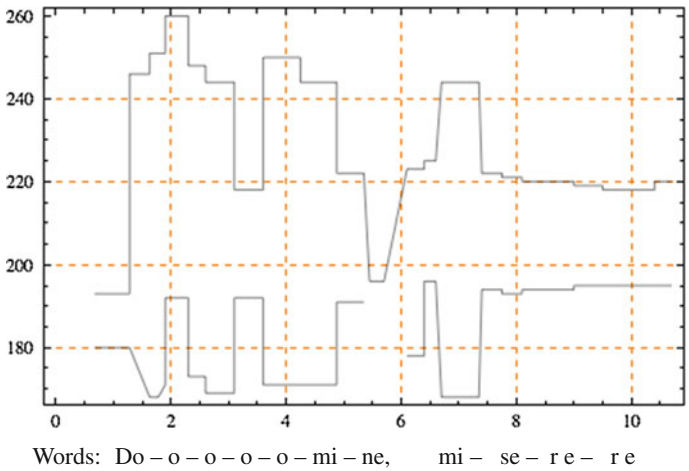


Fig. 3 Lombardic diaphonia, 2 male voices, tracks of fundamental frequencies/time

window of length N. The results of the time/frequency analysis have been tabled and then plotted as shown in Fig. 3. For reasons of readability, a linear frequency scale (ordinate) was chosen. The movements up and down (melodic contour) as well as musical intervals formed between the two voices by their fundamental frequencies over time are clearly visible. However, the relatively poor time resolution of the analysis is also quite obvious since the 'pitches' sung (represented by their respective fundamental frequencies f_1) are indicated according to the transform size that has been employed. For example, at $F_s = 44100$ samples, a window of 8192 samples means a time interval of 185.76 ms for which a spectrum is calculated that contains information as to the 'average pitch' that, in our example, was realized by two singers within this span of time. In reality, there can be marked shifts of fundamental frequency within one frame or window of length N. In fact, the intonation practiced by the two singers in recording this piece of music shows far more subtle fluctuations than shown in Figs. 2 and 3 as became obvious in a more detailed analysis carried out with high resolution tools (Wigner transform and FFT combined with LPC pitch tracking and very small hop ratios).

What is evident from Fig. 3 is that the two singers didn't start in unison (what the notation provided by Gafurius would have demanded) but at an interval of about a semi-tone (193:180 Hz $\sim$ 122 Cents). Also, one can see that at the end of the phrase (from 7.5″ to 10.7″ on the time scale) a long dissonant interval, namely a major second based on the notes G_3 and A_3 occurs. While singing their respective notes/tones forming the major second, the singers adjust their intonation several times (the interval size varies from an initial 233/234 cents to ca. 201 and even 193 cents towards the end). There are some more details one can study with the data condensed in Fig. 3 at hand. Figure 3 can be regarded as kind of a descriptive 'notation' derived ex post from an actual performance. This notation, by the way, could be transformed back into a symbolic notation (e.g., western staff notation).

If one would need to improve temporal resolution of the analysis, there are methods at hand in digital signal processing (DSP) which permit to achieve this goal without sacrificing adequate frequency resolution. One of the most basic and at the same time most efficient procedures is to overlap consecutive frames of analysis (what has been done to some degree also for the present analysis). In case overlap is almost complete and the so-called 'hop ratio' therefore very small, a sequence of signal spectra will result following one another at a short delay of n samples while the frequency resolution of each spectrum is determined by N. Such an analysis technique is well suited for transients where the rate of change in the signal per time often is significant. We will show an example for such an analysis below. The point of interest with respect to choosing a certain method of analysis of course is this: what is the degree of exactitude necessary in regard to (a) auditory perception and relevant psychoacoustic parameters? Further, which technique should be used if (b) the study of musical structure is an issue (e.g., when studying music not well documented yet)? In addition, signal analysis could also be pursued in regard to (c) acoustics of certain instruments where the aim often is to investigate processes of vibration, sound production and sound radiation. The precision needed under (c) is certainly much higher than that required for (a) or wanted for (b).

Taking Fig. 3 as an example, one may call the analysis plausible in regard to musical structure since the melodic contours of the two voices and the intervals formed between them can be followed with ease. What is less accessible to intuitive understanding in this plot, though, is the exact size of the intervals realized by the two voices. Of course, musicians and musicologists will have an idea as to the fundamental frequencies of notes in a diatonic scale (at least in regard to main intervals). However, a number of deviations in intonation that were documented in the signal analysis are difficult to read from the tracks in Fig. 3. In regard to auditory perception, the precision achieved in the plot in Fig. 3 probably is above that ordinary listeners might achieve by using their ears only for analysis (even trained musicians might find it difficult to separate the two voices which are quite close in register, and in the recording at hand do not differ much as to their respective timbre). In sum, one could argue that the analysis as shown in Fig. 3 is sufficient to illustrate a musical structure as was put to sound by two male singers, and it represents about the result trained listeners might obtain from an aural analysis of the musical phrase as recorded on CD.

In regard to time and frequency resolution as are most relevant for signal analysis, it should be noted at this point that the 'uncertainty relation' (or 'relation of indeterminacy') yields $\Delta f \, \Delta t \geq 1$ for linear systems such as analogue band filters.[4] For the auditory system, it has been shown in experiments based on biophysical cochlea models (cf. Mammano and Nobili 1993, Nobili and Mammano 1999)

[4] There are several definitions as to 'linear'. In electronics, linear refers to circuits (like LRC filters) in which linear relations exist between physical magnitudes (induction, capacity, resistance, gain) and where all voltages and current are proportional to the electromotive force driving the system (cf. Küpfmüller 1968, 12f.). In signals and systems theory, linearity is defined by Bachmann (1992, 9) like this: superposition at the input has the same effect as superposition at the output.

that time/frequency analysis of the cochlea for the range of speech signals above 200 Hz already for a passive model comes close to $\Delta f\ \Delta t \approx 0.55$ (Russo et al. 2011), that is, very close to the theoretical limit of 0.5 as defined by Heisenberg's 'uncertainty relation' or the equivalent formulation Gabor (1946) has given for time/frequency resolution as a relevant parameter for communication systems. The general concept Gabor advanced was that for every type of resonator a *characteristic rectangle* of about unit area can be defined in a time/frequency plane. For a sharp resonator such as a narrowband filter $\Delta f\ \Delta t \approx 1$ can be assumed. From mathematical considerations as well as from properties of some elementary signals (sine or cosine wave, Dirac-δ) Gabor (1946, 435) concluded that the signal for which $\Delta f\ \Delta t = 1/2$ *applies is the modulation product of a harmonic oscillation of any frequency with a pulse of the form of a probability function*. (For an 'ideal' bandpass filter he calculated the value 0.571). Gabor suggested that a time/frequency space (understood as an information diagram with the axes time and frequency) can be divided into rectangles which have sides defined by Δf and Δt, respectively. According to Gabor, each area $\Delta f\ \Delta t$ *represents one elementary quantum of information*; he therefore proposed to call such an area a *logon*.

Remarkably, Gabor (1946, Part 2) included hearing into his study, where he is making reference to several empirical studies on difference limens for pitch and time (as had been published by Shower and Biddulph in 1931, and by Bürck et al. 1935; see below). Gabor argued that the ear (or, rather, the sense of hearing) disposes of a *threshold information area* in regard to frequency (pitch) and time, and of an *adjustable time constant at least between 20 and 250 ms*. Thus he regards hearing a most relevant field where his concept of time/frequency areas or *logons* is of practical significance.

It is obvious that basic ideas as formulated by Gabor for signal and systems theory also underlie some other approaches, notably wavelet analysis (cf. Dutilleux et al. 1988; Mertins 1999, Chap. 7; Evangelista 1997). In fact, it can be demonstrated that, in regard to fundamental mathematical concepts, formal equivalence exists for the Wigner transforms, Gabor coefficients, and Weyl-Heisenberg wavelets (see Dellomo and Jacyna 1991). Gabor's concept and related concepts by Eugene Wigner and J. Ville have led to a systematic treatment of linear and non-linear time/frequency analysis of signals (see Cohen 1995; Flandrin 1999; Mertins 1999). Application of the Wigner transform (WiT) to acoustical signals is possible with some modification of the original formulation (cf. Yen 1987) and can yield high-resolution time/frequency representations. For a complex-valued signal $s(t)$, WiT can be calculated according to

$$W(t,\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau} s\left(t+\frac{\tau}{2}\right) * \left(t-\frac{\tau}{2}\right) d\tau \tag{6}$$

where * denotes the complex conjugate. For practical applications in DSP, the integral comes down to a summation, and a window function is applied since the WiT is a bi-linear transform that produces cross terms between spectral energy

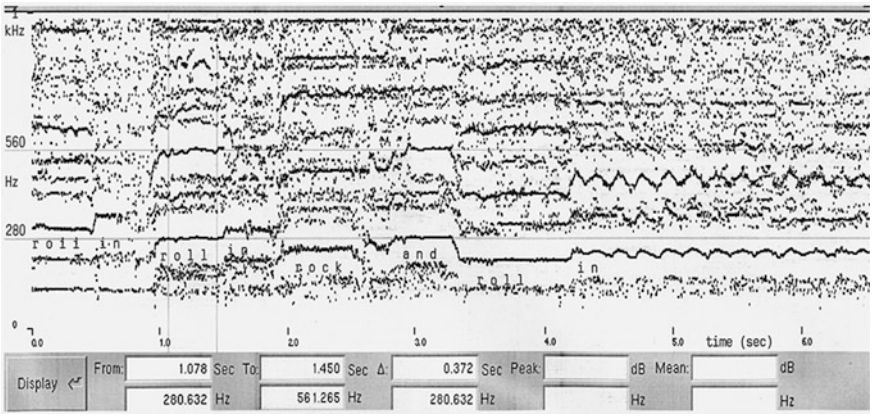


Fig. 4 Joni Mitchell *In France they kiss on mainstreet*, WiT+FFT+LPC

peaks resulting from a real-valued signal. The cross spectrum appears in the time and in the frequency representation and contains sum and difference of the original spectral components. The window function helps to cancel out cross terms. Also, a good comprise solution suited to suppress spurious spectral components is a combination of FFT and WiT for which parameters can be set so as to cancel out most of the unwanted cross terms while improved resolution (as compared to FFT alone) is maintained. As an example, an analysis of a phrase sung and played by Joni Mitchell in a demo version of her song *In France they kiss on mainstreet* is presented (Fig. 4). For the analysis, a combination of WiT and FFT as well as a spectral peak picking algorithm (linear predictive coding, LPC, see Markel and Grey 1976) was used.[5] One can easily trace the fundamental frequency as well as the second partial (i.e., the first harmonic an octave above the fundamental) of Mitchell's voice. In regard to intonation, some pitches within the phrase "roll-in, roll-in, rock and roll-in" are more stable than others; Mitchell goes into a marked vibrato on the last, long held syllable "in".

3 Time/Frequency Analysis: Some Applications and Examples

There are quite many time/frequency analysis techniques that have been applied to musical signals (for an overview see Kostek 2005). In retrospect, sonagrams derived with analogue filtering were a valuable tool for sound analysis and also for

[5] Analysis for 0–1 kHz was performed with the *Sonogram* software (Hiroshi Momose 1991); settings were FFT+Wigner, time window 2048 pts, Hanning weighting, time increment 85 pts = 1.77 ms; LPC, sideband suppression 80 Hz, dynamic range of analysis and graph representation −20 to −1 dB.

musical transcription (see Schneider, this volume). The output of the analysis was plotted on special paper as kind of a 2½-D graph indicating spectral energy for quasi-continuous frequency bands over time (with relative amplitude per frequency or small frequency band marked as grayscale or, rather, "blackscale"). With DSP tools, sonagrams (now often labelled son*o*grams or spectrograms, see Fig. 1) were typically calculated by means of FFT algorithms operating in the time or in the frequency domain (or in both). In regard to time and frequency resolution, the common Fourier analysis effected by means of a FFT implementation (cf. DeFatta et al. 1988) bears to the fundamental relation of $\Delta f\, \Delta t = 1$ if we neglect weighting functions and other possible restrictions. In practice, the result of the analysis can be improved in many details by zero padding and interpolation of data. In addition, overlap of frames (typically, blocks of samples of length 2^n) allows to account for changes a signal undergoes in time (e.g., frequency and amplitude modulation). Further, peak-picking algorithms which detect peaks in spectral envelopes and create tracks of such peaks from one spectrum to the next are very useful tools in particular for the analysis of transient or modulating signals (see Kostek 2005; Beauchamp 2007).

For a demonstration of alternative techniques of analysis, sounds composed of two sounds produced from quite different instruments, namely a pipe organ and a carillon bell have been processed with several tools. The two sounds employed in analysis consist (1) of an organ tone followed by a bell, and (2) a bell followed by the organ. Two organ tones (C_2, C_3) have been played with a Quintadena 16′ stop of a historic organ.[6] The bell is part of the historic carillon of Bruges in Flanders.[7] For the fundamental frequencies and the prominent partials of the organ sounds, mind that the Quintadena stop is covered (*Gedackt*), and that a pipe length of 16′ means each tone played sounds one octave below the actual note name. Due to historic tuning (before a 'standard pitch' had been established), the fundamental of the C_2 played with the Quintadena 16′ is at $\sim$36 Hz, and C_3 is at $\sim$72 Hz, respectively.

The sound where the pipe organ starts at C_2 develops slowly in amplitude (Fig. 4) because few harmonic partials are actually excited in the covered pipe where excitation of modes and built-up of standing waves takes about 200 ms before the process is complete. Of the partials, the fundamental at 36 Hz is strongest in amplitude. An interval of 333 ms was chosen from the (measurable, barely audible!) onset of the pipe sound for the point where the bell sound starts (Fig. 5).

The bell sound, because of the excitation of the instrument by an impulse, builds up very fast with a considerable number of modes some of which are in a harmonic and others are in an inharmonic frequency ratio to the fundamental. The first second of sound (organ plus bell), if subjected to a standard Fourier spectral analysis, can be represented in a 3D-plot as in Fig. 6 which shows 20 spectra

[6] Historic organ of St. Bartholomäus, Mittelnkirchen, Altes Land, build by Arp Schnitger, Jacob Albrecht and Johann Matthias Schreiber 1688–1753. The Quintadena 16′ pipe rank is in the Hauptwerk of the organ.

[7] Built by Joris du Mery 1742–1748.

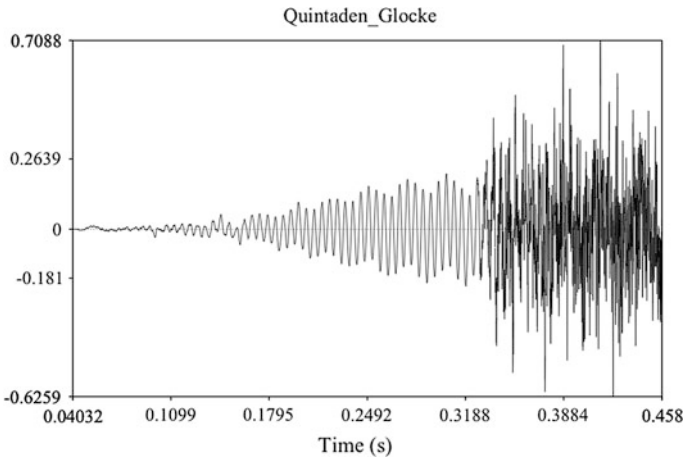


Fig. 5 Oscillogram of organ (Quintadena 16′, Pipe C_2) plus bell sound

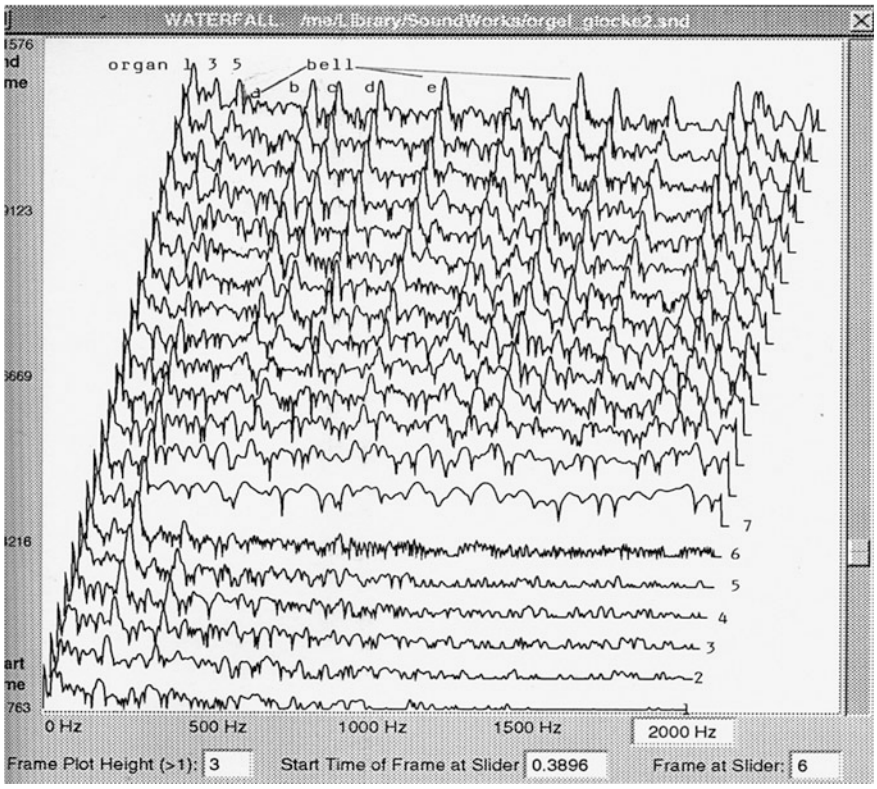


Fig. 6 3D-spectrogram of a complex sound (organ plus bell), 20 spectra

calculated with parameter settings for appropriate time and frequency resolution.[8] For readability, the frequency range displayed is 0–2 kHz though the bell sound contains spectral energy up to about 5 kHz. The 3D-plot, which covers about one second of sound, seems sufficient to study the evolution of two complex sounds that do have but little spectral overlap since the three most significant partials nos. 1, 3 and 5 of the organ sound have average frequencies of a 36, 112 and 181 Hz, respectively while the bell has its lowest partial (the so-called hum note) at about 208 Hz. From the 3D-plot, one can see that the organ sound except for the fundamental and partials nos. 3 and 5 (of which no. 3 has a long transient and comes into play not before spectrum no. 6) is quite noisy (air is streaming through the pipe before standing waves for more modes of vibration are established). Also, one can see that, with spectrum no. 7, the bell sound sets in, which is percussive and therefore has a fast buildup of modes of vibration and of corresponding spectral energy (the display is band-limited at 2 kHz for reasons of readability). The bell sound has a quite weak fundamental (the so-called hum, marked 'a' in the plot) at ca. 208 Hz yet a very regular spectrum typical of a minor third bell; in this sound, major components representing the prime, tierce, quint, and nominal (marked b, c, d, e in the plot) are found at ca. 411, 493, 627 and 829 Hz, respectively. It is evident that the bell sound carries significant energy from 400 Hz to the upper limit of the range on display, and that the 'watershed' that divides the organ sound and the bell sound in the spectrum is at about 200 Hz.

Given the two sounds have practically no spectral overlap they should be perceived as two separate objects (or as falling into two 'streams' in regard to auditory scene analysis, cf. Bregman 1990) as they excite different areas of the BM filter bank. This might support stream segregation as used for object identification along the auditory pathway. Moreover, the two sounds superimposed into one have different onsets in time as well as different attack features in regard to their wave shape and envelope. If processed by a filter bank that measures excitation of the BM per Bark (excitation per Bark [phon]; see Zwicker and Fastl 1999, Chap. 6), the analysis done with the Praat software (version 5323; Boersma and Weenink 2011) yields the following cochleagram (Fig. 7):

Since we know already from the FFT analysis presented in Fig. 6 that the organ sound has its energy concentrated at low frequencies, we find this distinctive feature also in the cochleagram where excitation at the onset is restricted to Bark bands 1–5. By contrast, the bell sound with many spectral components in the frequency band from about 400 Hz to 4.5 kHz mostly engages Bark bands 4–18. From its onset for an interval of ca. 150 ms the bell sound is so strong in energy that it masks the soft organ sound which, however, resurfaces later in the cochleagram (after time point 0.5 s) and becomes audible as such because many of the bell's higher partials have a fast decay so that the envelope of the bell sound

[8] FFT: 8192, Hanning, Hop ratio 0.25, zero pad factor 2.0. Analysis performed with Spectro 3.01 (Perry Cook, Gary Scavone).

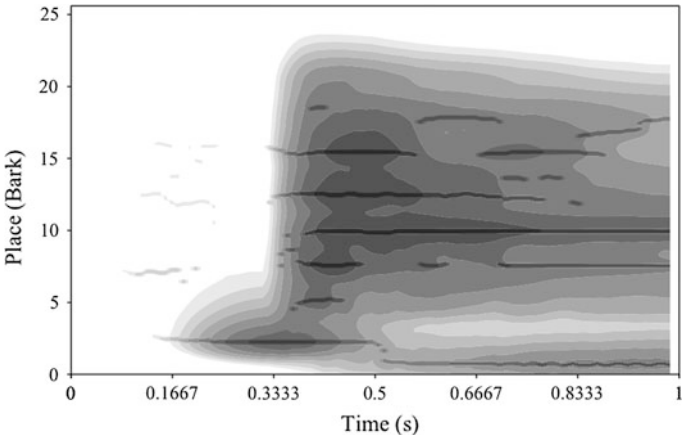


Fig. 7 Cochleagram of a complex sound containing organ and bell sound

shows a clear exponential decay (intensity [SPL dB] of the bell sound decays by ca. 8 dB in the first 500 ms, and by ca. 14 dB within a second from onset).

The purpose of presenting an analysis of the same sound performed with two different, if related tools is to underpin the usefulness of complementary methods where information obtained with one tool can help in interpreting output data generated with the other. In this way, one can often expand analyses by going into more details; in addition, applying different tools to the analysis of the same sound samples can help to minimize the risk of artifacts. To this end, two methods of analysis applied to another sound example will be evaluated in brief. We will analyze one sound played again with the Quintadena 16′ stop with two methods suited to achieve high resolution in time and frequency. One is autoregressive modeling (AR), the other is a complex-valued filter bank with the option of calculating the so-called instantaneous frequency for any sample point.

AR (see Marple 1987, Kostek 2005) is a family of methods developed for calculating spectral estimates for short or even very short segments of signals x(n) representing, for example, sound that may be transient or modulating in frequency and amplitude. For such sound segments usual Fourier techniques which are directed at frequency values for more or less steady-state sound signals may yield unclear results or even fail. In regard to DSP implementation suited to signal analysis, the AR approach rests on an all-pole filter model since the aim is to find such frequency bands in a signal where energy exists (see Marple 1987, Chap. 8). The transfer function of an AR model system (LTI = linear, time-invariant; cf. Bachmann 1992, Chap. 13) implemented as a recursive IIR filter can be given as

$$H(f) = \frac{1}{1 + \sum_{k=1}^{k} a_k \exp[-j2\pi fkT]} \tag{7}$$

The issue that makes AR techniques difficult is that one must choose a certain model as well as the order of the model (i.e., the number of poles in the complex z-plane). In practice, one must have some knowledge about properties of the signal to be analyzed in beforehand, or otherwise check various models and prediction order settings to find a good solution. 'Good' in this respect means the signal should neither be *under*analyzed (for this will lead to missing part of the relevant spectral information) nor *over*analyzed (which will result in spurious peaks in the spectra that do not represent energy at frequencies actually contained in the signal). Experimenting with various models (such as Burg, Autocorrelation, Covariance, Modified Covariance [ModCov]; see Marple 1987) and block lengths in processing sounds recorded from bells and harpsichords, Keiler et al. (2003) found that a stable analysis valid with respect to a mathematically defined signal which includes both FM and AM could be achieved best with the ModCov model (which did yield more precise and valid results than the Burg model at identical prediction orders and block sizes); a condition that must be met for stable AR analyses with ModCov is that the prediction order does not exceed a limit of 2/3 of the block length of samples used for analysis. Accordingly, AR[9] applied to the analysis of a transient organ sound produced by the pipe C_4 (in the Helmholtz system, this is c') of the Quintadena 16′ stop uses ModCov on a block of $N = 355$ samples with a prediction order of $p = 192$. For the analysis, a sequence of blocks was processed to yield data for one second of sound sampled at 16 bit/48 kHz. One should note that 355 samples at 48 kHz sampling mean $\sim$7.4 ms of the sound signal. The organ sound put to AR analysis is peculiar in that harmonic no. 7 audibly sets in first (what is a rather rare case for an organ pipe). The issue to be checked with AR analysis was (a) whether the auditory sensation is correct, and if so, (b) what the exact onset time as well as (c) the estimated frequency position for the partials might be as they appear in the sound one after another. The result of the AR analysis is shown in Fig. 8:

One can see that the fundamental is at ca. 143 Hz, and that only odd harmonics (1, 3, 5, 7, 9) are present with noticeable energy (as it should in a covered Quintadena pipe). Partial no. 7 indeed sets in first and builds up fast to a stable vibration (with a corresponding strong line in the AR spectrum marking its frequency at ca. 1015 Hz).[10] However, after ca. 250 ms, this mode of vibration starts to modulate (initially, almost in a periodic fashion) and then disintegrates. Conversely, the fundamental mode undergoes a transient phase of about 100 ms

[9] The code used for analysis was programmed in MatLab by Can Karadogan and Florian Keiler while working in the Department of Signal Processing and Communications of the Helmut-Schmidt-Universität Hamburg. The AR tool was developed to be used in a joint project directed at the study of transients in the sound of musical instruments (cf. Keiler et al. 2003).

[10] Fourier transforms of the steady-state part of the sound show that partial frequencies for higher harmonic partials are not exactly at integer ratios. Moreover, frequencies for partials including the fundamental fluctuate over time as can be seen from increasing values for variance of frequencies in longer FFT transforms (e.g., 65536). However, ACF analysis clearly gives a single 'pitch' for this pipe tone corresponding to 143 Hz.

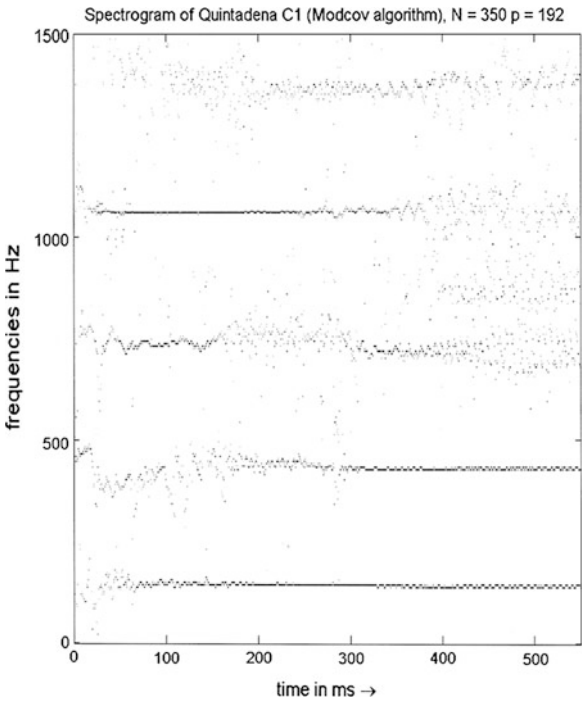


Fig. 8 Quintadena 16′, pipe/tone C_4, AR-Analysis (ModCov) 0–1.5 kHz

and then reaches a fairly stable regime of vibration (the frequency in the spectrum from then on shifts only slightly over time). Partial no. 3 is very unstable for about 200 ms and only after 300 ms begins to reach the harmonic frequency at ca. 431 Hz. Partial no. 5 sets in with a swing around the expected harmonic frequency of 715 Hz and after 150 ms disintegrates (not to recover within the time window of 520 ms under review). Partial no. 9 sets in weakly in a frequency range that is above the expected harmonic frequency range; after ca. 200 ms, this partial gets somewhat more stable for about 100 ms to undergo heavy modulation thereafter. The AR analysis indicates that partials 1, 3, 5, 7 set in almost at the same time, however, partial no. 7 indeed becomes audible first so prominently because it is the only partial for which a stable vibration and a corresponding frequency exist for at least 150 ms from onset.

Since reliability and validity of AR analyses are often difficult to assess (this holds true in particular for unknown types of signals where one must make assumptions as to the structure of the signal), it is always wise to check the results with another method. This has been done with a high-resolution filter bank making use of a complex-valued, quasi-continuous wavelet transform that offers calculation of instantaneous frequencies (Solbach et al. 1998). A complex-valued signal has the advantage that the instantaneous frequency can be determined for very

short segments (or even single sample points).[11] For the present analysis covering four octaves each of which was separated further into four bands in order to simulate the bandwidth of the auditory filter, a gammatone filter was used as mother wavelet. The gammatone filter is considered a good approximation to the human auditory filter (cf. Patterson et al. 1992) and has been implemented in many auditory models (see, e.g., Meddis and O'Mard 1997). For the gammatone filter defined in the time domain the impulse response is given as

$$g_\gamma(t) = \gamma(n, \lambda) \cdot \varepsilon(t) t^{n-1} \cdot e^{-\lambda t} \cdot \cos(2\pi f_0 t), \quad n \geq 1, \lambda > 0, \tag{8}$$

where n is the filter order, $\lambda > 0$ is the damping factor, f_0 is the center frequency of the filter, $\varepsilon(t)$ is the unit step function, and $\gamma(n, \lambda)$ is a normalization constant. For the present analysis, a 4th order IIR filter with a relative bandwidth of 0.05 is used. The upper limit frequency of analysis was set to 1600 Hz. The results of the analysis are displayed in Fig. 9. The frequency axis has logarithmic spacing (the distance between frequencies printed on the y-axis is 400 cents; ticks on the x-axis are at a distance of 100 ms):

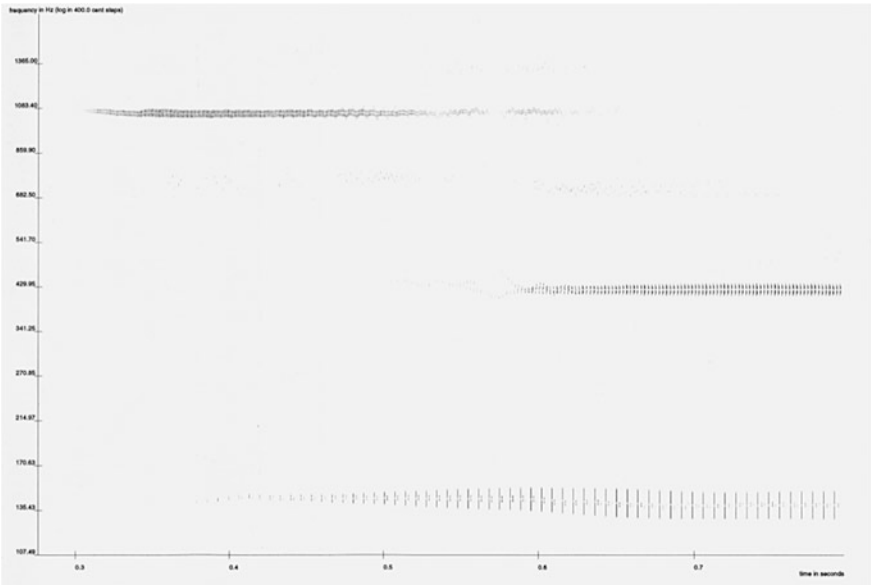


Fig. 9 Quintadena 16′, tone/pipe C_4, wavelet gammatone filter

[11] The usual approach (cf. Cohen 1995, 30ff., Flandrin 1999, 26ff.) is to calculate the so-called analytic signal by means of a Hilbert transform (Flandrin rightly calls the analytic signal a "complexified" signal).

The analysis clearly shows partial no. 7 to appear as a stable spectral component of definite pitch before the fundamental sets in weakly a hundred ms later fluctuating somewhat in frequency. Even more delayed is partial no. 3 which is 300 ms behind partial no. 7 yet quite stable in frequency. The wavelet analysis has been repeated with a Gaussian as mother wavelet for five octaves and twelve filter bands per octave; this fine-grain analysis detected partial no. 5 in addition to partials 1, 3 and 7. The two wavelet analyses are in good agreement with the AR analysis though the latter is even more detailed in very short signal segments while the wavelet analysis based on the gamma-tone filter might be closer to the actual behaviour of the auditory periphery (see below).

4 'Perceptually Adequate' Analysis and the Fourier-Time-Transform (FTT)

In the following, some fundamentals of psychoacoustics will be considered and compared to parameters found in DSP-based analysis and auditory modeling. The latter aims at a realistic 'emulation' of the auditory system in regard to basic functions and actual performance (cf. Meddis et al. 2010). Signal-analysis tools such as WT and FTT are less complex than full-grown auditory models (e.g. Meddis and Lopez-Poveda 2010), however, they can be viewed as representing the initial stage of BM filtering and thus are important as auditory 'preprocessors' (cf. Solbach et al. 1998; Terhardt 1998) that generate output used further in pitch and loudness perception as well as in auditory scene analysis. It should be underpinned that effective neural processing of complex sound naturally depends on the quality of (peripheral) BM filtering; the faster and the more precise this stage operates, the better neural processing along the auditory pathway can be achieved.

4.1 Frequency and Time Resolution; Discrimination and Recognition Tasks

The Fourier integral (see Bracewell 1978, Chap. 2; Meyer and Guicking 1974, 70ff) which is fundamental to Fourier analysis can be viewed as presenting a time function $x(t)$ in terms of frequency (or, rather, angular frequency ω). The Fourier integral considers frequency in an infinite interval ($-\infty \leq T \leq \infty$) and thus, as Gabor (1946, 431) has put it, *sub specie aeternitatis*. In musical signal analysis, however, one has to work with sounds that change over time, and often abruptly so. The answer to this situation was to consider applicability of Fourier theory to signals of definite length as well as to signals that lack clear periodicity and which are inharmonic in spectral composition. For practical reasons, techniques such as STFT (see Mertins 1996, Chap. 4, 1999, Chap. 7) were developed. The basic

concept for STFT is to multiply a sound signal $x(t)$ by an analysis window $g(t)$ and then compute the Fourier transform. For the analysis of a time signal, typically windows of length $N = 2^n$, $n = 8, 9,\ldots, k$ are chosen. If the signal to be analyzed is longer than N, the signal is processed frame by frame (with an overlap of 50 % or more to ensure continuity). Hence the window "slides" along the time axis by an amount defined by a shift parameter τ. The result thus obtained can be displayed in 2D or in (quasi) 3D-images such as Fig. 6 above. Though the STFT is regarded a good analysis tool that has been widely applied in acoustics and in particular in musical acoustics, it has a certain disadvantage in that conventional Fourier-transform algorithms operate on fixed values for N, which defines both Δf and Δt in a two-dimensional time/frequency plane (with f [Hz] as ordinate and t [ms] as the abscissa). Hence, time and frequency resolution are constant over the total bandwidth of analysis. In terms of Gabor's *logons* (see above), a uniform rectangle as "analysis box" results for low as well as for high frequency bands. An analysis window of constant length $N = 2^n$ samples applied to the full bandwidth of human auditory perception (ca. 25 Hz–16 kHz) seems unfortunate because our auditory system apparently needs a certain number of signal periods rather than a fixed time interval for pitch analysis (see below). Since the period duration T (ms) varies with frequency, the analysis window (either expressed in ms or in the number of samples) should be longer for low frequencies as compared to middle and high frequency bands.

In regard to temporal resolution relevant to hearing, a range of 'time constants' basic to temporal integration has been issued. It has been critically remarked that *"time* constants" *estimated from different experimental tasks range over three order of magnitude, from 250 to 200.000 μs* (Eddins and Green 1995, 207). In fact, there are different time constants relevant for different perceptual tasks as well as in regard to triggering motor responses, etc. In view of acuity achieved in discrimination tasks, minimum integration time in hearing appears to be 2–5 ms, depending to some extent on types of stimuli and conditions (see, e.g. Bilsen and Kievits 1989 who used so-called white flutter pulses). The data, which have been obtained in gap detection as well as in other experiments, are uneven (cf. Moore 2008, Chap. 5). Among relevant factors, time-intensity trades have to be taken into account (temporal integration depends on intensity or sound level; see Eddins and Green 1995). If minimum integration time of ca. 2–5 ms is interpreted in terms of response time of the auditory filter (as has been done), it appears that the response time perhaps plays a small role at low frequencies ($100 < f_{gr} < 500$ Hz) but not for frequencies above 1 kHz.

Other 'time constants' refer to noticeable asynchronies in the onset of the same tone played by two instruments (typical values seem to be $10 < t < 20$ ms), to "smearing" of several discrete echoes that occur in a room within a certain time span ($t < 50$ ms) into a sensation of quasi-continuous reverberation, and to temporal integration of energy in the sensation of loudness (most experimental data suggest an interval of $100 < t < 200$ ms). In regard to such 'time constants', one of course has to distinguish between discrimination and identification tasks, not to forget temporal organization of sound objects on a higher level such as grouping

and chunking in music cognition (see Snyder 2000). Discrimination for example in 2fc-experiments simply calls for responding if a certain 'event' did happen or not irrespective of what the informational 'content' of such an event may be. A very short pulse or noise burst will be sensed as a 'knack' but is not accessible for detailed auditory analysis. Even decisions subjects have to make whether a stimulus presented in a pair of sine tones is 'higher' or 'shorter' than the other (a design typical of experiments directed to difference limens for Δt and Δf relative to frequency bands) might just require a modicum of information on the side of the subject as to the nature of the stimuli. In contrast, identification of a stimulus in regard to one or several properties needs considerably more time since sound input that has been transformed into neural spike trains must be processed along several stages of the auditory pathway before, for example, a certain 'pitch' can be assigned to a stimulus. If one accepts periodicity detection and temporal processing for pitch as the predominant principle (notwithstanding significant evidence for rate-place representations and tonotopicity), the periods of time signals that might occur in musical sound are roughly from 33 ms (30 Hz) to 0.067 ms (15 kHz). Therefore, a maximum lag of 33 ms has been implemented in an ACF model suited to account for very low frequencies down to 30 Hz (Pressnitzer et al. 2001). In addition, time needed for arbitrary pitch estimates has been suggested as being 66 ms, with possibly less time down to about 40 ms or even 20 ms needed for such signals where subjects have a certain knowledge as to their likely pitch range in beforehand (cf. de Cheveigné 2005, 205). If 66 ms is a correct 'time constant', for most of musical relevant frequencies it would cover several or even many periods. In some early experiments, the time needed for developing a clear sensation of pitch for a sine tone varied from about 60–100 ms for very low frequencies (50 Hz) and ca. 30 ms for 300 Hz to about 15 ms for a frequency range of ca. 0.5–5 kHz (Bürck et al. 1935). From the empirical data as well as from considerations concerning the physics of the signal (that was switched on and off in an electronic circuit) and conditions of measurement, Bürck and colleagues calculated curves of tone recognition times as a function of frequency where about 80–100 ms would be required for a sine tone of 100 Hz but only ca. 5–10 ms for a sine tone in the range 1–5 kHz. Taking these approximate figures, one may hypothesize that pitch estimates for sine tones require about 5–8 periods of the time signal. The estimate figures mentioned above (to which several more from various experiments can be added) can be taken as tentative time constants in computational models of auditory perception.

In regard to frequency discrimination in hearing, for frequencies of two pure (sine or cosine) tones presented one after another, and with constant sound pressure level (SPL), the difference limen (DL) or just noticeable difference (jnd) has been estimated to be of the order of 1/30 of the Critical Bandwidth (CB). The concept of CB (see Moore 1995; Zwicker and Fastl 1999, Chap. 6) refers to BM excitation and filtering. From empirical data, a cochlear tonotopic frequency map has been proposed (cf. Greenberg 1990) where one CB corresponds to ca. 0.89 mm of BM. Hence, 1/30 of this unit would have to be considered as the jnd in regard to place theories of pitch and BM excitation patterns. However, one has to

see that hearing is a dynamic process based on feedback regulation and fast adaptation to stimulus conditions (otherwise, extremely sharp frequency discrimination as observed in trained musicians and very short recognition times for pitch and timbre of complex sounds would not be possible). Therefore, it seems only natural to see that center frequencies, bandwidths and shape of auditory filters (AF) vary with BM excitation level and bandwidth of input signals. Further, it is obvious that CB models such as have been proposed for loudness summation and place theories of pitch should be taken as a basic concept that must be validated with empirical data since a number of assumptions pertaining to CB models do not hold in a strict sense (cf. Moore 1995). Empirical data on CBs indicate that the Bark scale comprising 24 or 25 (in theory: non-overlapping) filter bands is not quite appropriate in particular for low frequencies ($f_c < 500$ Hz) since the bandwidth of the AF increases significantly with decreasing frequency. This effect is most prominent for $f_c < 200$ Hz (cf. Jurado and Moore 2010; Schneider and Tsatsishvili 2011). Compared to the Bark scale (cf. Zwicker and Fastl 1999), the so-called ERB scale (ERB = Equivalent Rectangular Bandwidth) comprising about 40 filter bands fits better to perceptual data though it does not fully account for pronounced increase of bandwidth at low frequencies. Each ERB is calculated by taking $4f_c/p$, where f_c is the center frequency and p is a filter parameter that determines the passband and the slope of the filter. In regard to modeling, the "effective bandwidth" for each AF along the BM depends on place and center frequency (that apparently is not fixed yet variable within a certain range), on sound level as well as on spectral energy distribution and spectral flux within audio signals. Very roughly, one can approximate CBs by 1/3 octave band pass filters. In reality, the "effective bandwidth" of AFs seems to vary from about one octave at very low frequencies to close to 250 cent around 1–3 kHz.

4.2 Wavelets and FTT

Wavelet analysis is one of several methods that have been developed to account for Gabor's *logon* concept and to provide equally good time and frequency resolution over the bandwidth of auditory perception. Wavelet analysis basically can be viewed as a Fourier approach where the window of analysis $g(t)$ is shifted in frequency by Ω_0, that is, multiplied in the time domain by $e^{i\Omega}0^t$. Similar to STFT, a sliding process along the time axis is part of the analysis with an increment of τ. Wavelet analysis (cf. Dutilleux et al. 1988) further includes a part equivalent to the 'window' $g(t)$, namely the analyzing wavelet $h(t) = e^{i\Omega}{}_0{}^t\, g(t)$ that is dilated in frequency by a parameter a so that

$$h^{(a,\tau)}(t) = \frac{1}{\sqrt{a}} h\left(\frac{t-\tau}{a}\right). \tag{9}$$

The wavelet transform (WT) of a continuous time signal $s(t)$ then is

$$W_h(\tau, a) = \frac{1}{\sqrt{a}} \int h\left(\frac{t-\tau}{a}\right) s(t)dt \tag{10}$$

The wavelet transform is computed by convolving the signal with a time-reversed and scaled wavelet (see Evangelista 1997). In regard to sound analysis, WT can be considered as a kind of band pass filter where the center frequency and the bandwidth of the filter can be varied by different values for the parameter a (cf. Mertins 1999, Chap. 9). In this respect, WT effectively computes a *constant-Q* filter analysis as has been employed in the gammatone filter analysis shown above (Fig. 9) where WT was performed for a frequency band of 0–1.6 kHz divided into four octaves each of which was subdivided into four bands of 250 cents to approximate the bandwidth of the auditory filter (AF) with respect to CB concepts.

A concept similar to STFT as well as to WT in certain respects is the Fourier-Time-Transform (FTT) as proposed by Terhardt 1985. In an article in which he considered properties of several different Fourier transforms, Terhardt argued that Fourier transforms are not restricted to periodic signals, and that the actual analysis window must not be identical with a period (or several periods) of a time signal $p(x)$ to yield valid spectral representations (a criterion to check validity of course is whether or not restoration of the time signal from the spectral data by an inverse transform can be achieved). Without going into details (many of which relate to linear systems theory rather than to "plain" spectral analysis), the argument put forward by Terhardt is that, for causal systems and signals, analysis of a physical signal such as sampled sound can be confined to time intervals from $t = 0$ to t so that the FTT for one-sided signals is given by

$$P(w, t) = \int_0^t p(x)e^{-wx}dx;\ \mathrm{t} > 0 \text{ and } \mathrm{w} = \mathrm{j}\,2\pi\mathrm{f} = \mathrm{j}\omega \tag{11}$$

The spectrum $P(w, t)$ for every instant t represents the time signal within a time interval that is defined as $-\infty < \mathrm{x} \leq t$. Also, $p(x) = 0$ for $x < 0$. For practical applications, signal values that are far in the past are of little relevance as to the current state of a system or signal[12]; therefore, the signal is multiplied by an exponential weighting function $\exp(-a(t - x))$ where $a \geq 0$ is a damping factor that can have values of 0–1. Consequently, with the exponential weighting included, Eq. (11) becomes

[12] The same consideration was made in "running" autocorrelation algorithms, which typically "slide" along a time signal and include a weighting function to successively discard past sample values so that ACF in fact is computed from an "effective time window" of N samples up to the sample point t moving with time. As to the equivalence of "running" ACF and FTT, see Terhardt 1998, 94f.

$$p(w,t) = \int_0^t p(x) e^{-a(t-x)}\, e^{-wx}\, dx;\ t > 0 \tag{12}$$

FTT applied to one-sided signals yields two parts, one steady-state and one transient (cf. Terhardt 1985, Eqs. 32 and 33)[13]; the transient part vanishes with ongoing time; also, amplitude density distribution narrows with time passing, and approaches a steady-state bandwidth of $\Delta\omega = a$ (3 dB cutoff frequency). After signal onset, the steady-state is reached at about $t = 1/a$ ($1/a$ is also the time constant of the exponential weighting). The damping factor a can be employed to control the steady-state bandwidth (that can be narrowed, however at the cost that the time needed to attain the steady-state proportionally increases). For simple cosine signals of sufficiently high frequency, the FTT magnitude spectrum according to Terhardt (1985, 254) is *largely similar to the output of a simple-resonance filter* for which the 3 dB bandwidth is $B = a/\pi$. Given that the boundary between transient part and steady-state part can be taken as the "effective time window" of the analysis defined by $1/a$, the product of the effective time window and the steady-state bandwidth would be as small as $1/\pi = 0.3183$.

If this product would be viewed in terms of the uncertainty relation in regard to signals and systems, it would clearly be far below Gabor's theoretical limit of $\Delta f\ \Delta t = 1/2$. In this context, it might be noted that, for signals of given (rms) duration and energy (set to a value of 1), the uncertainty product has been calculated by Papoulis (1962, 62f., Eqs 4-39–4-46) as

$$D_t * D_\omega \geq \sqrt{\frac{\pi}{2}} \tag{13}$$

where the equality holds for Gaussian signals (i.e., the product numerically yields 1.2533). The difference between products $\Delta f\ \Delta t \geq 1$ (Eq. 2) postulated from mathematical analysis and values much smaller than 1 calculated for FTT and other filter models results from the 3 dB bandwidth parameter, which is common to filter design and performance tests yet must not necessarily apply to auditory perception. The bandwidth of the AF as determined in hearing experiments involving subjects of different age (Patterson et al. 1982) can be roughly given as 11 % of the center frequency for young adults who have not yet suffered hearing loss. For a f_c of 0.5, 2 and 4 kHz (as were employed in the experiments of Patterson et al. 1982), this means a relative filter bandwidth of ca. 191 cents (corresponding to the musical interval of a major second). Alternatively, the normalized width of the equivalent rectangular filter (roex[p, r]) has been given as $\mathrm{BW}_{\mathrm{ER/fc}} = 4/25 = 0.16$ (Patterson et al. 1982, 1801).

In FTT analysis, parameter values for bandwidth B and damping factor a can be set so as to simulate performance of the auditory periphery. To this end, the bandwidth should be that of the CB (cf. Zwicker and Fastl 1999, Chap. 6) divided

[13] A more detailed analytic formulation of the FTT is given by Mummert 1997.

by 25, which would not be too far away from the jnd for pure tones[14] Referring to analytical expressions designed to approximate critical-band rate and critical bandwidth (Zwicker and Terhardt 1980), Terhardt suggested that an "audio FTT" could be performed with the parameters set like

$$B = a/\pi = 1 + 3\left(1 + 1.4(f/\text{kHz})^2\right)^{0.69}\text{Hz} \tag{14}$$

Assuming that there are 24 CBs (expressed as a Bark scale), the frequency resolution for the FTT is $24 \times 25 = 600$ frequency samples per spectrum deemed sufficient and necessary to model peripheral auditory analysis (cf. Terhardt 1985, 255). In regard to the effective window length (i.e., the analysis interval T_A) relative to frequency bands, Terhardt (1992, 378) has given these figures:

f/kHz	0.1	0.5	1	2	4	8
TA/ms	24	22	16	8	2.7	0.74

Numerically, for a sampling rate at 44.1 kHz, an effective window length of 24 ms would correspond to 1058 samples falling into this time interval. A cosine signal of $f = 0.1$ kHz and a period of 10 ms would cover 441 samples per period so that the analysis interval will have access to, on the average (as the analysis window slides along the time signal), two periods of the signal. The ratio is much better at higher signal frequencies and shorter periods where the analysis window would hold (at best, if no truncation occurs) 16 periods at 1 kHz as well as at 2 kHz. The effective window length of the FTT has been calculated (Vormann and Weber 1995, 1191) as

$$T(\omega) = 2.988/a(\omega) \tag{15}$$

where $a(\omega)$ is the frequency-dependent transformation parameter. Correspondingly, the bandwidth is given as

$$B(\omega) = \frac{\sqrt{\sqrt{2}-1}}{\pi} \cdot a(\omega) \tag{16}$$

whereby an uncertainty product $T \times B \approx 0.61$ has been calculated. This of course would outperform a conventional Fourier transform analysis by far so that time/frequency resolution close to the cochlear filter bank can be expected from the FTT analysis (see below). In some of the relevant publications (Heldmann 1993; Vormann 1995), values as to T and B as well as to their product differ somewhat; parameter values as found in the literature for the 1st and 2nd order as well as estimates for the 4th order are given in Table 1:

[14] For example, one CB included in the table given by Zwicker and Terhardt (1980, p. 152), ranges from 920 to 1080 Hz with $f_c = 1000$ Hz and is 160 Hz wide; divided by 25, the frequency step would be 160/25 = 6.4 Hz as compared to the jnd at 1000 Hz, which is ca. 3 Hz.

Table 1 FTT parameters

Order	1	2	4
Window function	e^ − ax	x e^ − ax	$x^3/6$ e^ − ax
Resolution *dT*	1/a	2.988/a	4.990/a
Bandwidth (*B*)	a/π	0.6436 a/π	0.4350 a/π
dT * *B*	1/π	1.923/π	2.171/π

In this table, a denotes the scaling factor $a(\omega)$, and t denotes the time axis. For practical reasons, parameter values may be rounded like

Order	1	2	4
Window function	e^{-at}	$t \cdot e^{-at}$	$\frac{t^3}{6} \cdot e^{-at}$
dT	$1/a$	$3/a$	$5/a$
B	a/π	0.644 a/π	0.435 a/π
dT * *B*	$1/\pi \approx 0.32$	$1.93/\pi \approx 0{,}61$	$2.17/\pi \approx 0{,}69$

The bandwidth B for any order of analysis n can be calculated according to

$$B = \frac{a}{\pi}\sqrt{2^{\frac{1}{n}} - 1} \tag{17}$$

The original FTT algorithm (see Terhardt 1985) has been improved later on in regard to the weighting function (cf. Schlang and Mummert 1990, Terhardt 1998, 97) where a form $a\, t\, e^{-at}$ has been proposed. Also, weighting of the form $h(t) = t^3 e^{-at}$ has been introduced for a 4th order FTT (as $h(t)$ in this case is equivalent to the Laplace transform of a 4th order low-pass filter, see von Rücker 1997).

For comparison of conventional Fourier transform and FTT analysis, a number of natural sounds were chosen; in addition some complex sounds based on FM and AM processes were generated with Mathematica. In the following, the results for the organ sound (Quintadena 16′, pipe/note C2) on which a bell sound has been superimposed (see Figs. 5–7) will be presented.

In the FTT algorithm applied to analysis, a 4th order weighting function had been implemented. Since the effective time window for the standard FTT has been given as 24 ms at 0.1 kHz, corresponding to 1058 samples at 44.1 kHz sampling (see above), a comparison to an FFT of 1024 sample points seems a reasonable choice. However, the FFT also employed a weighting function for which a Blackman window was chosen.[15]

The analysis obtained with a FFT of 1024 and Blackman weighting is shown in Fig. 10:

The same sound subjected to 4th order FTT analysis is displayed in Fig. 11:

[15] The ENBW for the Blackman window is 1.73 bins in DFT and the 3.0 dB bandwidth is 1.68 bins.

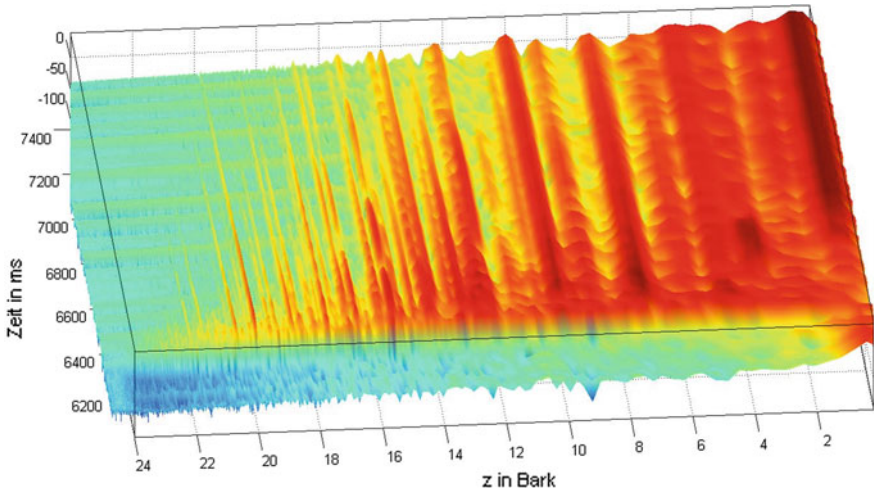


Fig. 10 Organ (Quintadena 16′ C_2) plus bell, FFT 1,024 pts, Blackman

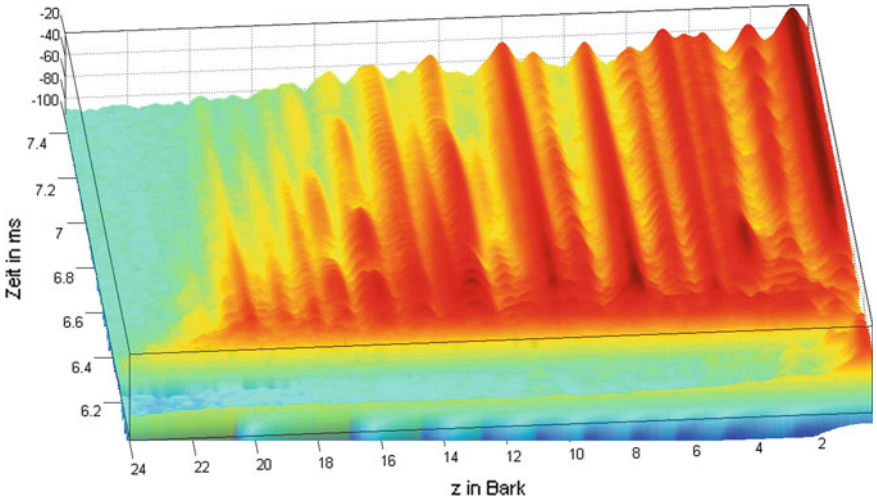


Fig. 11 Organ (Quintadena 16′ C_2) plus bell, 4th order FTT

From a comparison of both analyses presented as 3D-plots (were the abscissa [x] is in Bark[z], the ordinate [y] is in dB, and time (ms) is in the z-dimension) one can see that time and frequency resolution for the FTT at low frequencies is considerably better than with the 1024 point FFT subjected to Blackman weighting. Note that with a FFT length of $N = 1024$ and sampling at 44.1 kHz, frequency resolution (Eq. 3) nominally is ca. 43 Hz. As this is the constant bandwidth of the FFT

analysis (a DFT can be viewed as equivalent to a filter bank), the signal is under a fine-grain analysis at higher frequencies (Bark[z] 10–20) so that the FFT analysis picks many small spectral components corresponding to higher modes of vibration of the bell while the FTT analysis is more condensed since it relates to the concept of CBs, and therefore integrates such components which are closely spaced in frequency into broader "spectral ridges" (Fig. 11). A similar picture would be obtained with a WT-based analysis. One can argue that auditory perception of complex sounds basically is directed at picking spectral peaks that are present during a reasonable time interval (relevant as 'integration constant' in regard to hearing). In this respect, a limited number of clearly expressed "spectral ridges" may be more relevant to actual hearing as this must be performed in quasi-real time, and consequently calls for some temporal as well as spectral integration (as reflected in CBs and 'integration constants'). Algorithms directed to finding peaks in spectral envelopes are quite common as in LPC (see Fig. 4) or similar source-filter analysis models (cf. Rodet and Schwarz 2007); if a sequence of frames is processed so that spectral envelope peaks can be separated and extracted, the next step is to connect such peaks from one frame to the next so that 'tracks' for harmonic partials or inharmonic components result over time. Such tracks then can be used for finding quasi-continuous pitch contours or for separation of 'sound objects' in a computational auditory scene approach (cf. Kostek 2005).

Comparison of the two types of analysis ("plain" Fourier, FTT) may indicate an advantage on the side of the FTT as one would expect from uncertainty products reported in the literature. However, the difference obtained in several analyses (of which but one example is included in the present article) seems gradual rather than principal. To optimize analysis, one often has to experiment with parameter settings. In addition, it is always revealing to apply different methods and models to the analysis of particular sound samples because in this way one can try to extract as many distinctive features as is needed for a certain problem, and at the same time the results obtained with one method can be tested for validity and reliability by using a second or even a third tool.

As far as 'perceptually adequate' analysis is concerned, comparison of several models including Gabor filtering, a linear, simplified but functional cochlear model (first published by Netten and Duifhuis 1983), WT and gammatone filtering tested for their impulse responses resulted in kind of a ranking (Hut et al. 2006) where Gabor filtering was leading in regard to the uncertainty product, but also the linear cochlea model performed well. WT was judged to be unsuited to auditory modeling because an 'auditory wavelet' would not exist, and, therefore, Hut et al. (2006, 633) concluded that *wavelet analysis methods cannot be used in perception research*. The gammatone filter (implemented in many auditory models) according to these tests did well in terms of *general purpose linear time–frequency filtering, but does not give a good cochlear representation* (Hut et al. 2006, 635). Since an advanced cochlear model (Mammano and Nobili 1993; Nobili and Mammano 1999) seems to provide extremely good resolution in both time and frequency

(Russo et al. 2011) with $\Delta f\,\Delta t \approx 0.55$, and hence close to the Gabor limit of 0.5, this approach perhaps could be the most promising to approximate performance of the auditory system even further (for recent developments, see Meddis et al. 2010). It should be noted, in this respect, that known values for the 'uncertainty relation' have been questioned to hold for the human auditory system (see, e.g. Kral and Majérnik 1996). The reason for such an assessment based on empirical data in most cases was that the performance of the auditory system in discrimination tasks (where stimuli were varied in frequency, level, and duration) was better than accepted values for the 'uncertainty product', on the one hand, and the relation between band-width and duration apparently was not linear, on the other. An explanation for this system behaviour can be found on the level of functional neuroanatomy and neurophysiology since hearing is effected by a complex network involving ascending and descending pathways as well as feedback regulation loops (as in OHC motility and BM/TM adjustment necessary for sharp frequency discrimination and 'pitch' processing; OHC = outer hair cell, BM = basilar membrane, TM = tectorial membrane; see Pickles 2008).

5 Conclusion

The present article intends to shed light on several approaches to digital sound analysis that are viewed (a) as tools useful for research in musical acoustics and organology, and (b) in regard to auditory perception. Besides the proven Fourier analysis techniques such as STFT, especially for the study of transient or impulsive sounds other methods such as WT (see Zhu and Kim 2006) or AR can be applied for time/frequency representations. To account for characteristics of the auditory systems, namely different resolution power relative to the period length (ms) of nearly periodic as well as quasi-periodic sound signals (meaning spectral structures ranging from harmonic to inharmonic; see Schneider 1997, 2001), algorithms simulating peripheral filtering must be designed which offer appropriate filter bandwidth and time constants. WT and gammatone filter banks are among such algorithms that can be applied to many sounds, and can thus be considered versatile tools. If an approach is needed which is closer to functions found implemented in the auditory system, computational models such as developed by Meddis and O'Mard (1997, 2006) should be applied to the study of musical sound in regard to psychoacoustics and perception (see Schneider and Frieler 2009). The FTT model that was proposed already in 1985 still can be a useful method for time/frequency analysis that is close to basic parameters of the auditory periphery.

References

Bachmann, W. (1992). *Signalanalyse. Grundlagen und mathematische Verfahren*. Braunschweig: Vieweg.

Beauchamp, James. (2007). Analysis and synthesis of musical instrument sounds. In J. Beauchamp (Ed.), *Analysis, synthesis, and perception of musical sounds* (pp. 1–89). New York: Springer.

Bilsen, F. & Kievits, I. (1989). The minimum integration time of the auditory system. Preprint 2746, AES Convention Hamburg March 1989.

Boersma, P. & Weenink, D. (2011). *Praat*. Doing phonetics by computer (version 5232). Amsterdam: University of Amsterdam, Institute of Phonetics.

Boersma, P. (1993). *Accurate short-term Analysis of the fundamental frequency and the harmonic-to-noise ratio of a sampled sound*: *Proceedins of Institute of Phonetics, University of Amsterdam* (Vol. 17 pp. 97–110).

Bracewell, R. (1978). *Fourier transform* (2nd ed.). New York: McGraw-Hill.

Bregman, A. (1990). *Auditory scene analysis*. Cambridge: MIT Press.

Bürck, W., Kotowski, P. & Lichte, H. (1935). Der Aufbau des Tonhöhenbewußtseins. *Elektrische Nachrichtentechnik*, *12*, 326–333.

Cohen, L. (1995). *Time-frequency analysis*. Upper Saddle River, N.J.: Prentice-Hall.

de Boer, E. (1976). On the "residue" and auditory pitch perception. In W. D. Keidel & W. D. Neff (Eds.), *Handbook of sensory physiology* (Vol. 3, pp. 479–583). New York: Springer.

de Cheveigné, A. (2005). Pitch perception models. In C. Plack, A. Oxenham, R. Fay, A. Popper (Eds.). *Pitch. neural coding and perception* (pp. 169–230). New York: Springer.

DeFatta, D., J. Lucas, & Hodgkiss, W. (1988). *Digital signal processing. A system design approach*. New York: Wiley.

Dellomo, M., & Jacyna, G. (1991). Wigner transforms, Gabor coefficients, and Weyl-Heisenberg wavelets. *Journal of Acoustical Society of America, 89*, 2355–2361.

Dutilleux, P., Grossmann A. & Kronland-Martinet, R. (1988). Application of the wavelet transform to the analysis, transformation and synthesis of musical sound. Preprint 2727, AES Convention 85, November 1988.

Eddins, D., & Green, D. (1995). Temporal integration and temporal resolution. In B. C. J. Moore (Ed.), *Hearing* (pp. 207–242). San Diego: Academic Press.

Evangelista, G. (1997). Wavelet representations of musical signals. In C. Roads, St. Pope, A. Piccialli, G. de Poli (Eds.), *Musical signal processing* (pp. 127–153). Lisse: Swets and Zeitlinger.

Flandrin, P. (1999). *Time-Frequency/Time-Scale Analysis*. San Diego: Academic Press.

Gabor, D. (1946). Theory of communication. *Journal of Institution of Electrical Engineering, 93*, 429–457.

Gafori, F. (1496/1967/1968). *Practica Musicae*. Milan (Reprint Farnborough, Hants.: Gregg Pr. 1967); [English translation and transcription of musical examples by Clement Miller]. American Institute of Musicology 1968).

Greenwood, D. (1990). A cochlear frequency-position function for several species—29 years later. *Journal of Acoustical Society of America, 87*, 2592–2605.

Heldmann, K. (1993). Wahrnehmung, gehörgerechte Analyse und Merkmalsextraktion technischer Schalle. Ph.D. thesis, Technical University of Munich.

Hut, R., Boone, M., & Gisolf, A. (2006). Cochlear modeling as time-frequency analysis tool. *Acustica, 92*, 629–636.

Jurado, C., & M, Brian. (2010). Frequency selectivity for frequencies below 100 Hz: Comparison with mid-frequencies. *Journal of Acoustical Society of America, 128*, 3585–3596.

Keiler, F., Karadogan, C., Zölzer, U. & Schneider, A. (2003). *Analysis of transient musical sounds by auto-regressive modeling*: *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)* (pp. 301–304). London: St. Marys.

Kostek, B. (2005). *Perception-based data processing in acoustics*. Berlin: Springer.

Kral, A., & Majérnik, V. (1996). Neural networks simulating the frequency discrimination of hearing for non-stationary short tone stimuli. *Biological Cybernetics, 74*, 359–366.
Küpfmüller, K. (1968). *Die Systemtheorie der elektrischen Nachrichtenübertragung* (3rd ed.). Stuttgart: Hirzel.
Mammano, F., & Nobili, R. (1993). Biophysics of the cochlea: Linear approximation. *Journal of Acoustical Society of America, 93*, 3320–3332.
Markel, J., & Gray, A. (1976). *Linear prediction of speech*. Berlin: Springer.
Marple, S. L. (1987). *Digital spectral analysis*. Englewood Cliffs, N.J.: Prentice-Hall.
Meddis, R., & O'Mard, L. (1997). A unitary model of pitch perception. *Journal of Acoustical Society of America, 102*, 1811–1820.
Meddis, R., & O'Mard, L. (2006). Virtual pitch in a computational physiological model. *Journal of Acoustical Society of America, 120*, 3861–3869.
Meddis, R. & Lopez-Poveda, E. (2010). Auditory periphery: From pinna to auditory nerve. In R. Meddis et al. (Eds.), *Computational models of the auditory system* (pp. 7–38). New York: Springer.
Meddis, R., Lopez-Poveda, E., Fay, R., & Popper, A. (Eds.). (2010). *Computational models of the auditory system*. New York: Springer.
Messner, G. (2011). Du krächzt wie ein Rabe…, singst wie eine Nachtigall…In A. Schmidhofer, St. Jena (Eds.), *Klangfarbe. Vergleichend-systematische und musikhistorische Perspektiven*. Frankfurt/M.: P. Lang, pp. 205–217 (plus sound examples on a CD in the book).
Mertins, A. (1996). *Signaltheorie*. Stuttgart: Teubner.
Mertins, A. (1999). *Signal analysis*. Chichester: Wiley.
Meyer, E., & Guicking, D. (1974). *Schwingungslehre*. Braunschweig: Vieweg.
Momose, H. (1991). *Sonogram*. Davis, CA: University of Cal.
Moore, B. (1995). Frequency analysis and masking. In B. Moore (Ed.), *Hearing* (pp. 161–205). San Diego: Academic Press.
Moore, B. (2008). *An introduction to the psychology of hearing* (5th ed.). Bingley: Emerald.
Mummert, M. (1997). *Sprachcodierung durch Konturierung eines gehörangepaßten Spektrogramms und ihre Anwendung zur Datenreduktion*. Ph.D. thesis, Technical University of Munich.
Netten, S., & Duifhuis, H. (1983). Modelling an active, nonlinear cochlea. In E. de Boer & M. Viergever (Eds.), *Mechanics of Hearing*. Delft: Delft University Pr., 143–151.
Nobili, R., & Mammano, F. (1999). Biophysics of the cochlea II: Stationary nonlinear phenomenology. *Journal of Acoustical Society of America, 99*, 2244–2255.
Oertel, D., Fay, R., & Popper, A. (Eds.). (2002). *Integrative functions in the mammalian auditory pathway*. New York: Springer.
Papoulis, A. (1962). *The Fourier Integral and its applications*. New York: McGraw-Hill.
Patterson, R., Nimmo-Smith, I., Weber, D., & Milroy, R. (1982). The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *Journal of the Acoustical Society of America, 72*, 1788–1803.
Patterson, R., Robinson, K., Holdsworth, J., McMcKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. *Advances in the Biosciences, 83*, 429–443.
Pickles, Ja. (2008). *An Introduction the Physiology of Hearing* (3rd ed.). Bingley: Emerald.
Pressnitzer, D., Patterson, R., & Krumbholz, K. (2001). The lower limit of melodic pitch. *Journal of the Acoustical Society of America, 109*, 2074–2084.
Rodet, X., & Schwarz, D. (2007). Spectral envelopes and additive+residual analysis/synthesis. In J. Beauchamp (Ed.), *Analysis, Synthesis, and Perception of Musical Sounds* (pp. 174–227). New York: Springer.
Rossing, T. (1982). *The Science of Sound*. CA: Addison—Wesley.
Rücker, C. (1997). Berechnung von Erregungsverteilungen aus FTT-Spektren. *Fortschritte der Akustik—DAGA 1997*, pp. 484–485.
Russo, M., Rožić, N., & Stella, M. (2011). Biophysical cochlear model: Time-frequency analysis and signal reconstruction. *Acustica, 97*, 632–640.

Schlang, M. & Mummert, M. (1990). Die Bedeutung der Fensterfunktion für die Fourier-t-Transformation als gehörgerechte Spektralanalyse. *Fortschritte der Akustik*, *DAGA 1990*, Bad Honnef 1990, pp. 1043–1046.

Schneider, A. (1997). *Tonhöhe, Skala, Klang. Akustische, tonometrische und psychoakustische Studien auf vergleichender Grundlage*. Bonn: Orpheus-Verlag für Syst. Musikwiss.

Schneider, A. (2001). Complex inharmonic sounds, perceptual ambiguity, and musical imagery. In R. I. Godøy & H. Jørgensen (Eds.), *Musical imagery* (pp. 95–116). Lisse: Swets and Zeitlinger.

Schneider, A. & Frieler, K. (2009). Perception of harmonic and inharmonic sounds: Results from ear models. In S. Ystad, R. Kronland-Martinet & K. Jensen (Eds.), *Computer music modeling and retrieval. Genesis of meaning in sound and music* (pp. 18–44). Berlin: Springer.

Schneider, A., von Ruschkowski, A., & Bader, R. (2009). Klangliche Rauhigkeit, ihre Wahrnehmung und Messung. In R. Bader (Ed.), *Musical acoustics, neurocognition and psychology of music* (pp. 103–148). Frankfurt: P. Lang.

Schneider, A., & Tsatsishvili, V. (2011). Perception of musical intervals at very low frequencies: Some experimental findings. In A. Schneider & A. von Ruschkowski (Eds.), *Systematic musicology: Empirical and theoretical studies* (pp. 99–125). Frankfurt: P. Lang.

Solbach, L., Wöhrmann, R., & Kliewer, J. (1998). The complex-valued continuous wavelet transform as a preprocessor for auditory scene analysis. In D. F. Rosenthal & H. G. Okuno (Eds.), *Computational auditory scene analysis* (pp. 273–292). Mahwah, N.J.: Erlbaum.

Snyder, B. (2000). *Music and memory*. Cambridge, MA: MIT Press.

Terhardt, E. (1985). Fourier transformation of time signals: Conceptual revision. *Acustica, 57*, 242–256.

Terhardt, E. (1992). From Speech to language: on auditory information processing. In M.E.H. Schouten (Ed.). *The Auditory Processing of Speech. From Sounds to Words* (pp. 363-380). New York: Mouton de Gruyter.

Terhardt, E. (1998). *Akustische Kommunikation*. Berlin: Springer.

Vormann, M. (1995). *Psychoakustische Modellierung der virtuellen Tonhöhe*. Diploma thesis (Physics), Carl von Ossietzky University, Oldenburg.

Vormann, M. & Weber, R. (1995). Gehörgerechte Darstellung von instationären Umweltgeräuschen mittels Fourier-Time-Transformation (FTT). *Fortschritte der Akustik—DAGA 1995*, pp. 1191–1194.

Winer, J., & Schreiner, C. (Eds.). (2011). *The Auditory Cortex*. New York: Springer.

Yen, N. (1987). Time and frequency representation of acoustic signals by means of the wigner distribution function: Implementation and interpretation. *Journal of the Acoustical Society of America, 81*, 1841–1850.

Zhu, X., & Kim, J. (2006). Application of analytic wavelet transform to analysis of highly impulsive noises. *Journal of Sound and Vibration, 294*, 841–855.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth. *Journal of Acoustical Society of America, 68*, 1523–1525.

Zwicker, E., & Fastl, H. (1999). *Psychoacoustics. Facts and models* (2nd ed.). Berlin: Springer.

Performance Controller for Physical Modelling FPGA Sound Synthesis of Musical Instruments

Florian Pfeifle and Rolf Bader

1 Introduction

In Musical Acoustics, realising real-time solutions for Physical Models is a hot topic. The sound quality and variability of musical instrument sounds calculated by whole-body formulations will be strongly pushing this field towards a variety of different solutions in the near future. As the session of real-time implementations of Physical Modelling in Musical Acoustics at the Joint Meeting of the American Acoustical Society with the European Acoustical Society in Paris 2008 showed (Smith 2008), many attempts are discussed, yet none of them is working with satisfactory results in real-time. Yet, the first working solution of such a real-time implementation is by using a Field-Programmable Gate Array (FPGA) hardware with a Finite-Difference Implementation, like presented for the Banjo (Pfeifle and Bader 2009), the violin (Pfeifle and Bader 2011), or other musical instruments (Pfeifle and Bader 2011, 2012). The advantage of using a Finite-Difference scheme compared to e.g. a Finite-Element solution is that high frequencies are represented much more realistically. Still, Physical Models using Finite-Differences with whole-body implementations are far from real-time on a standard Personal Computer. On the other hand, simplifications of the model to regain real-time performance is not satisfying if one wants to simulate physically accurate models. So e.g. reducing the number of Degrees of Freedom (DOF) in these models does effect the higher partials of sounds considerably. Also changing the instrument shapes, e.g. through spatial distortion, is too time consuming to work in real-time. Reducing the complexity of instruments in terms of geometrical simplifications clearly leads to simplified solutions (Smith 2008). Also, these models can scarcely be varied in terms of the geometrical fine structure of the instruments and therefore in modelling the constituent parts crucial for musical expression and articulation.

F. Pfeifle (✉) · R. Bader
Institute of Musicology, University of Hamburg, Neue Rabenstr. 13, 20354 Hamburg, Germany
e-mail: fkra003@uni-hamburg.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_14,

But, using a FPGA hardware, such complex models can be formulated and solved in real-time for instantaneous sound production. This chapter introduces the basic idea of physical models and their FPGA implementation. Additionally we present a real-time control implementation for interacting and manipulating the models on the FPGA. With this, arbitrary parameters of the model, like material constants, the geometry, etc. as well as performance parameters for playing the instrument can be changed in real-time. This allows for direct interaction with the instrument model in a musical setting or an immediate auralisation of parameter changes.

1.1 Physical Modelling

Physical Modelling of Musical Instruments has become a well-established technique when researching important properties for musical performance or instrument building (Bader and Hansen 2008). It has been applied to many instruments. With guitars it could be found, that the coupling of bending and in-plane waves in the guitar body plays a crucial part in the radiated sound (Bader 2005a, b). Also the need to build the guitar back plate under tension to enhance the sound's brightness could be shown. When systematic changes are applied to the thickness of top plate, back plates, rims, fan bracing, and ribs, the change in brightness of the sound radiated from the different parts do sometimes show a linear but mostly a nonlinear behaviour, caused by the complex coupling of these parts. So e.g. if the fan bracing is added to the top plate and so reduces the flexibility of the plate, resulting in reduced radiation energy of the higher harmonics, this stiffer top plate is able to transport the higher modes to the neck much stronger. This again leads to a stronger radiation of the high frequencies with increasing fan bracing thickness (Bader 2008b). Other investigations show the possibility to model the guitar employing modal synthesis with respect to the top plate only (Bécache et al. 2005). The precision of modelling is shown in (Elejabarrieta et al. 2002), where a Finite-Element Analysis of a guitar top plate is compared to a real top plate through different stages of construction.

Other instruments have been investigated in a similar way, too. The pianos' sound board has been modelled in (Giordano 2006), where the fan bracing and the stiffness of the plate play crucial roles in the overall sound behaviour. The kantele, a finish dulcimer, is investigated in (Erkut et al. 2002). Also a labium is studied in a similar way, taking fluid dynamics and turbulence into consideration. It could be shown, that the turbulent damping of the flute is crucial for its impedance, where only because of this damping of air vortices in the flute at the embouchure hole causes the flow energy into the tube only to be about 3 % of the whole blowing energy (Bader 2005d). With saxophones, the role of the thickness distribution of the reed shape could be shown to produce the sound expected by different reeds of a commercial reed producer (Bader 2008a). Other investigations concerning the reed show a very similar air vortex distribution in the mouth piece (Da Silva et al. 2007).

Percussion instruments of many kinds have also widely been studied using Physical Modelling techniques. Bells show complex behaviour within their

transient phase as well as in terms of damping or radiation directivity (Schoofs 2002; Lau et al. 2004). Optimization algorithms have been added to a Finite-Element calculation for bells to find perfect shapes for the bell (Özakča 2004). Other Percussion Instruments have been investigated, too, like a round drum (Essl and Cook 2000) discussing travelling waves along the rim of the instrument important for the transient sound behaviour. In a model/measurement comparison of the complete bass drum, coupling the drums membranes to the wooden box and the inclosed air, it could be shown, that the higher frequencies are radiated from the wooden shell showing its importance in sound production (Bader 2006a). When investigating a Balinese percussion instrument, the gender dasa played in a gamelan orchestra, it could be shown, that the trapezoid shape of the plates is producing additional modes through scattering of waves during the initial transient phase of the sound (Bader 2009, 2004). This effect is so substantial that it can be termed as the responsible mechanism producing the overall sound quality of the instrument.

Violins have been studied using whole body Finite-Difference Methods also, implementing the string-bow interactions with changing bowing pressure and velocity (Bader 2005c). Here, also the changes in thickness of different body parts show a complex interaction structure, meaning, that linear changes in the geometry of the violin body mostly results in nonlinear changes in the radiated sound. This effect is also confirmed by investigations of the whole string instrument family, most prominently in Bissingers research of a Hutchings-Schelleng violin octet (Bissinger 2003).

Summarising these investigations, Musical Instruments do show properties in their sound design which can only be treated when examining the fine structure of the instrument geometry. These fine structural elements can be the precise shape of the instrument body, coupling between different body parts, thickness distribution of plates like violin top plates, the shape of fan bracing common with guitars, the distribution of turbulence during blowing, or the thickness distribution along reeds with saxophones or clarinets. All these parameters influence the overall sound and timbre in such a way that they can be responsible for the subtle differences that distinguishes one piece from another. So one can say that these fine structures are responsible for differences in musical expressivity and articulation in musical performance, and in turn are the parts of musical instruments, where “the music is happening”.

Investigation of these fine structures is often only possible when a geometrical model of the instrument is built, which includes nearly all details of the instrument (Bader 2006a, 2009, 2007b; Elejabarrieta et al. 2004). These details are responsible for the fine tuning of the sound under different playing conditions. For example the quality of a violin may only be judged after using it under many different conditions, in a chamber orchestra, in large concert halls, with folk music, as a solo instrument etc. Under all these conditions the violin has to fit the needs of presence, loudness, or timbre flexibility. These changes are difficult to detect with methods only concentrating on basic characteristics of the instrument.

The methods used in the field of modelling of whole instrument geometries are basically Finite-Element Methods (FEM) (Barthe 2002; Knothe and Wessels 1999), Finite-Difference Methods (FDM) (Bader 2005a; Bilbao 2009), methods of flow dynamics, like Lattice-Boltzmann method (LBM) or simplified methods, like Waveguides, Delay-Lines or wave front methods. All these methods use basic differential equations for bending, in-plane movement, or flow dynamics with appropriate boundary conditions as are well known from the literature (Leissa 1969a, b; Wagner 1947; Hutchings 1981; Fletcher 2000; Flügge 1962). The advantage of FEM or FDM is that geometries of any complexity can be perfectly resolved and any kind of differential equations found to govern the system can be used in the underlying spatial discretisation. Some disadvantages are the need for many nodal points to resolve the geometry, and the arising computational cost, both in memory and computation time. Here a method that could provide a high spatial resolution with shortened computation time (real-time or close to real-time in the ideal case) would clearly be desirable.

2 Real-Time Synthesis Solutions

For many reasons, a real-time implementation is of great advantage.

1. Musical instrument shapes are so complex, that many changes need to be tested. Some sound improvements can only be found by trial and error. Until now, there is no theoretical framework enabling one to design a desired sound and use an algorithm which calculates the instrument shape which produces this sound. So as the possibilities of instrument changes are nearly endless, a fast, real-time tool is needed to try out changes of instruments while immediately judge the differences aurally.
2. Experts for musical instruments are instrument builders. As they are not familiar with Physical Modelling and computer implementations, they would need a soft- or hardware tool to try different shapes and listen to the results. Although instrument builders would agree to wait for several minutes to get a resulting sound, the calculation time of several hours for one sound is too much to be practical in musical instrument building environments. Here a fast, if not real-time implementation could be used and would be a great convenience in instrument building practice.
3. Modern music production is based on real-time sound producing algorithms. Only in rare cases would a sound designer or musician wait for hours before a calculated sound can be played. In a realistic musical setting they are in need of real-time sound synthesis tools. Physical Model sounds do change sounds within musically reasonable regions and therefore enlarge the creative possibilities to a great extent. A whole geometry tool would be a commercial product on a still growing market. Many attempts have been made to include Physical Modelling in sound production by commercial soft- and hardware musical

instrument building companies, and hundreds of 'Virtual Musical Instruments' are on the market. Most of them are just sampled versions of real instruments which are filtered and manipulated using modern musical signal processing techniques.

Virtual Instruments that try to get as close to a Physical Model as possible again use many simplifications. Though the resulting sounds are often very close to the real instruments they often lack timbre components in the transient or steady state. Hence, a method using accurate whole geometry Physical Modelling would be a huge step towards realistic sounding Virtual Instruments.

2.1 FPGA-Hardware

A Field Programmable Gate Array (FPGA) is a special form of a Very Large Scale Integrated Circuit (VLSI) consisting of matrix-like ordered Logic Blocks, Input/Output-Blocks, and Routing Channels (interconnection network), and is therefore similar to an Application Specific Integrated Circuit (ASIC). Logic Blocks are built out of different kinds of logical units, mainly Look-Up-Tables (LUT) with AND- and OR-Gates and other vendor specific logic such as FLIP-FLOPs or Multiplexers. All Logic Blocks can be utilised in parallel and connected with an interconnection network to yield any function expressible by Boolean logic. Besides this, a FPGA chip consists of physical input and output ports to connect the implemented logic to other devices.

In contrast to an ASIC and other forms of VLSIs, all elements of the FPGA are programmable by the user. This means, FPGAs can be programmed and reprogrammed using a vendor specific flashing tool, with differing logic as often as needed. This results in the possibility to prototype, test and implement hardware algorithms from a standard Personal Computer.

These capabilities of free programmability and of parallel rather than sequential processing are the great advantages of FPGAs compared to CPUs, DSPs, or other sorts of Micro Processors. On an FPGA it is therefore possible to compute massive numbers of instructions in parallel within one clock cycle. To benefit from this advantage it is crucial to implement an algorithm that processes as many parallel instructions as possible. It has been shown that an optimized parallel FPGA algorithm is always superior to a similar sequential algorithm in means of calculation time and maximum clock rates (Brassail et al. 2007; Subasri et al. 2006; Lis and Kowalski 2008; Zou et al. 2006; Inoguchi 2004). In most cases this massive parallel computation can be implemented on a single FPGA chip located on a small mother board, so it is much smaller and much cheaper than a multi-core server system, where MPI can be used to parallelise the algorithm (Karniadakis and Kirby II 2003). These systems are very expensive even if only about 64 knots are used. Furthermore, more knots do not result in a linear speed up anymore. The FPGA on the other hand is capable of processing large amounts of calculations in

parallel on one single, small board! FPGA chips can also be implemented in a commercial solution as e.g. in RME sound cards.

A software version of a model that is intended for implementation on a FPGA device is written in a special Hardware Description Language such as VHDL or Verilog. These Programming Languages contain many High Level constructs like if-, for-, or, while-loops, object oriented programming and some Low Level aspects like bitwise declaration of signals, variables, and constants. The main difference in comparison to other programming languages is the concurrent processing of instructions. This means that unlike e.g. with the C programming language, where code is evaluated and compiled sequentially, here the code is processed in parallel and needs to be coded accordingly. Although the FPGA structure is parallel by nature, sequential statements can also be realised using a Finite State Machine (FSM) as mentioned above and discussed in more detail below.

Before flashing a VHDL design to a FPGA board, the behaviour of the model can be simulated using a simulation tool. Similar to a software debugger, the FPGA code can be simulated using all kinds of different external constraints, like clock speed, changing temperature specifications or varying input values to the model. After validating if the simulated model behaves as expected, the VHDL source code can be compiled and build into a bit-file which then can be flashed onto the hardware device.

The build-process includes different steps like the translation to hardware language, mapping of the IO-Ports and finally placing and routing the design, so that all hardware constraints are met. With these advanced programming techniques the engineer is capable to model and simulate complex Digital Hardware Systems completely in software before putting them into use on a real hardware device. This allows very fast and stable development of hardware designs without the need for a physical Integrated Circuit (IC) development.

2.2 FPGA—Implementations

As discussed above, because of their high speed processing capability and flexibility, FPGA devices are used in many fields of digital signal processing.

In the area of sound analysis and sound source localization, implementations have been realized for real-time noise source identification (Veggeberg and Zheng 2009), high-speed Direction of Arrival (DOA) algorithms (Hao and Ping 2002), or Delay And Sum (DAS) beamforming (Chen et al. 2008), besides countless other applications in this field. All of these works either show real-time realization of the problem for the first time or find a tremendous sped up using an FPGA with its parallel processing capabilities.

FPGA implementations in the field of music have been reported, too. Gibons synthesized a singing voice using Wavetable Synthesis (Gibbons et al. 1998). Physical Modelling was performed by Gibbons for a mass-spring system (Gibbons et al. 2005). Motuk solved the 1D model wave equation (Motuk et al. 2005) and a

membrane (Motuk et al. 2007) on an FPGA board. One recent chapter discusses the use of FPGA to synthesize industrial sounds (Martins et al. 2008).

Typical signal processing applications for FPGAs include various implementations of IIR- or FIR-filter designs (Meyer-Baese 2004). In (Madanayake et al. 2004) 2D/3D Plane-Wave Filters are realised using IIR/FIR-Filters. (Shuang et al. 2008) focuses on converting analog to digital controllers using a filter-design with FPGA. Maslenikow et al. and Brich et al. discuss a method how DSP Filter-Designs (IIR/FIR) can easily be implemented on a FPGA chip (Maslennikow and Sergiyenko 2006; Brich et al. 2006).

In the field of sound production, several publications in which an FPGA is used as a function generator for simple signals like a sine wave or a rectangle signal have been published (Meyer-Baese 2004; Reichardt and Schwarz, 2007; Kilts 2007). Most of these works focus on high frequency signals, like an AM-demodulation chain for a digital radio receiver (Meyer-Baese 2004).

Solving differential equations with FPGAs using finite differences have been reported before, too. Suzuki et al. discuss the use of a FPGA to simulate an electric field solving Maxwell's equations using a FDTD (Finite Differences in the Time Domain) algorithm (Kolodko et al. 2005). This algorithm is widely used for the analysis of electromagnetic problems (Shlager and Schneider 1995). Other approaches to solve differential equations on a FPGA implicitely are e.g. the implementation of the conjugate gradient method (Strzdoka and Göddeke 2006) or the Euler Method (Jayalakshmi and Ramanarayanan 2006). For parallel algorithms solving linear equation systems see also (Karniadakis and Kirby II 2003).

Chapters focusing mostly on the real-time aspect of FPGA implementations are e.g. particle track recognition (Liu et al. 2008), Direction Of Arrival (DOA) algorithms (Hao and Ping 2002), High Speed cross correlation (Von Herzen 1998), Noise Source Identification (Veggeberg and Zheng 2009), or Digital Beamforming (Wang et al. 2005).

3 Performance Control of the Physical Model

3.1 PCIe Interface

Among multiple other In–Out-Ports that are available on the XILINX ML-605 development board, a standard host-device communication port is the PCIe (Peripheral Component Interface express) interface. In this project the PCIe-port is used for high-speed data transfer between the instrument model calculated on the FPGA-Board (Device) and a Graphical User Interface (GUI) running on a Personal Computer (Host). In this section an introduction to the basic functionalities of the PCIe protocol and a short overview on the realised model and the communication protocol is given.

Table 1 Shows some key differences of PCIe compared to PCI-X and PCI

	PCI-Express (PCIe)	PCI-X	PCI
Bus-structure	Serial	Parallel	Parallel
Maximal data-width	64Bit	32/64 Bit	32/64 Bit
Communication protocol	Point-to-point protocol	Split transaction protocol	Delayed transaction protocol
Maximum bandwidth [megabytes/second]	16000 (Rev.: 3.0)	4260 (Rev.: 2.0)	532 (Rev.: 2.1)
Clock signal	Retrieved from data-stream	External clock lane	External clock lane

3.2 PCIe Fundamentals

In 2002 the PCI interest group, the PCI-Sig consortium[1] published the first specifications of the PCIe protocol as an extension to the established PCI and PCI-X protocols (Solari 2003). To this day the basic protocol has undergone several revisions and currently has the version number 3.0.[2] See Table 1 for an overview.

One of the most notable differences of the PCIe interface compared to the older PCI and PCI-X protocols is the serial structure of the data transfer instead of the prior parallel structure. This fundamental design change arose from the insight that a Point-to-Point interconnection between devices on a bus permitted higher transmission rates but where too costly when implemented in a parallel structure (Solari 2003). The implementation of a serial Point-to-Point communication protocol had another advantage over the older structure: time consuming bus arbitration, a standard in the older protocol specifications, was eliminated, enabling the bus to handle higher clock rates and thus a higher data bandwidth. Another major difference between the protocols is the transmission of the clock signal. In the older protocols (PCI and PCI-X) the clock signal was transmitted over a discrete clock line, this slowed the transmission-rates due to settling time delays in the reference clock signal. In PCIe the clock signal is retrieved directly from the data stream. This is accomplished by using 8b/10b coding for physical transmission of a signal. This coding can represent 8 bit of data in a 10 bit data word eliminating any voltage offset and assuring a maximal number of zero crossings of the signal, making it easier to retrieve a clock signal by a Phase Locked Loop (PLL) circuit. Among many other changes, these three design differences ensure an increase in bandwidth of data communication in computer systems. So today the PCIe interface has become the standard interface for high data throughput communication of peripheral devices with the central processing unit in Personal Computers.

A PCIe platform usually consists of specific features shown in Fig. 1.

[1] A consortium consisting of almost 1000 hardware and software development companies.

[2] Some specifications for Revision 4.0 where published in August 2012.

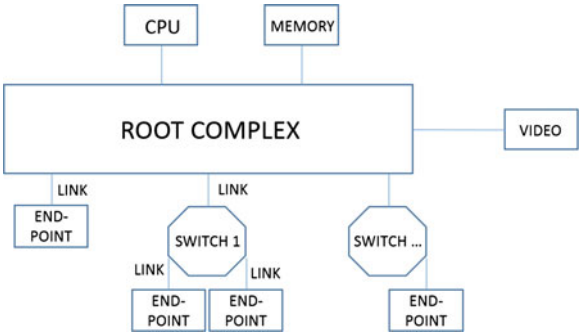


Fig. 1 Overview of basic PCIe platform components

The Root Complex is the interface between all connected PCIe Devices and the Central Processing Unit of the Host as well as the memory of the host and an additional graphics adaptor. Connected to the Root Complex are the PCIe-Devices or Endpoints via Links. A Link consists of at least one PCIe-Lane up to 32 Lanes. Each Lane consists of one differential line for each direction. So a 16 × PCIe Link (a standard Link size for graphics cards) consists of 16 Lanes with 4 lines for every Lane. To connect more Endpoints to a Root Complex, a Switch is used to handle the data transfer and switch between the attached devices.

3.3 PCIe Layer Communication

Following the Open Systems Interconnection (OSI) model[3] standard, the PCIe communication protocol implements the three bottommost layers:

1. The Transaction Layer (TL)
2. The Data Link Layer (DLL)
3. The Physical Layer (PL).

Figure 2 shows a schematic overview of the data communication path of the presented model. The data payload from the FPGA model is packed into Transaction Layer Packets (TLP). The TLP's are then transported to the Data Link Layer and further on to the Physical Layer. From the Physical Layer the data is transmitted to the PC. Each Layer has several specific functions on the data path, the most important are listed in Table 2:

There are four types of PCIe transactions:

1. Memory write/read transaction
2. In/Out transaction
3. Configuration transaction
4. Message transaction

[3] ISO/IEC 7498-1.

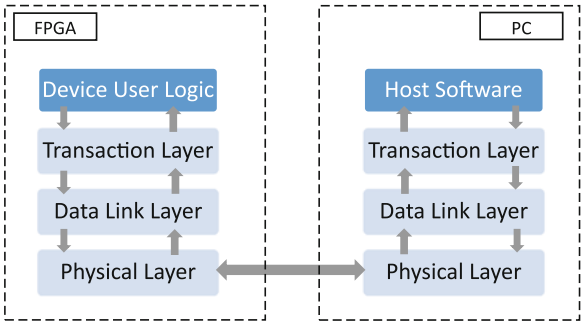


Fig. 2 Communication between the FPGA and the PC via the physical layer of the PCIe

Table 2 Layers of the PCIe transaction

Transaction layer	Handles the interaction with the user logic and builds transaction packets with right header information, like type or direction of interaction
Data link layer	Handles error detection and ensures correct Link functionality
Physical layer	Converts the data into a serial bit stream with integrated reference clock

The IO-transaction is provided for backwards compatibility with PCI so the most common method for transferring data between a host and a device is using memory transaction. Configuration transactions are used for writing and reading the configuration space of PCIe devices. Message transactions include error messages, interrupt requests or power management event signals.

3.4 Hardware Design

The final hardware design is based on a XILINX Core Designer Project for the ML605 FPGA Development-Board which implements four on-board Block-RAM's with 2048 Kbyte each as addressable Memory- or IO-Space. All data transactions in the model are implemented as memory read/write transactions. The control data from the GUI is sent to the FPGA Board as data with a destination address specifying the type of the control data. When for instance two strings are implemented on the FPGA board, each string has a different address for the value of its initial length or its material property. The output data of the model for auralisation is written to a memory space RAM Block on the FPGA. After one time-step of the implemented model is calculated, the software on the PC host reads the content of the memory space and handles all further processing (e.g. communication with the sound card).

3.5 Software, Drivers, and Controllers

The control of the FPGA Physical Model via the PCIe interface is realized by several drivers and software. The drivers ensures communication between the software and the PCIe Interface, while the software communicates with the drivers to downstream parameters of the physical model to the FPGA and to upstream the sound produced by the virtual instrument, as well as other calculated values, from the hardware. Additionally, the software communicates with a soundcard implemented on the root complex, to play the sound produced by the FPGA in real-time to the performer. This soundcard can be attached to the root complex via another PCI or PCIe slot, or via a USB port. Of course the calculated parameters and sound can also be stored on the computer and used for analysis later on. The system described here is working on a WINDOWS platform, although a similar implementation can also be built on a Linux, MAC, UNIX, or another system.

3.6 Drivers

The Application cannot communicate directly with the Device, it needs the Windows Driver Foundation (WDF) class. This is split into User and Kernel Mode, where the Application has only access to the User Mode. From a software perspective, e.g. within the programming language C++, the communication with the kernel mode is through subroutines of the Windows API, historically first implemented in the file system. So all data transfer is realized using the CreateFile, WriteFile, etc. commands. So all devices, from a software standpoint, are similar to a file system, sending or receiving data or control parameters.

The Kernel Mode, communicating between the Application and the Device, has several subsystems, the I/O system or a Plug-and-Play Power system for inserting USB devices on the fly and managing power consumption of PCIe devices, etc. These subsystems communicate with the Framework and the WDF objects, which are objects defined by the drivers of the devices. The Framework contains default objects for devices, which can be overwritten by drivers implementing their own callbacks to the WDF Framework. So the Framework manages all requests from the upper edge with the Applications and the lower edge of the Devices. Most of this communication is default, so the Framework automatically takes care of them. Still all drivers installed may add additional functionality, which the Framework considers.

So all specifications of the device need to be implemented in the Drivers which can be installed in the User Mode or in the Kernel Mode. User and Kernel Mode driver can have additional drivers adding some functionality which are therefore called Function Drivers. So three kinds of drivers exist to communicate between the Application and the Device: User Mode, Kernel Mode, and Function Drivers.

The separation between a Kernel and a User Mode is necessary for security reasons as well as for programming convenience. So e.g. the OS runs on many different hardware motherboards with different memory, bus, and CPU configurations. Therefore, writing code would mean to know all hardware specifications of all possible CPUs, busses, etc., which is not practical. The Hardware Abstraction Layer (HAL) of the OS therefore knows all hardware on the market and translates the commands to the specific code needed by the hardware. Also, to ensure stability of the system, it is necessary for the OS to not be disturbed by loops in a code to end in a deadlock and crash. Therefore, the user is only allowed to suggest code which is then checked by the OS in terms of its operability. If the code would lead to a crash it is not performed by the OS. All code written in the User Mode therefore need not to know anything about the hardware of the computer, and also need not consider a general crash of the system. Still this also means that certain operations, like reading or writing memory directly to devices, power up or down devices, etc. are not allowed in the User Mode and only can be performed in the Kernel Mode. Therefore, drivers which do not need such operations, like USB drivers, can run in User Mode. Still direct access to memory, as needed with the FPGA implementation here, needs at least one Kernel Driver. But writing and testing a Kernel Driver is much more demanding and have serious security risks, therefore User Mode Drivers are preferred whenever possible.

The Drivers are organized in a stack, a User-mode and a Kernel-mode driver stack. Communicating between an Application and a Driver works through the Windows API functions for file operations, as discussed above. The Windows API calls the Kernel Subsystems, e.g. for an I/O-request it calls the I/O subsystem. This subsystem then builds an interrupt (IRP) for the request and sends it to the first Kernel Mode driver on the Kernel-mode stack. This driver, as all drivers on the stack, is a specific driver written for one device. Therefore it determines if it is responsible for the request. If it is, it performs the operations with the driver via the HAL and sends the results back to the Kernel subsystem. If not it checks if the driver may be a User Mode driver. If so, it sends the request to the User-mode stack. If the driver is not responsible for the request and the device does not use a User Mode driver, the first Kernel driver sends the request to the next Kernel Mode driver in the stack. This continues until one driver answers the request. Then the flow changes direction and the request is send to the next higher driver, which is still allowed to add functionality, if a routine in the driver is set to such a case. This may happen when several drivers operate on one device with different tasks. If the IRP has reached the highest driver on the Kernel-mode stack, this driver than passes the IRP to the kernel subsystem. This subsystem creates an error-code (a no-error at best) and may perform necessary memory copies. This is the case with the Kernel Driver used in this application. There, device memory is read by the kernel subsystem and written into the memory space accessible by the Application. It then returns control to the Application via the Windows API again, returning the error code and maybe also a pointer to the memory space the memory copy is accessible by the Application.

3.7 Memory

The OS has two kinds of memory space, physical and virtual. The physical space addresses represents all available physical memory on the system, while the virtual memory maps this physical memory to different layers in the OS. This mapping, or paging, is necessary, as e.g. two Applications with two different memories may use the same physical memory of a driver. Also in User Mode, the physical memory is not available because of security reasons. Therefore, to access the physical memory of the FPGA, its memory must be mapped in such a way to be accessible by the Application. Three memory types are needed to achieve this, the physical memory on the FPGA, a virtual memory space in der Kernel Mode, and the virtual memory space in the User Mode. The mapping is achieved using a Memory Descriptor List (MDL).

3.8 Implementation

To communicate between the software and the FPGA via the PCIe, reading and writing to its physical memory needs three drivers, a soundcard handling, and a control software.

3.8.1 Kernel Mode Driver

To map the physical memory to virtual memory, Kernel Mode drivers can use the MmMapIOSpace function. This function needs the physical memory address and the size of the memory. It maps the physical memory to a virtual memory space in the Kernel Mode and returns a pointer to this virtual memory space. With this pointer, a MDL is allocated and built with the MmBuildMdlForNonPagedPool and MmMapLockedPages functions. These Kernel Mode functions then map the virtual Kernel Mode memory to a virtual User Mode memory space. This User Mode memory space can then be accessed in User Mode.

These functions are implemented in the Kernel Mode Driver in the DeviceIOControl callback, in which a control ID is defined, which is provided by the User Mode Driver calling the Kernel Mode driver with a device control request.

3.8.2 User Mode Driver

The User Mode driver asks the Kernel I/O subsystem for a handle to the Kernel Mode driver via a CreateDevice function of the Windows API. This function calls the Kernel Mode driver installed on the system and asks for the handle. It then accesses the memory mapping function of the Kernel Mode driver by calling the DeviceIOControl Windows API function using the control ID defined in the

Kernel Driver. The Kernel I/O subsystem calls the Kernel driver, which returns the pointer of the virtual User Mode memory space to it in the IRP. The Kernel subsystem copies the physical memory of the FPGA device into the virtual Kernel memory and provides the Application with the mapped memory address of the virtual User Mode memory space.

3.8.3 Filter Driver

The Filter Driver is not necessary to read and write memory from and to the FPGA. Still it is necessary to access the PCI configuration space, where parameters of the model can be set or obtained. Therefore the PCI bus need to be accessed which cannot be addressed, as it is unnamed. Still when installing a function driver adding additional functionality to the bus, this driver is installed by the OS in such a way to directly have access to the bus driver. This access also includes the name of the bus, which then can again be used by the Kernel Driver to access PCI configuration space.

3.8.4 Soundcard Handling

The time series computed by the Physical Model on the FPGA is to be played in real-time by a soundcard. As users may have any kind of soundcard available on the market, standard drivers to play back sound are reasonable to implement with this application. Although in previous investigations, e.g. when building a microphone array digital input, the ASIO standard was used to be able to record 128 channels with 48 kHz each in real-time, here a more simple solution was implemented using a DirectX sound I/O setup.

Therefore, the DirectSound library in version 1.0.2902.0 was used. To play back sound, a circular buffer was allocated and attached to a device object of this library. The sound object was set to 96 kHz with 24 Bit sampling rate. A short buffer length was used to realize real-time performance without noticeable delay between players' performance control and sound changes. A notification thread was attached to the buffer to notify the application when the soundcard asked for new data, to refresh the buffer with new FPGA data. So the soundcard plays back the buffer which is filled with new data right after filling. When starting the device to start playback, the notification thread is started, too.

3.9 Control Software

The control software includes all the threads necessary to perform, the control input from a performer, the parameter streaming to the FPGA, the up-streaming of the sound produced by the FPGA, and the real-time playback of this sound to a soundcard to playback the performance. Therefore, three threads are implemented in the software.

3.10 Performance Thread

This thread continuously waits for performance data from any kind of controller, which is digitalised and supported to the software via and I/O driver or via mouse or keyboard control. The control parameter is then mapped to a parameter in the FPGA Physical Model. This could be speed or pressure of a violin bowing, it could be the position of a finger on the fretboard, it could be a knocking on the instrument, a detuning of a string or a membrane tension, or any other kind of parameter which is a parameter in the Physical Model. This parameter is mapped to the memory or configuration space of the PCIe hardware, implemented on the FPGA and written to the hardware. There, the FPGA hardware accesses it at the next time point of calculation of the Physical Model, changing this parameter in real-time during performance.

3.11 Sound Upstream Thread

The FPGA Physical Model solves a complex differential equation system for adjacent time points. The solution of one or several nodal points of the Physical Model, chosen by the user, is stored in the PCIe memory. The model continues to calculate the differential equation system for a finite amount of time points. This amount is a compromise between performance speed, PCIe package handling, and soundcard buffer size. From the standpoint of performance, a very short soundcard buffer size reduces the delay time between performance data input and sound output. Still, very short buffer sizes, like 8 or 16 words, are not efficiently sent through the PCIe packaging system, both for the upstream from the FPGA to the application as well as from the application to the soundcard, which is also a PCIe device (or USB, which is a PCIe on a lower level, too). So in this application the VHDL code calculates 32 time points, stores them into the PCIe memory space, and sets a flag on the PCI configuration space that it finished another buffer of data. The application is checking this flag, accessing the PCI configuration space within the sound upstream thread of the application. If the flag is set, the software knows that a new buffer of data is ready, reads it from the PCIe memory space, and stores it in an internal buffer. Then the software sets back the flag and the FPGA starts calculating another buffer of time points.

3.11.1 Sound Play Buffer Thread

The play buffer thread depends on the soundcard. It is a notification thread, as discussed above, which notifies, when the soundcard has read data from the buffer up to the next point of notification within the buffer. So the buffer need to be refilled with data, which is available from the internal buffer filled by the FPGA data stream. After copying this data, the thread sets the flag in the configuration

space of the PCIe so that the FPGA knows that it needs to continue calculating new data.

Most of the time the FPGA is indeed waiting, and not calculating, as the speed of calculation is much faster than the playback of sound. So the internal sampling rate of the Physical Model can be much higher than the playback rate of the soundcard, maybe 10–15 times higher, which increases calculation precision and stability.

4 Conclusions

The system for real-time performance of playing FPGA Physical Modelling of virtual musical instruments is flexible in a way to input any kind of control parameter. So e.g. the software could also be compiled as a dynamic library in the VST PlugIn approach and included in a sequencer software. Then the standard controllers known from recording studios, like keyboards and other MIDI inputs, like breath controller, guitar-to-midi controller, etc. could be used. Still it can also be used to input data from AD-converters with a USB interface attached to the computer. Its implementation of Physical Models is flexible, too, as different virtual musical instrument Physical Models can be uploaded via the PCIe interface. The compiled VHDL files are mostly so small that an upload of a new instrument, or a partial reconfiguration on the FPGA, is so fast that a performer will not need to wait, but experiences an instantaneous change of the instrument. Still this is not discussed within this chapter and subject to future work.

Although a systematic investigation about the experiences of users of this system is still to be done, the first experience of performers is the wide range of possible sounds of this system. As musical instruments have many nonlinearities, where sudden phase changes of the sound appear when linearly changing a control parameter, playing the virtual violin or banjo is often a great pleasure, as the performer can produce very different and interesting sounds easily. Indeed, as the playing parameters are manifold, a selection need to be done, which can also be changed in real-time. Then, playing such instruments, although they might be traditional ones like violins or plucked string instruments, are often like learning the same instrument in a new way, often leading to very new and fresh musical experiences.

Literature

Bader, R. (2002). Fraktale Dimensionen, Infomationsstrukturen und Mikrorhythmik der Einschwingvorgänge von Musikinstrumenten. Doktorarbeit der Universität Hamburg. http://www.sub.uni-hamburg.de/volltexte/2002/598.

Bader, R. (2003). Physical model of a complete classical guitar body. In R. Bresin (Ed.). Proceedings of the Stockholm Music Acoustics Conference, (Vol. 1, pp. 21–124).

Bader, R. (2004). Additional modes in transients of a balinese gender dasa plate. *Journal of the Acoustical Society of America, 116*, 2621.
Bader, R. (2005a) Computational Mechanics of the Classical Guitar. Springer (http://www.springeronline.com/3-540-25136-7).
Bader, R. (2005b). Complete geometric computer simulation of a classical guitar. Lay-Language paper of the American Acoustical Society 05, http://www.aip.org/149th/bader_Guitar.htm.
Bader, R. (2005c). Whole geometry finite-difference modeling of the violin. In Proceedings of the Forum Acusticum 2005, (pp. 629–634).
Bader, R. (2005d). Turbulent k-e model of flute-like musical instrument sound production. In E. Lutton & J. Lévy-Véhel (Ed.), Fractals in Engineering. New trends in theory and applications. Springer, (pp. 109–122).
Bader, R. (2006a). Finite-element calculation of a bass drum. Journal of the Acoustical Society of America, 119, 3290.
Bader, R. (2006b). Characterization of guitars through fractal correclation dimensions of initial transients. *Journal of New Music Research, 35*(4), 323–332.
Bader, R. (2007b). Spatial cognitive timbre dimensions of physical modelling sounds using Multi-Dimensional Scaling Techniques (MDS). *Journal of the Acoustical Society of America, 121*, 3184.
Bader, R. (2008a). Individual reed characteristics due to changed damping using coupled flow-structure and time-dependent geometry changing finite-element calculation. In Forum Acusticum joined with American Acoustical Society Paris 08. (pp. 3405–3410).
Bader, R. (2008b). Fine tuning of guitar sounds with changed top plate, back plate and rim geometry using a whole body 3D finite-difference model. In Forum Acusticum joined with American Acoustical Society Paris 08. (pp. 5039–5044).
Bader, R. (2009). Additional modes in a Balinese gender dasa plate due to its trapezoid shape. In R. Bader & Ch. Neuhaus & U. Morgenstern (Eds.), Studies in systematic musicology. Peter Lang Verlag (in print).
Bader, R. & Hansen, U. (2008). Acoustical analysis and modeling of musical instruments using modern signal processing methods. In D. Havelock & M. Vorländer & S. Kuwano (Eds.), Handbook of Signal Processing in Acoustics (pp. 219–247). Berlin: Springer.
Barthe, K. J. (2002). Finite-element methoden. Springer.
Bécache, E., Chaigne, A., Derveaux, G., & Joly, P. (2005). Numerical simulation of a guitar. *Computers and Structures, 83*(2–3), 107–126.
Bilbao, S. (2009). Numerical sound synthesis. Finite-difference schemes and simulation in musical acoustics. Wiley.
Bissinger, J. (2003). Modal analysis of a Violin Octet. *Journal of the Acoustical Society of America, 113*(4), 2105–2113.
Brassail et al. (2007). FPGA Parallel Implementation of CMAC Type Neural Network with on Chip Learning. *IEEE* 111–115.
Brich, T., Novacek, K. & Khateb, A. (2006). The digital signal processing using FPGA. ISSE'06, 29th International Spring Seminar on Electronics Technology. (pp. 322–324).
Chen, P., Tian, X., Chen, Y., & Yang, X. (2008). Delay-sum Beamforming on FPGA. ICSP 2008 Proceedings.(pp. 2542–2545).
Da Silva, A., Scavone, G., & van Waltijen, M. (2007). Numerical simulations of fluid-structure interactions in single-reed mouthpieces. *Journal of the Acoustical Society of America, 122*(3), 1798–1809.
Elejabarrieta, M. J., Ezcurra, A., & Santamaria, C. (2002). Coupled modes of the resonance box of the guitar. *Journal of the Acoustical Society of America, 111*, 2283–2292.
Elejabarrieta, M. J., Ezcurra, A., & Santamaria, C. (2004). Vibrational behaviour of the guitar soundboard analysed by means of finite element analysis. *Acustica united with Acta Acustica, 87*, 128–136.
Erkut, C., Karjalainen, M., Huang, P., & Välimäki, V. (2002). Acoustical analysis and model-based sound synthesis of the kantele. *Journal of the Acoustical Society of America, 112*(4), 1681–1691.

Essl, G., & Cook, P. (2000). Measurements and efficient simulation of bowed bars. *Journal of the Acoustical Society of America, 108*(1), 379–388.

Fletcher, N. & Rossing, Th. (2000). Physics of Musical Instruments. Springer.

Flügge, W. (1962). Statik und Dynamik der Schalen. Springer-Verlag.

FPGA Parallel Implementation of CMAC Type Neural Network with on Chip Learning. IEEE 2007. (pp. 111–115).

Gibbons, I. S., Howard, D. M. & Tyrrell, A.M. (1998). Real-time singing synthesis using a prarllel processing system. IEE Colloquium on Audio and Music Technology. The Challenge of Creative DSP, 8/1–8/6.

Gibbons, J. A., Howard, D. M., & Tyrrell, A. M. (2005). FPGA implementation of 1D wave equation for real-time audio synthesis. *IEE Proceedings, Computers and Digital Techniques, 152*(5), 619–631.

Giordano, N. (2006). Finite-difference modeling of the piano. *Journal of the Acoustical Society of America, 119*, 3291.

Hao, C., & Ping, W. (2002). The high speed implementation of direction-of-arrival estimation algorithm. *International Conference on Communication, Circuits and Systems and West Sino Expositions, 2*, 922–925.

Hutchings, C. M. (1981, October). Klang und Akustik der Geige. In Spektrum der Wissenschaft, Dezember 1981, (pp. 112–122) (original: Scientific American).

Inoguchi, Y. (2004). Outline of the ultra fine grained parallel processing by FPGA Seventh International Conference on High Performance Computing and Grid in Asia Pacific Region, 2004 pp. 434–441

Jansson, E. V. (2003). The BH-Hill and tonal quality of the violin. In Proceedings of Stockholm Music Acoustical Conference (SMAC 03), (Vol. 1, pp. 71–4).

Jayalakshmi, K. & Ramanarayanan, V. (2006). Real-time simulation of electrical machines on FPGA platform, IICPE pp. 259-263. doi:10.1109/IICPE.2006.4685378

Karniadakis, G. E., & Kirby II, R. M. (2003). Parallel Scientific Computing in C++ and MPI. Cambridge University Press.

Kilts, S. (2007). *Advanced FPGA Design: Architecture, Implementation, and Optimization*, isbn: 0470054379, Wiley-IEEE Press

Knothe, K., Wessels, H. (1999). Finite Elemente. Eine Einführung für Ingenieure. 3. Auflage, Springer.

Kolodko, J., Suzuki, S. & Harashima, F. (2005). Eye-gaze tracking: an approach to pupil tracking targeted to FPGAs, IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 344–349

Lau, A., Spiess, P., Wriggers, P., Schneider, A., & Bader, R. (2004, Feb). Analysis of bell vibrations. In 2nd Conference on Advances and Application of GiD. Preprints. International Center for numerical methods in engineering, Barcelona.

Leissa, A. W. (1969a). Vibration of Plates. Publication of the American Acoustical Society (pp. 1993).

Leissa, A. W. (1969b). Vibration of Shells. Publication of the American Acoustical Society (pp. 1993).

Lis, J. D. & Kowalski, C. T. (2008). Parallel fixed point FPGA Implementation of sensorless induction motor torque control. 13th Power Electronics and Motion Control Conference 2008, EPE-PEMC, (pp. 1359–1364).

Liu, M., Kuehn, W., Lu, Z. & Jantsch, A. (2008). System-on-an-FPGA design for real-time particle track recognition in physics experiments. 11th Euromicro Conference on Digital System Design Architectures, Methods and Tools. (pp. 599–605).

Madanayake, A., Bruton, L., Comis, F. & Comis, C. (2004). FPGA Architectures for Real-Time 2D/3D FIR/IIR Plane Wave Filters. Proceedings of the 2004 International Symposium on Circuits and Systems ISCAS'04, (Vol. 33 613–616).

Martins, G., Barata, M. & Gomes, L. Low cost method to reproduce sound with FPGA. IEEE International Symposium on Industrial Electronics 2008. ISIE 2008, (pp. 1932–1936, 208).

Maslennikow, O. & Sergiyenko, A. (2006). Mapping DSP Algorithms into FPGA. Proceedings of the International Symposium on Parallel Computing in Electrical Engineering, (pp. 208–213).
Meyer-Baese, Uwe. (2004). *Digital signal processing with field programmable gate arrays* (3rd ed.). Berlin: Springer.
Motuk, E., Woods, R. & Bilbao, S. (2005). Implementation of finite-differece schemes for the wave equation on FPGA. IEEE International Acoustics Speech and Signal Processing ICASSP'05, (Vol. 3, 237–240).
Motuk, E., Woods, R., Bilbao, S., & McAllistere, J. (2007). Design methodology for real-time FPGA-based sound synthesis. *IEEE Transactions on Signal Processing, 55*(12), 5833–5845.
Musical Acoustics: Virtual Instruments I & II', Session at the Joined Meeting of the American Acoustical Society with the European Acoustical Society, Paris 2008, Journal of the Acoustical Society of America. 123 (5, 2) 3521–23, 3664–66.
Özakča, M. T. G. (2004). Structural analysis and optimization of bells using finite elements. *Journal of New Music Research, 33*, 61–69.
Pfeifle, F. & Bader, R. (2009). Real-time physical modelling of a real Banjo geometry using FPGA hardware technology. In Bader, R. (Ed.). Musical Acoustics, Neurocognition and Psychology of Music Hamburg Yearbook for Musicology 25, (pp. 71–86).
Pfeifle, F., & Bader, R. (2011). Real-time finite-difference string-bow interaction Field-Programmable Gate Array (FPGA) model coupled to a violin body. *Journal of the Acoustical Society of America, 130*, 2507.
Pfeifle, F., & Bader, R. (2012). Real-time finite-difference physical model of musical instruments using a Field-Programmable Gate Array (FPGA). Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12), York, (p. 1–8) 2012.
Reichardt, J. & Schwarz, B. (2007). *VHDL-Synthese: Entwurf digitaler Schaltungen und Systeme*, isbn: 9783486581928, Oldenbourg. http://books.google.de/books?id=v93aGAAACAAJ
Schneider, A., & Bader, R. (2003). Akustische Grundlagen Musikalischer Klänge. Mitteilungen der mathematischen Gesellschaft Hamburg.
Schoofs, A. (2002). Computer-aided bell design and optimization in The Quality of bells. Eurocarillon 2002, Brugge, Sept. 6th, 2002. Proceedings of the 16th Meeting of the FWO Research Society on Foundations of Music Research, Ghent University:Ghent: IPEM.
Shlager, K. L., & Schneider, B. (1995). A Selective Survey of the Finite-Diffrence Time-Domain Literature. IEEE Antennas and Propagation Magazine, (Vol. 37 pp. 39–56).
Shuang, K, Yankai, X., Shan, J. & Hongwu, Z. (2008). Converting analog controllers to digital controllers with FPGA. ICSP2008 Proceedings. (pp. 486–489).
Smith, (2008). Musical Acoustics: Virtual Instruments I & II', Session at the Joined Meeting of the American Acoustical Society with the European Acoustical Society, Paris 2008, *J. Acoust. Soc. Am.*, (Vol. 123 (5, 2) pp. 3521–3523, 3664–3666).
Smith, J. III: Digital waveguide modeling and signal processing. http://ccrma.stanford.edu/∼jos/wg.html.
Solari, E., & Congdon, B. (2003, September). The Complete PCI Express Reference. Intel Press.
Strzdoka, R., & Göddeke, D., (2006). Pipelined mixed precision algorithms on fpgas for fast and accurate pde solvers from low precision components. 14th annual IEEE Symposium on Field Programmable Custom Computing Machinges, (pp. 259–270).
Subasri, V., Lavanya, K., & Umamaheswari, B. (2006). Implementation of digital PID controller in Field Programmable Gate Array(FPGA). *Proceedings of India International Conference on Power Electronics, 172–176*, 2006.
Veggeberg, K., & Zheng, A. (2009). Real-time noise source identification using programmable gate array(FPGA) technology. *Proceedings of Meetings on Acoustical Society of America, 1–10*, 2009.
Von Herzen, B. (1998). Signal processing at 250 Mhz using high-performance FPGA's. IEEE Transactions onvery large scale integration (VLSI) Systems. (Vol. 6 (2) pp. 238–246).
Wagner, K.W. (1947). Einführung in die Lehre von den Schwingungen und Wellen. Wiesbaden.
Wang, Z., Jin, R., Geng, J. & Fan, Y. (2005). FPGA implementation of downlink DBF calibration. Antennas and Propagation Society International Symposium.

Woodhouse, J. (2005). On the Bridge Hill of the Violin. *Acta Acustica united with Acustica, 91*, 155–165.

Zou, Z., Hongyuan, W. & Guowen, Y. (2006). An improved MUSIC algorithm implemented with high -speed parallel optimization for FPGA. 7th International Symposium on Antennas, Progpagation & EM Theory, ISAPE '06, (pp. 1–4).

Multisymplectic Pseudo-Spectral Finite Difference Methods for Physical Models of Musical Instruments

Florian Pfeifle

1 Introduction

In musical acoustics most aural perceivable features and properties of radiated sound resulting from structural vibrations can be described to a satisfactory grade with Differential Equations (DE), in most cases the wave equation. The order and dimension of this Partial Differential Equation (PDE) is given by the respective geometry of the instrument and the vibrational behaviour of the underlying structure. Beside the linear behaviour, also non-linear properties can be described by these equations, giving accurate results in most cases. Over the last years there has been a tremendous amount of work regarding physical models of musical instruments and PM for musical applications in general (Bilbao 2009; Pfeifle and Bader 2012b; Bader 2005a; Giordano 2006). Most of these works show that physically based models of musical instruments have a very realistic sound and timbre quality compared to instruments synthesized with other methods. Still, one of the major drawbacks of all PM methods is their computational cost which is directly linked to the accuracy of the model and the sound quality. So in most cases these instrument models are not capable of real-time sonification or are simplified in many ways. In a recent work, real-time implementations of physical models of several musical instruments where presented (Pfeifle and Bader 2011a, 2012b). A real-time controller was implemented, making it possible to play and configure the instrument models. Among many other findings two central insights where gained over the course of that work:

1. It is absolutely necessary to have a stable algorithm that yields steady results under changing conditions and parameters.
2. A high accuracy and small error of the used method benefits the overall stability and sound quality of the models.

F. Pfeifle (✉)
Institute of Musicology, University of Hamburg, Neue Rabenstr. 13, 20354 Hamburg, Germany
e-mail: fkra003@uni-hamburg.de

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_15,

In this work we research properties of two methods that could facilitate these factors and result in more stable and accurate real-time models of musical instruments.

All PM-based sound synthesis techniques take a similar approach, the governing PDEs are discretised and solved using numerical methods like e.g. the Finite Element Method (FEM) or Finite Difference Method (FDM) (Bilbao 2009). Both methods are used in many fields of numerical mathematics and physics (Bathe 2002), Peiró and Sherwin (2008) and are a de facto standard in many commercial PM software tools.[1] A progression of these methods are Spectral and Pseudo-spectral (PS) methods. They where developed in the 1970s and 1980s basing in FDMs and FEMs. The development of the mentioned methods can roughly be categorised into three eras:

- 1950s: Finite Difference Methods
- 1960s: Finite Element Methods
- 1970s: Spectral and Pseudo-spectral Methods (Fornberg 1998).

Even though spectral methods have been shown to have superior properties for a large range of discretised DE solutions, like for instance spectral accuracy resulting in a smaller discretisation error, they initially where only suited for problems of regular geometry and periodic boundary conditions. To reduce some of these imposed restrictions many different sub-methods have been proposed among which PS methods take a prominent role (Lyons et al. 2005; Patera 1984; Komatitsch and Tromp 2002; Gottlieb and Orszag 1987; Hamman et al. 2007; Wang 2002; Trefethen 2000; Chaljub et al. 2007; Lee et al. 2000; Komatitsch et al. 2005).

PS methods have been applied with success in many fields of research like fluid dynamics (Patera 1984), medical research (Tong and Krozer 2002) or for solving the Korteweg-de-Vries equation (Ascher and McLachlan 2004; Fornberg 1998). In room acoustics there are several works that use PS methods for solving large 3-dimensional FD problems with success (Spa et al. 2010). To the best of our knowledge there is only one work where a spectral method is used to solve a physical model of a musical instruments numerically, but only for calculating mode shapes, not for sound synthesis purposes (Sathej and Adhikari 2008). This work tries to fill this gap and present a methodology for implementing such methods for musical instruments.

There is a large body of work for numerical problems of many kinds and in depth research of pseudo-spectral methods and its properties since the late 1970s (Fornberg 1990; Tong and Krozer 2002), so in this work we can build upon the results of such important works for pseudo-spectral methods like the monographs covering this topic by Fornberg (1998) and Trefethen (2000). In their work they show many applications and implementations of PS methods for solving differential equations in various dimensions and for linear and nonlinear problems (Lesage et al. 2008).

[1] COMSOL, ANSYS and many others.

One conjunctive property to all discrete solution techniques of DEs is the discretisation of all independent variables over the given problem domain.[2] There are several approaches of solving discrete equations but in the case of PM most methods parrallelise the problem in space and calculate the time progression of the solution with implicit or explicit time-stepping methods (also known as numerical integrators) (Bilbao 2007).

Over the last decade a new classification paradigm for qualitative behaviour of time stepping methods (numerical integrators) has been developed by comparing basic properties of standard solution algorithms. Besides fundamental properties like energy conservation or reversibility an important quality of a numerical integrator is the area preservation of the Hamiltonian flow. If an algorithm satisfies this requirement it is called *symplectic* (Feng and Qin 1987; Hairer et al. 2002).

Hairer et al. showed the beneficial properties of symplectic integrators for ODEs on long time stability, energy conserving properties and other advantages over non-symplectic methods. Even though a symplectic method does not always preserve the energy completely (McLachlan et al. 2006) it is in many cases far superior to non-symplectic methods.[3] Over the last years several works extended symplectic ODE methods to PDEs yielding *multisymplectic* integrators (Feng and Qin 1987; Moore 2009; Moore and Reich 2003b). Even though there is ongoing research of several attributes of multisymplectic methods, the results that have been found to this point are encouraging and can be applied in many fields of numerical mathematics[4] (Moore 2009; Schober and Wlodarczyk 2008; Sha et al. 2008; Kong et al. 2007).

2 Mathematical Methods

2.1 Pseudo Spectral Methods

Spectral methods are linked to collocation methods like Galerkin and Chebycheff methods. The basic idea behind all these methods is to calculate discrete differentiation on a global grid defined by the underlying geometry and the specific interpolation functions. There are two conventional methods of calculating the derivative of a function with pseudo-spectral methods:

1. Calculate the derivation function $\hat{\phi}$ in Fourier Space given by the Fourier Transform equalities. Transform $\hat{\phi}$ back to the time domain and perform the

[2] For the wave equation all space variables and the time variable is discretised.

[3] Moore and Reich (2003a) show that in many cases the reversibility of an algorithm is far more important than an exact energy conservation.

[4] To this point many results for multisymplectic integrators are only shown numerically. Some properties have yet to be generalised (Moore 2009; Moore and Reich 2003b).

derivation of function ψ in the time domain via a matrix multiplication with ϕ (Trefethen 2000).
2. Transform function ψ to the frequency space, perform a derivation in Frequency space (dot product with $\hat{\phi}$), then transform the resulting function back to the time domain (Fornberg 1998).

Another point of view can be deduced by properties of the convolution theorem. We start with the example of a finite difference approximation for the 1-dimensional wave equation. One can write the discrete version of the wave 1-d wave equation as follows (Bilbao 2009)

$$\delta_t^2 \cdot \mathbf{u} = c^2 \cdot \delta_x^2 \cdot \mathbf{u} \tag{1}$$

with δ^2 the second order centered finite difference operator defined as

$$\delta_x^2 = \frac{[1, -2, 1]_x}{\Delta x^2} \tag{2}$$

and $\mathbf{u}$, the dependent variable in vector form. If we set the values $\Delta x = \Delta t$ and the wave propagation factor $c = 1$ we can rewrite Eq. 1 with 2 to

$$[1, -2, 1]_t \mathbf{u} = [1, -2, 1]_x \mathbf{u} \tag{3}$$

Reorganizing Eq. 3 to $\mathbf{u}[t+1]$ yields the well known explicit finite difference equation

$$\mathbf{u}[t+1, x] = [1, -2, 1]_x \mathbf{u}[t, x] - u[t-1, x] + 2 \cdot \mathbf{u}[t+1, x] \tag{4}$$

For the boundary value problem on the domain x = x $\in \{x|0 <= x <= L\}$ with $\mathbf{u}[t, 0] = 0$ and $\mathbf{u}[t, L] = 0$ the first term on the right side of the equality can be rewritten as a convolution

$$[1, -2, 1]\mathbf{u}[t, x] \equiv \{1, -2, 1\} * \mathbf{u}[t, x] \tag{5}$$

over the domain x, with $*$ denoting a convolution and $\{\}$ a vector. Now, using the convolution theorem and properties of the Fourier Transform it is possible to perform this time domain convolution as a multiplication in the frequency domain. With $\mathcal{F}\{\}$ and $\mathcal{F}\{\}^{-1}$ the Fourier Transform and the Inverse Fourier Transform respectively we can rewrite Eq. 4 to

$$\mathbf{u}[t+1, x] = \mathcal{F}\{\mathcal{F}\{\{1, -2, 1\}\} \cdot \mathcal{F}\{\mathbf{u}[t, x]\}\}^{-1} - \mathbf{u}[t-1, x] + 2 \cdot \mathbf{u}[t+1, x] \tag{6}$$

Now we have to recapitulate some properties of the finite difference operator δ_x^2. The centered difference approximates the derivative at a local point $u[x]$ using a Taylor expansion around x. Because we use the centered approximation which is composed of the left-sided and the right-sided finite difference approximation, the error is of second order. If we use higher order approximations of the finite difference operator we thereby minimize the error term proportionately to the

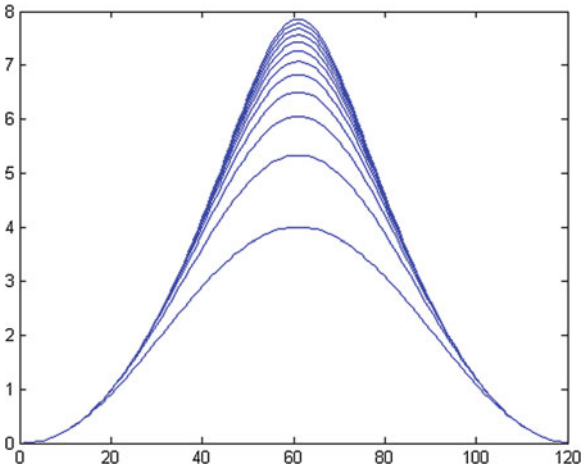


Fig. 1 Absolute value of Fourier transformed FD-operators with increasing order. From bottom $N = 2$ to $N = 64$ the topmost

approximation order N. A fourth order approximation has an error of order $\mathcal{O}_4$ and so on. If we approximate the derivative over the whole domain with finite differences using the highest possible order, we can minimize approximation errors up to the order of discrete points -1.[5] The fourth order central FD operator for the second derivative is given as

$$\delta_x^2 = \frac{\left[-\frac{1}{12}, \frac{4}{3}, -\frac{5}{2}, \frac{4}{3}, -\frac{1}{12}\right]}{\Delta x^2} \tag{7}$$

with an error of $\mathcal{O}_4$. If we now increase the order of the approximation we can get a high order FD approximation extending over the whole domain. Following Eq. 6 we can perform the convolution in frequency space as a multiplication with Fourier transformed FD operators $\widehat{\delta_x^2} = \mathcal{F}\{\delta_x^2\}$. As one can see in Fig. 1 with increasing order of the FD operator function $\widehat{\delta_x^2}$ approaches the analytical derivation function $\hat{\phi}$.

2.2 Symplectic Integrators

Numerical methods that exhibit symplectic properties are among the oldest solution methods for DEs and have been discovered and rediscovered throughout the centuries (Hairer et al. 2002, 2003) many times. Even though some advantageous properties where noted by physicists, like Newton or Störmer (Hairer et al. 2003) the concept of symplecticity was discovered only in the 1980s (Hairer et al. 2002)

[5] If this is physically correct is a question that can not be answered here. The analytical wave equation has no speed limit for transported information. Whereas discrete approximations have a speed limit set by the discretisation step width in space and time. This is the stability condition expressed by Courant et al. (1928).

and systematically researched over the last two decades. At first, symplectic methods were only developed for ODEs but since the end of the 1990s were extended to PDEs (Mclachlan 1994). Numerical methods are called symplectic if they preserve the Hamiltonian character of the Differential Equation (Hairer et al. 2002). In this section we present some basic features of symplectic methods and take a simple harmonic oscillator (SHO) as an example.

The ODE of the SHO can be written as

$$x_{tt} = -\omega^2 \cdot x \tag{8}$$

This equation can either be solved using a direct discretisation of the differential term on the left-hand side using a central FD approximation (Bilbao 2009) or using a P-Q splitting method (Mclachlan 1994) separating the Hamiltonian into

$$\mathcal{H} = T(p) + V(q) \tag{9}$$

with T the kinetic energy and V the potential energy. This results in methods of the form

$$p^{t+1} = p^t - \Delta t \cdot H_q\left(p^{t+1}, q^t\right) \tag{10}$$

$$q^{t+1} = q^t + \Delta t \cdot H_p\left(p^{t+1}, q^t\right) \tag{11}$$

with H_q and H_p vectors of the partial derivatives of H in respect to q and p and Δt the discretisation step width. Equation 11 approximates the true Hamiltonian flow for every time step explicitly and is called Symplectic Euler (SE) algorithm, one of the most widely used symplectic algorithms.

Another well-established method is called Verlet (Position Verlet/Velocity Verlet) method[6] (Verlet 1967; Hairer et al. 2003; Sha et al. 2008). The Velocity Verlet (VV) can be written as

$$p^{t+1/2} = p^t - \frac{\Delta t}{2} \cdot H_q\left(p^{t+1/2}, q^t\right) \tag{12}$$

$$q^{t+1} = q^t + \frac{\Delta t}{2} \cdot \left(H_p\left(p^{t+1/2}, q^t\right) + H_p\left(p^{t+1/2}, q^{t+1}\right)\right) \tag{13}$$

$$p^{t+1} = p^{t+1/2} - \frac{\Delta t}{2} \cdot H_q\left(p^{t+1/2}, q^{t+1}\right). \tag{14}$$

As one can see in comparison to the SE method, we have one evaluation step more. In practice this means a second force evaluation per time-step which leads to a more accurate calculation of the velocity (pulse) p and position q.

Figure 2 shows a comparison of the SE and the VV algorithm with the non-symplectic Euler method with increasing time discretisation step size.

[6] The Velocity Verlet method is similar to the Leapfrog algorithm.

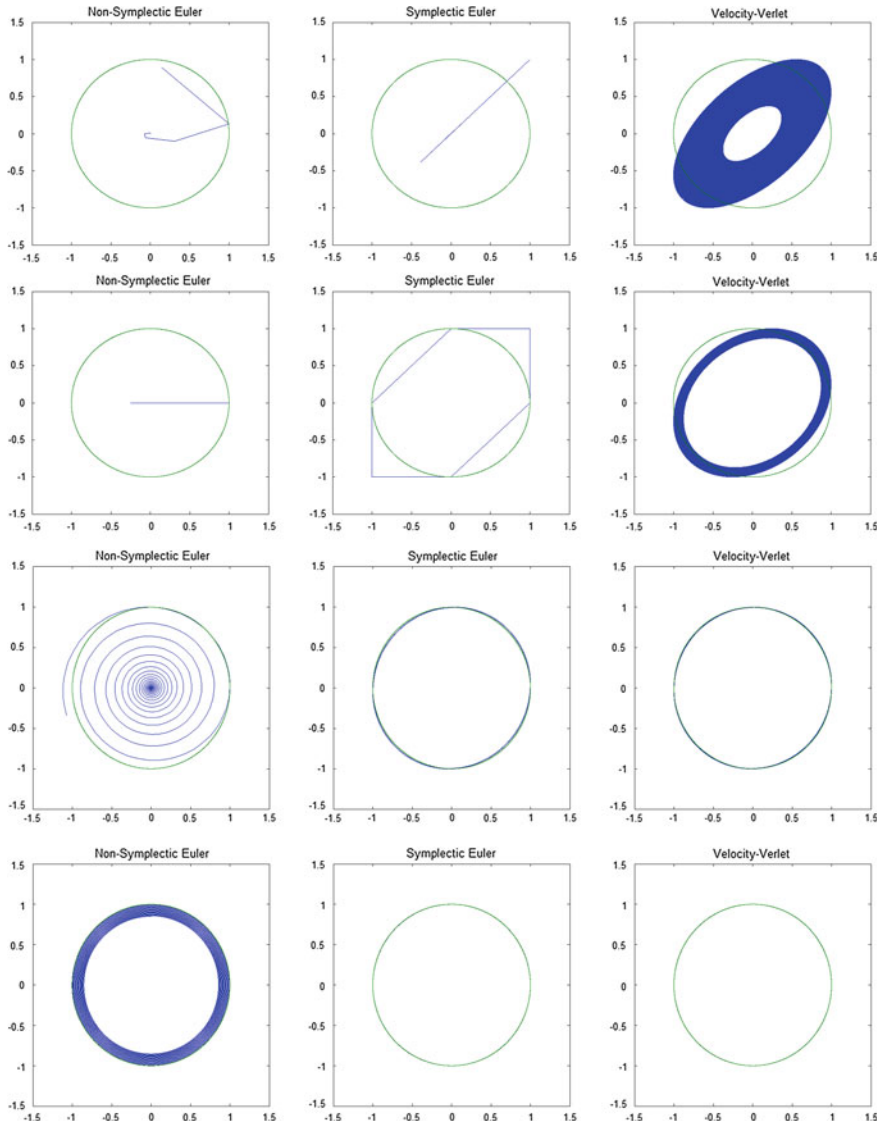


Fig. 2 Comparison of numerical methods (*blue*), analytical solution (*green*). *Row 1* $\omega^2 \cdot \Delta t = 5$, *Row 2* $\omega^2 \cdot \Delta t = 1$, *Row 3* $\omega^2 \cdot \Delta t = 0.005$, *Row 4* $\omega^2 \cdot \Delta t = 0.00005$

As one can clearly see, the VV algorithm (2. order symplectic) is stable for all time step sizes and the most accurate compared to the analytical solution. The SE (1. order symplectic) is stable for step sizes that satisfy the relation $\Delta t \leq \frac{1}{\sqrt{\omega^2}}$. The non-symplectic Euler shows the poorest behaviour of all three methods and is usable only for very small discretisation step widths.

3 Multisymplectic PS-FD Examples

3.1 1-dimensional Wave Equation

The motion of a string under small deflection and with linear mass distribution can be described by the 1-dimensional wave equation also known as the d'Alembert equation. In this section we present the model of a single linear string and show properties of the proposed method. Using a multisymplectic integrator for Eq. 1 with boundary conditions $u(0,t) = u(L,t) = 0$ as described in Moore and Reich (2003a), Feng and Qin (1987) and in Sect. 2 we get an explicit time stepping method called the Euler Box Scheme, similar to (11) as presented in Moore and Reich (2003a). As already mentioned there are several well researched symplectic algorithms for this kind of problem. The accuracy of the used algorithm depends in most cases on the number of force evaluations which directly influence the computational cost. In practical terms one can say: The more force evaluations performed per time step the more accurate is a method. But with more force evaluations the computational complexity rises. For this work three different numerical integrators are compared for solving the equations of motion. Besides a straightforward time-domain implementation all three methods are implemented using PS methods. For calculating the Fourier Transform of the PS part of the algorithm standard fft-methods are used.[7]

Because of the linear properties of the symplectic methods within one time-step, the values for p and q can be calculated in the time domain or the frequency domain giving similar results. This property is utilised to reduce the number of Fourier Transforms per time step. Table 1 lists a comparison of the implemented methods for the 1-dimensional wave equation.

Figure 3 shows the movement of a linear string discretised with 128 points and calculated with a Spectral Symplectic Euler method. The Blue string is computed with a spectral FD-kernel with $N = 128$ points, the green line uses a kernel of $N = 3$ points, which is the standard second order central FD grid as in Eq. 2.

As one can see, the string with the larger FD-kernel shows a smaller dissipation compared to the initial triangular excitation of the string. A comparison with the analytical solution of the wave equation for the 1-dimensional string with the presented initial value is not reasonable here because the movement of an ideal string is not a movement found in real strings of musical instruments. Nonetheless, for classifying the results gained with PS methods one can compare the frequency drift of the higher partials commonly found in FD methods (Bilbao 2009).

[7] In MATLAB the build-in fft()-function is used. For the C++ implementation the fftw-library (Frigo and Johnson 1997, 1998), the CUFFT-library (NVIDIA CUDA 2013) and a template based Fast Fourier Transform is used.

Table 1 Symplectic PS algorithms

Name of method	Force eval.	Fourier trans.
Symplectic Euler	1	2
Newton-Störmer-Verlet	2	4
PEFRL	3	6
Spectral Symplectic Euler	1	1
Spectral Newton-Störmer-Verlet	2	1

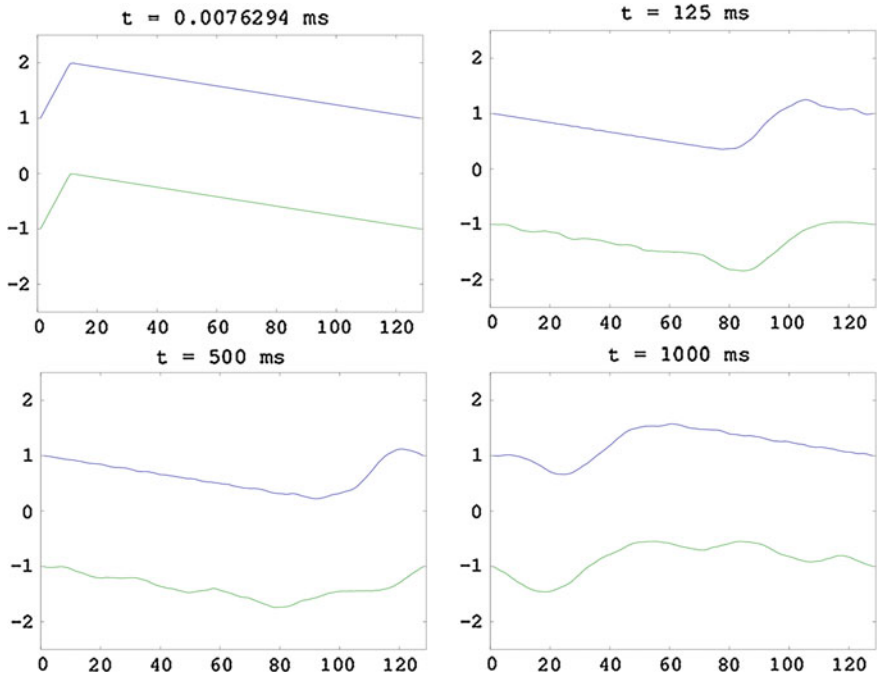


Fig. 3 Time series of PSFD-string with 128 discrete points. *Blue* 128-point FD grid, *Green* 3-point FD-grid

Table 2 Frequency drift of PS kernels of different order

Order N	1. partial	4. partial	8. partial	16. partial	20. partial
Ideal	2	5	9	17	21
3	1.9991	4.9809	8.9554	16.8599	20.7771
128	2.0	4.9936	8.9904	17.0191	21.0541

Table 2 shows the beneficial effect of a higher order spectral kernel on the accuracy of the higher partials. This minimises the effect of detuning of higher modes commonly found in FD solutions of the 1-dimensional wave equation i.e. the linear string.

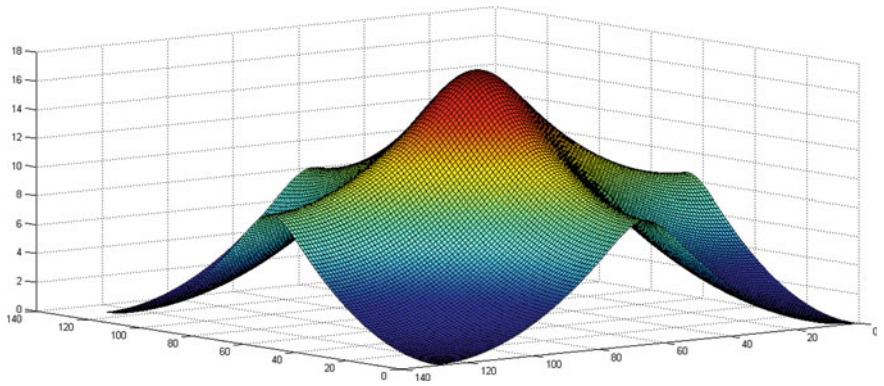


Fig. 4 Absolute value of Fourier transformed 2-d FD-operator

3.2 2-dimensional Wave Equation

Following the same multisymplectic discretisation scheme as used in the preceding section like in Moore and Reich (2003a, b), called the Euler Box Scheme, the differential equation can be integrated numerically with three time steps similar to algorithm 11 as shown above. The only difference between the two algorithms is the extension to two dimensions (integrating in the x and y direction) and a 2-dimensional version of the integration kernel. A central finite difference approximation of the space discretisation yields a 2-dimensional convolution kernel. The Fourier transformed kernel is depicted in Fig. 4. Examples of the 2-dimensional membrane can be seen in the next section in Fig. 5.

4 Coupled Geometries

Following the methodology described above, two simplified versions of musical instruments are modeled in MATLAB and C++ programming language. The models consist of strings and of a sound radiating front plate and are presented as a proof of concept and not as complete physical models of the whole geometry as in other works (Pfeifle and Bader 2011a, b, 2011c, 2012a). The modeled instruments are a five-string banjo and a violin. The banjo is used as a starting point for the following model because its "simple" geometry (basically a string coupled to a round membrane) makes it an instructive device for testing basic concepts of PM for musical instruments. The second instrument model is chosen to show that the method also works for higher order differential equations with non-linear mass distribution (the wooden front plate of the violin with orthotropic material properties) and for non-linear excitation (the interaction of the violin bow with the string). All instrument parts are modeled with the method described above. Sounds

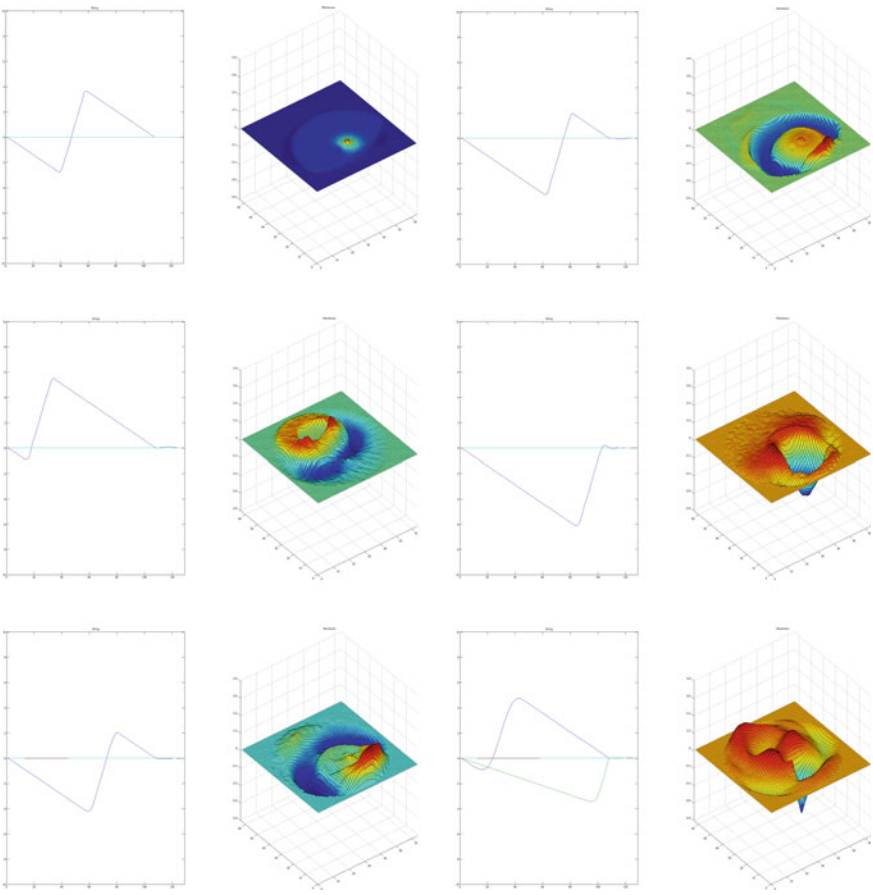


Fig. 5 Time series of a banjo model including PSFD-strings coupled to a PSFD-membrane

from both models and other instruments can be found at the *Systematic Musicology* web-site of the Institute of Musicology, Hamburg (Pfeifle and Bader 2013).

4.1 Model of a Banjo

The first instrument model is an American five-string banjo, consisting of five strings, a simplified bridge model and the membrane. Two main parts of the model, the string and the membrane have been described thoroughly in Sect. 2. The interesting part of this model is the coupling between the string and the membrane. Changes in strength and position coupling between the string and the membrane strongly influence the timbre and characteristic of the radiated sound. As we have explicit expressions for the force, velocity and deflection of the string

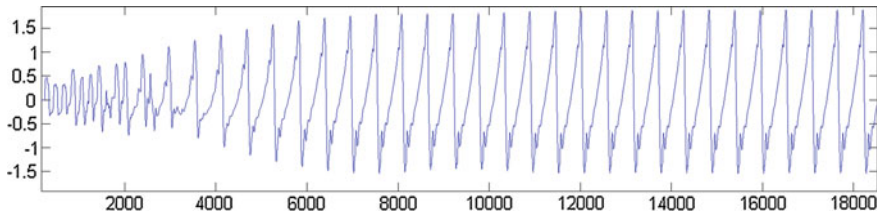


Fig. 6 The first 76 ms of a violin tone

and membrane of every point of the model, we can couple string and membrane via the impedance at the contact point as described in Bader (2005a). The characteristic motion of the transient movement of the membrane due to excitation by the string can be seen in Fig. 5. As one can see, the pulse travelling on the string is transferred via the bridge to the membrane.

4.2 Model of a Violin

The model of the violin consists of four strings, a bridge model similar to the bridge of the banjo and the front plate of the violin with two sound holes. The model for the string/violin-bow interaction is based on a model presented in Bader (2005b) and extended as presented in Pfeifle and Bader (2011a). The non-linear excitation of the violin string is one of the main reasons for the characteristic violin sound and the expressiveness of the violin. For this physical model the string is implemented as presented before but has additional conditions depending on the state of the string bow interaction. The first characterisation of the violins string movement was done by Hermann von Helmholtz (1896). In his honour, the generic motion of the string is called Helmholtz motion. The first 20,000 samples of the violin model is shown in Fig. 6.

As one can see, at the beginning of time series the string is deflected in a non-periodic manner. This is the initial scratching sound of the violin. After a short time, the model settles in the well-known Helmholtz motion as expected.

4.3 Sound Radiation

The sound radiation from the front-plate of each instrument is integrated into a virtual room and picked up at two positions approximately 60 cm above the instrument and 16 cm apart.[8] The integration includes all sound radiating points

[8] Mean length between two ears.

from the front plate and weight them depending on the respective position and distance to the receiver point. Sounds are available at Pfeifle and Bader (2013).

5 Conclusion

In this work we have presented a methodology for calculating physical models of musical instruments with high-order PSFD discretisation. Several positive features of this method, like smaller error and a stable long-term behaviour of the models could be shown. Although there are still open questions concerning multisymplectic methods from a mathematical point of view, this work has presented some insights into these methods for musical sound synthesis and computation of physical models. To the best of my knowledge this has been the first attempt to use multisymplectic PSFD methods for PMs of musical instruments so there is still a lot of research to be done in the future. A very important question in this scope is if the change in sound quality and timbre of higher order PS methods is perceivable by musicians and non-musicians. Another research question concerns the applicability of the proposed methods for specialised hardware implementations on a GPU or a FPGA. Furthermore this method can be extended to other instruments or other musicological problems.

References

Ascher, U. M., & McLachlan, R. I., (2004). Multisymplectic box schemes and the kortewegde vries equation. *Applied Numerical Mathematics*, *48*(34), 255–269. Workshop on Innovative Time Integrators for PDEs.

Bader, R. (2005, Oct) *Computational Mechanics of the Classical Guitar*. New York: Springer.

Bader, R. (2005). Whole geometry finite-difference modeling of the violin. *Proceedings of the Forum Acusticum*, *2005*, 629–634.

Bathe, K. J. (2002). *Finite-Element Methoden*. New York: Springer.

Bilbao, S. (2007). Robust physical modeling sound synthesis for nonlinear systems. *Signal Processing Magazine, IEEE*, *24*(2), 32–41. march.

Bilbao, S. (2009). *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics*. Chichester, UK: Wiley.

Chaljub, E., Komatitsch, D., Vilotte, J. P., Capdeville, Y., Valette, B., & Festa, G. (2007). Spectral-element analysis in seismology. In V. Maupin & R. S. Wu (Eds.), *Advances in Wave Propagation in Heterogenous Earth* (pp. 365–419). Amsterdam: Elsevier. volume 48 of Advances in Geophysics.

Courant, R., Friedrichs, K., & Lewy, H. (1928). Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische Annalen*, *100*, 32–74.

Feng, K., & Qin, M. (1987). The symplectic methods for the computation of hamiltonian equations. In Y. Zhu & B. Guo (Eds.), *Numerical Methods for Partial Differential Equations* (pp. 1–37). Heidelberg: Springer. volume 1297 of Lecture Notes in Mathematics.

Frigo, M., & Johnson, S. G. (1997). *The fastest Fourier transform in the west*. Technical Report MIT-LCS-TR-728, Massachusetts Institute of Technology, September 1997.

Frigo, M., & Johnson, S. G. (1998). FFTW: An adaptive software architecture for the FFT. In *Proceedings 1998 IEEE Intlernational Conference on Acoustics Speech and Signal Processing* (Vol. 3, pp. 1381–1384). IEEE.

Fornberg, B. (1990). High-order finite differences and the pseudospectral method on staggered grids. *SIAM Journal on Numerical Analysis*, *27*(4), 904–918.

Fornberg, B. (1998). *A practical guide to pseudospectral methods*. Cambridge: Cambridge university press.

Giordano, N. (2006). Finite-difference modeling of the piano. *The Journal of the Acoustical Society of America*, *119*, 3291.

Gottlieb, D., & Orszag, S. A. (1987). *Numerical analysis of spectral methods: theory and applications*. Philadelphia: Society for, Industrial Mathematics.

Hairer, E., Lubich, C., & Wanner, G. (2002). *Geometric numerical integration : structure-preserving algorithms for ordinary differential equations*., Springer series in computational mathematics. Berlin: Springer.

Hairer, E., Lubich, C., & Wanner, G. (2003). Geometric numerical integration illustrated by the Stoermer-Verlet method. *Acta Numerica*, *12*, 399–450.

Hamman, C. W., Kirby, R. M., & Berzins, M. (2007). Parallelization and scalability of a spectral element channel flow solver for incompressible navier-stokes equations. *Concurrency and Computation: Practice and Experience*, *19*(10), 1403–1422.

Komatitsch, D., & Tromp, J. (2002). Introduction to the spectral element method for three-dimensional seismic wave propagation. *Geophysical Journal International*, *139*(3), 806–822.

Komatitsch, D., Tsuboi, S., Tromp, J., et al. (2005). The spectral-element method in seismology. *Geographical Monograph-American Geophysical Union*, *157*, 205.

Kong, L., Liu, R., & Zheng, X. (2007). A survey on symplectic and multi-symplectic algorithms. *Applied Mathematics and Computation*, *186*(1), 670–684.

Lee, U., Kim, J., Leung, A. Y. T., et al. (2000). The spectral element method in structural dynamics. *Shock and Vibration Digest*, *32*(6), 451–465.

Lesage, A. C., Zhou, H., Araya-Polo, M., Cela, J. M., Ortigosa, F. (2008). 3d reverse-time migration with hybrid finite difference-pseudospectral method.

Lyons, W., Ceniceros, H. D., Chandrasekaran, S., & Gu, M. (2005). Fast algorithms for spectral collocation with non-periodic boundary conditions. *Journal of computational physics*, *207*(1), 173–191.

Mclachlan, R. (1994). Symplectic integration of Hamiltonian Wave Equations. *Numerical Mathematics*, *66*, 465–492.

McLachlan, R. I., Reinout, G., & Quispel, W. (2006). Geometric integrators for ODEs. *Journal of Physics A: Mathematical and General*, *39*, 5251–5285.

Moore, B. E. (2009). Conformal multi-symplectic integration methods for forced-damped semi-linear wave equations. *Mathematics and Computers in Simulation*, *80*(1), 20–28.

Moore, B., & Reich, S. (2003). Backward arror analysis for multi-symplectica integration methods. *Numerische Mathematik*, *95*(4), 625–652.

Moore, B. E., & Reich, S. (2003). Multi-symplectic integration methods for Hamiltonian PDEs. *Future Generation Computer Systems*, *19*(3), 395–402.

Nvidia, CUDA. (2013). *Compute unified device architecture*. Available at https://developer.nvidia.com/category/zone/cudazone,. Accessed 2 Jan 2013.

Patera, A. T. (1984). A spectral element method for fluid dynamics: laminar flow in a channel expansion. *Journal of Computational Physics*, *54*(3), 468–488.

Peiró, J., & Sherwin, S. (2005). Finite difference, finite element and finite volume methods for partial differential equations. *Handbook of Materials Modeling*, 2415–2446.

Pfeifle, F., & Bader, R. (2011). Real-time finite-difference string-bow interaction field programmable gate array (fpga) model coupled to a violin body. *The Journal of the Acoustical Society of America*, *130*(4), 2507–2507.

Pfeifle, F., & Bader, R. (2011). Measurement and physical modelling of sound hole radiations of lutes. *The Journal of the Acoustical Society of America*, *130*(4), 2507–2507.

Pfeifle, F., & Bader, R. (2011). Nonlinear coupling and tension effects in a real-time physical model of a banjo. *The Journal of the Acoustical Society of America*, *130*(4), 2432–2432.

Pfeifle, F., & Bader, R. (2012). Measurement and analysis of sound radiation patterns of the chinese ruan and the yueqin. *The Journal of the Acoustical Society of America*, *131*(4), 3218–3218.

Pfeifle, F., & Bader, R. (2012b). Real-time finite difference physical models of muical instruments ona field programmable gate array (fpga). In *Proceedings of the International Conference on Digital Audio Effects (DAFx-12)* (pp. 63–70), York, UK, Sept. 17–21.

Pfeifle, F., & Bader, R. (2013). *Systematic musicology hamburg*. Available at http://www.systmuwi.de/muwi_research_Physical_Modeling_of_Musical_Instruments.html,. Accessed 02 Jan 2013

Sathej, G., & Adhikari, R. (2008). *The eigenspectra of indian musical drums*. arXiv, preprint arXiv:0809.1320.

Schober, C. M., & Wlodarczyk, T. H. (2008). Dispersive properties of multisymplectic integrators. *Journal of Computational Physics*, *227*, 5090–5104.

Sha, W., Huang, Z., Chen, M., & Wu, X. (2008). Survey on symplectic finite-difference time-domain schemes for maxwell's equations. *Antennas and Propagation, IEEE Transactions on*, *56*(2), 493–500. feb.

Spa, C., Garriga, A., & Escolano, J. (2010). Impedance boundary conditions for pseudo-spectral time-domain methods in room acoustics. *Applied Acoustics*, *71*(5), 402–410.

Tong, M. & Krozer, V. A. (2002). *Non-Uniform Pseudo-Spectral Time Domain (PSTD) Method in One-Dimensional Applications*.

Trefethen, L. N. (2000). *Spectral methods in MATLAB* (Vol. 10). Society for, Industrial Mathematics.

Verlet, L. (Jul 1967). Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, *159*, 98–103.

von Helmholtz, H. (1896). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Friedrich Vieweg und Sohn. Available at http://www.uni-leipzig.de/psycho/wundt/opera/helmhltz/toene/TonEmpIn.htm

Wang, Z. J. (2002). Spectral (finite) volume method for conservation laws on unstructured grids. basic formulation: Basic formulation. *Journal of Computational Physics*, *178*(1), 210–251.

Human–Computer Interaction and Music

Isabel Barbancho, Alejandro Rosa-Pujazón, Lorenzo J. Tardón and Ana M. Barbancho

1 Introduction

In this document, we offer some insight into the potential of combining advanced interaction paradigms with sound and music for the development of innovative interactive audio applications. Therefore, in this chapter we present novel ways to interact with the music by means of using advanced natural human computer interfaces. Furthermore, the basics of specific applications which are currently being developed will be briefly presented. A motion-based paradigm is considered in the context of music signal processing to provide an innovative and immersive experience in audio and music applications.

1.1 Review of Previous Works

Human Computer Interfaces have long moved beyond the conventional setting of a single user sitting in front of a desktop computer. Nowadays, the latest technological advances in the fields of motion tracking or speech recognition have allowed for the definition of more complex, enriching interaction metaphors. Also, the development and proliferation of low-cost off-the-shelves devices, such as Microsoft's *Kinect* camera or the Nintendo *Wiimote*, make it possible to provide more intuitive ways of interaction with a given computing device (Villaroman et al. 2011).

An adequate fusion of auditory and visual cues is critical in the conception of natural human computer interfaces (Shivappa et al. 2010), since vision and hearing are the primary senses used by humans to comprehend and interact with the world. Thus, in order to create a fully immersive experience, it is necessary to carefully

I. Barbancho (✉) · A. Rosa-Pujazón · L. J. Tardón · A. M. Barbancho
Department of Ingeniería de Comunicaciones, Universidad de Málaga, Málaga, Spain
e-mail: ibp@ic.uma.es

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_16,

design the presentation of audiovisual information. As such, music is an integral part of the auditory modality. Moreover, musical interaction itself can constitute the basis and main focus of a certain interface for a wide assortment of applications: music learning (Wang and Lai 2011), musical instrument creation/simulation (Jorda 2010), music-guided rehabilitation (De Dreu et al. 2012), etc.

One of the most well-known types of applications regarding musical interaction is that of videogames. Games such as *GuitarHero* (GuitarHero 2011), *SingStar* (SingStar 2004) or *RockBand* (RockBand 2012) are some of the most prominent and popular applications of this kind, especially on western markets. *Guitar Hero III: Legends of Rock* became the first single game ever to surpass $1 billion in sales (Craft 2009). This proves the economic relevance of this genre. However, an analysis of 20 releases showed that these rarely respect the creative component of actually playing music (Grollmisch et al. 2009). Mostly, the player has to rhythmically trigger buttons according to predefined sequences. This mostly requires dexterity and swift reflexes, whereas the possibilities to actively participate in music creation are only marginal.

Some rip-offs of the most popular titles exist in the open source world with titles like *FretsOnFire* (FrestsOnFire 2006), *UltraStar* (UltraStar 2010). It is remarkable that the editing of the songs (alignment of notes and lyrics with the songs) for these titles has to be done manually, this is why *UltraStar* features a built in editor. For commercial titles, game studios use proprietary editors and formats. There are only a few attempts to use automatic music processing in these kinds of games. In (Barbancho et al. 2009) a description on how to generate game packages for *FretsOnFire* is presented. *AudioSurf* (AudioSurf 2008) or audio processing as proposed in (Migneco et al. 2009) are examples of action games based on automatic, real-time analysis of music features.

Nevertheless, it would be definitely unfair to judge music games as being nonmusical. They do foster the interest of children and adolescents in music, some gamers are even encouraged by these games to learn a real instrument. For example, the Beatles Rock Band Edition (*BeatlesRB* 2009) allows up to three singers to sing in harmony and rewards the players with high scores if they succeed. Such features are somewhat useful for self-study voice education (*Mikestar* 2009). Furthermore, recent studies seem to confirm that these kinds of applications can actually help children develop musical skills and knowledge (Gower and McDowall 2012). Thus, despite their limitations, they can serve as a gateway to musical concepts for potential users, especially with regards to rhythm.

Musical interaction has recently become a hot topic. Antle et al. (Antle et al. 2008) proposed a system to connect body movements to output sounds. The experiments conducted showed that learning processes could be improved through the use of such interaction metaphors. Further research was performed regarding the use of tangible interaction in order to manipulate the pitch, volume and tempo of ongoing tones (Bakker et al. 2011). Another study (Holland et al. 2010) made use of a haptic vibrotactile device set (the Haptic Drum Kit) to aid the learning process of rhythm and rhythmic patterns.

A good example of musical exploration and creation using physical body movements as an interface can be found in the works by Khoo et al. (Khoo et al. 2008). Their study depicts a system capable of mapping motion features onto acoustic parameters (such as frequency, amplitude, tempo, or time signature) for real-time expressive rendering of a given piece of music. At the same time, visual feedback was projected accordingly on a screen. Therefore, this system allowed users to interact with music in an intuitive and explorative way, lowering the barriers towards the understanding and appreciation of music features. A similar interaction metaphor was considered in (Castellano et al. 2007). In a similar fashion, it is possible to modify the visual patterns presented by means of speech or sung voice, making the voice "visible", as shown in the works conducted by Levin and Lieberman (Levin and Lieberman 2004), or directly interacting with a virtual character (Taylor et al. 2005; Mancini et al. 2007).

Musical interaction can act as a strong motivator, but it can also fulfill an important role as a way to allow users to acquire or expand their knowledge on music theory. In general terms, the mechanics and abstract concepts of music are not usually known to most lay people. Furthermore, in order to learn the different aspects of music theory it is necessary to devote a considerable amount of time to such purpose. On the other hand, visitors to physical museums are often overwhelmed by the vast amount of information available (Karimi et al. 2012). In this regard, musical interaction allows for a learning-by-action exploration, lowering the barriers of the inherent abstract nature of many of these concepts and making the global experience much more accessible and enjoyable.

1.2 Chapter Organization

As previously stated, the aim of this text is to portray how innovative user-computer interfaces can be used to provide new experiences and forms of interaction with music. More concretely, Sect. 2 presents a brief overview of the technologies available for the design and implementation of interaction paradigms revolving around human body motion tracking. After the different alternatives have been identified, the chapter will cover the option chosen to implement the natural human–computer interface as well as the justification behind this decision and a succinct review of the capabilities for motion tracking of the Open Natural Interaction (*OpenNI*) framework. Then, Sect. 3 will present two applications which use motion-tracking to allow the user to modify some musical features in the sounds played. In addition, an overview of alternative motion-based interaction paradigms will be described in order to more widely show the potential of these interaction techniques in musical applications. Finally, Sect. 4 will briefly present the conclusions gathered from this work.

2 Motion Tracking: A Camera-Based Approach

2.1 Technologies for Motion Tracking

In the next lines, a brief summary of the most commonly used motion capture technologies is given. As it is not the scope of this chapter to exhaustively cover the distinct technological aspects behind these technologies, only the most prominent and relevant aspects will be summarized. For further details, the following works should be reviewed: Welch and Foxlin (Welch and Foxlin 2002), Baratoff and Blanksteen (Baratoff and Blanksteen 1993), Perry et al. (Perry et al. 1997).

Traditionally, the most commonly used alternatives for motion tracking can be summarized into the following five categories, depending on the main technology employed: optical, electromagnetic, acoustic, inertial and mechanical sensors.

- Optical trackers in general show high update rates and short lags. The data acquired are usually quite accurate, and electromagnetic noise and room temperature have little to no effect on them. However, they suffer from the line of sight problem: any obstacle between sensor and source will create an occlusion which can seriously degrade the overall performance of the tracking system. Ambient light and infrared radiation can also adversely affect the performance. As a result, the environment must be carefully designed to reduce these interferences as much as possible.
- Electromagnetic trackers have been quite popular, but they can give inaccurate measures. They usually show latency problems and they are also very sensitive to the presence of large amounts of metal and conductive materials in the surrounding area or other electromagnetic fields, such as those that would be generated by large computer equipment, displays, etc. They have some important advantages, however, such as not requiring direct line-of-sight between the trackers themselves and the source, and they usually have a small and ergonomic size.
- Acoustic sensors lie somewhat in-between the previous ones. They have a restricted workspace volume and require direct line-of-sight (although they are not as affected by this as optical trackers). Time-of-flight trackers usually have a low update rate, and phase-coherence trackers are subject to error accumulation over time. Additionally and more importantly, both types are affected by temperature, pressure changes and the humidity level of the environment. Furthermore, the effects of echoing and the presence of other nearby sonic sources can have very negative effects on their measurement.
- Inertial systems are already available in chip form. Inertial sensors are completely self-contained. There are no line-of-sight requirements, no emitters to install, and no sensitivity to interfering electromagnetic fields or ambient noise. They also have very low latency (typically a couple of milliseconds or less) and high sampling rates. The main disadvantage of inertial trackers is that they are

very sensitive to drifting errors, since they do not actually measure position and/or orientation directly, which makes them ineffective when absolute measures are needed.
- Mechanical trackers offer many advantages, such as almost complete environment independence, and potentially highly accurate, fast and low latency measurements. Unfortunately, mechanical trackers are very cumbersome. They can be bulky, heavy, and severely limit the motion of the user, giving rise to ergonomic issues.

As a conclusion, there is no optimal solution when it comes to making a choice. The advantages and disadvantages of each technology must be carefully assessed according to the requirements of the task at hand.

We have decided to use a camera-based system. A camera-based tracking system can be seen as a passive subtype of the optical tracking technology, since the tracking points of interest will not actually emit light by themselves, instead the system will rely on the light reflected on natural surfaces. We have followed this approach because it offers an off-the-shelf, inexpensive solution that minimizes intrusiveness (Gleicher and Ferrier 2002), constituting a good solution to implement high precision motion tracking. Nevertheless, while humans can easily identify motion from a recorded sequence telling from their own experience in the real world, computers face some difficulties (Gleicher and Ferrier 2002), for mapping a 3D world into a 2D movie involves an important loss of information. In order to circumvent such limitation, we have opted to use a Z-camera, that is, a camera that is capable of capturing the 3D scene, by means of a depth map which assigns to each pixel a value related to the distance between the camera and the point in the 3D scene which was projected onto that pixel. In particular, a Microsoft *Kinect* sensor is used.

2.2 The OpenNI Framework

Kinect is a composite device consisting of an infrared (IR) projector, an IR camera and a RGB camera. The infrared camera projects a pattern of infrared points which are then used to triangulate points in space (Smisek 2011), thus allowing to determine the depth of the objects in the scene. In November 2010, the company *Primesense* released their own open source drivers, as well as a middleware motion tracking module called *NITE*, both integrated as part of the *OpenNI* framework. By using the *OpenNI/NITE* framework, it is feasible to achieve advanced user identification and motion tracking capabilities. This section aims to briefly present the key features to consider in the development of our interface using a *Kinect* device.

The purpose of the *OpenNI* framework is to provide the means to develop human–computer interfaces that achieve "natural interaction", that is, an interaction based on human senses, especially on human hearing, motion and vision. In

order to actually accomplish this objective, it is necessary to resort to hand and body gestures and motion tracking, speech and voiced command recognition, etc. *OpenNI* offers a flexible framework to make use of natural interaction devices (such as 3D sensors) and the middleware needed to control those devices.

The layer view of the *OpenNI* framework is shown in Fig. 1. It includes complex components, such as Production Nodes, Production Chains or Capabilities. However, it is not the aim of this section to thoroughly cover the possibilities of the *OpenNI* API. Let it suffice to say that a Production Node is a component that outputs a certain type of data to be used by either other Production Nodes or the application itself. Examples of such components would be an Image Generator (which generates colored image-maps) or a Depth Generator (which produces a depth map). In the particular case of the *Kinect* sensor module, both the RGB data and depth image data are accessible through the use of Production Nodes, where examples are shown in Fig. 2.

One of the key features of the *OpenNI/NITE* framework is the Production Node called Scene Analyzer. This component analyzes a depth image produced by a lower-level node to look for figures in it. Thus, the foreground, background, the floor plane, as well as human figures are identified (PrimeSense 2011a). The data are then processed into a labeled depth map as output. *NITE* processes this data to generate a label map that allows for specific user tracking in the scene, using a user segmentation algorithm (PrimeSense 2011b). This label map holds information regarding the detection of human shapes, in such a way that each pixel is associated

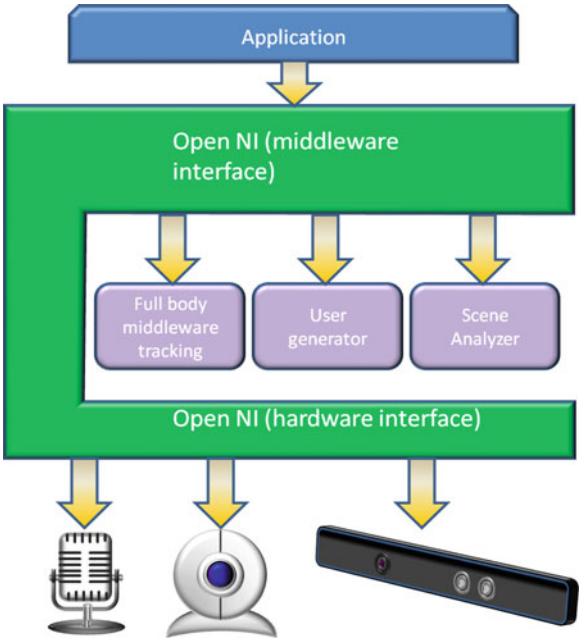


Fig. 1 An example of the abstract layer view of *OpenNI* (based on image in (PrimeSense 2011a))

Fig. 2 Example of Kinect's production nodes output, depth image in rainbow scale (*left*) and RGB image (*right*)

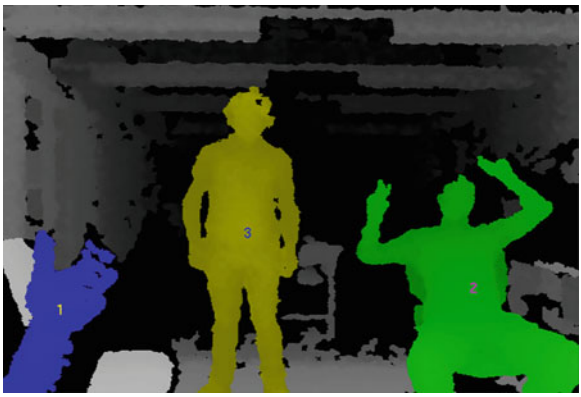


Fig. 3 User segmentation labelled depth image, identifying three different users from non-user pixels (background)

with a label which in turn indicates whether that pixel corresponds to the background or to one of the figures identified. An example can be found in Fig. 3.

Finally, the most critical feature for the implementation of a human–computer interface for musical interaction applications could be considered to be *NITE's* skeletal tracking production node (PrimeSense 2011b).

NITE's skeletal model makes use of a total of 15 skeletal nodes (of 24 possible nodes supported by the *OpenNI* API). The correspondence of such nodes with human joints can be seen in Fig. 4. This production node processes the labeled depth image generated by the user segmentation node to calculate the skeletal joints position and orientation values for the fifteen nodes of one of the users identified in the previous step. Note that joint position coordinates are far more stable than joint orientation ones, which are actually quite noisy. Thus, it is rather preferable to rely on joint position measures to prevent incorrect measures.

After a sufficiently stable model of the user's joint positions has been found, it is possible to implement different kinds of interactive models based on the motion detected for each of the nodes.

In the next section, we will cover the implementation of specific examples of musical interaction by means of skeletal motion tracking and musical signal processing.

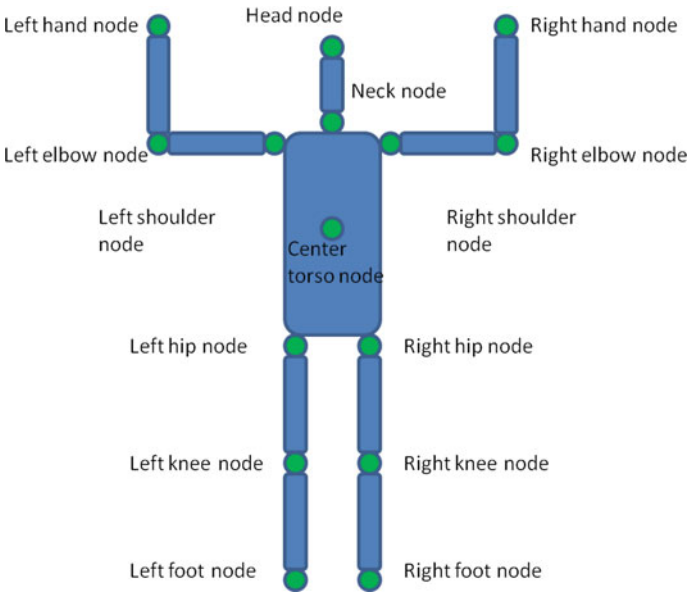


Fig. 4 *NITE's* skeletal model

3 Interaction with Music: Examples of Applications

So far, we have discussed the advantages and drawbacks of different tracking technologies and we have briefly skimmed over some of the possibilities that the *OpenNI/NITE* framework opens for the use of the *Kinect* camera in an application with stress in music interaction. However, we have not yet covered any specific example. This section provides examples of actual applications developed using a natural interaction interface based on *Kinect*. More specifically, this section will present two of such applications: first, an application where the user's hand movements induce pitch shifting in a sound source being reproduced, and secondly, an application that detects a drum-hitting-like gesture as a cue to actually play a drum-like sound. Finally, a collection of other possible paradigms for musical interactive applications is portrayed.

3.1 Pitch-Shifting by Motion Tracking

Pitch shifting is the process of changing the pitch of a given piece or sample without affecting its duration or speed. This process, along with time-stretching, is used, for instance, to match pitches and tempos of two sound excerpts that are to be mixed, to correct the intonation of instruments or singers, or to produce special effects such as increasing the range of a given instrument or adding a chorus-like

effect to a single singer performance (by mixing transposed copies of the original monophonic voice).

The most straight-forward way to implement pitch shifting would be to simply resample the clip to be processed, that is, to rebuild the original continuous waveform and sample it again at a different rate. However, this process also changes the duration of the signal. This means that if the signal were to be played at the original sampling frequency, the audio clip would sound faster or slower. In order to keep the duration unchanged, it would be necessary to time-scale the input signal by the same factor as it is going to be oppositely scaled by the resampling process. This stage can be implemented by using a time-stretching algorithm that does not introduce changes to the original pitch, such as the synchronized overlap-add method (Zölzer 2002). The order in which these two operations are applied can be interchanged (example in Fig. 5).

Another alternative is to use a phase vocoder to apply a pitch transposition transformation to a representation of the audio data based on a time–frequency model. One possible implementation of the phase vocoder multiplies the input audio signal by a sliding window which is nonzero for only a finite period of length N samples. This divides the input audio waveform in chunks or frames, which are transformed to the frequency domain by using the Fast Fourier Transform (FFT). This process is called Short-Time Fourier Transform (STFT), and it is mathematically represented by the equation Eq. 1.

$$X(n,k) = \sum_{m=-\infty}^{+\infty} x(m)w(m-n)e^{\frac{-j2\pi km}{N}} = |X(n,k)|e^{j\phi(n,k)} \quad (1)$$

Equation 1: Short-Time Fourier Transform for time–frequency representation.

In this equation, x(n) represents the digital audio samples, w(n) is the sliding window chosen, and X(*n, k*) represents the short-time spectra of each frame.

After the short-time spectrum has been calculated, it is possible to apply an operation or transformation by either manipulating the magnitude or the phase of each sample at each of the N frequency bins given by the FFT. Once the desired operation has been performed, the data is transformed back to the time-domain by using an Inverse FFT (IFFT), and then the processed output signal is conformed by adding and overlapping all the different frames previously calculated.

The summation of the IFFTs of all the *X*(*n, k*) allows the synthesis of the transformed signal back to the time domain. This implementation of the phase vocoder is refered to in (Zölzer 2002) as Direct FFT/IFFT approach. An alternative

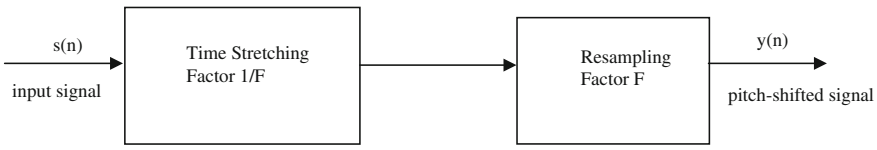


Fig. 5 Pitch shifting by resampling and time-stretching

way to implement phase vocoders is that of the filterbank or "Sum of Sinusoids" approach (Zölzer 2002).

When partitioning the signal into frames it is important to select an appropriate window. In particular, we have used a Hanning window in our implementation, and the windowing process is applied with an overlapping factor of 87.5 %, using a frame and FFT size of 1,024 samples. For each $X(n, k)$ calculated, real frequencies represented by each of the 1,024 bins are extracted (taking into account the phase $\varphi(n, k)$ and the delay introduced by the overlapping windowing process). Then, the frame is pitch-shifted by multiplying each frequency value by the corresponding pitch-shifting value. Afterwards, the IFFT is calculated using the same magnitude values as the original frame, but the phase values for each sample at each frequency bin are extracted from the transposed frequency values. Thus, the synthesis process yields an effectively pitch-shifted signal (an example can be found in Fig. 6).

After the selection and implementation of the pitch-shifting scheme, the next stage is the integration of such a process into a human motion tracking system to actually perform the pitch-shifting effect. To this end, we designed and programmed a simple virtual environment integrating the *OpenNI/NITE* API to get access to user motion data detected by a *Kinect* device in real time. The virtual environment and the application were programmed in C/C++ using a set of different libraries and frameworks: OpenAL for audio management, OGRE3D for graphics, OpenCV to manage the bitmaps extracted from the *Kinect* device, and *OpenNI/NITE* to implement natural interaction.

A screenshot of the virtual environment developed is portrayed in Fig. 7. Each of the fifteen nodes of the user's tracked skeleton was represented visually by spheres with a metallic texture. The information extracted from the depth image and the user segmented image was also used to create two small billboards to offer additional information to the user.

Users interact in the virtual world by using both their left and right hands. More specifically, during the start-up, the application loads a previously selected audio file. Left hand motion is used to start playing the audio or to pause the playback. This action was implemented by performing a simple "push-button-like" gesture to toggle the playing state of the song. Thus, playback switches between on and off whenever the user's left hand is moved forward so that the distance between the hand and the torso becomes larger than a certain amount over the Z axis, towards the camera.

While the music is playing, the user has the possibility of altering its pitch in real time by simply raising or lowering his right hand. In particular, if the right hand is kept approximately at the same height (Y axis) as the user's head height, then the song is played without any alterations. If the right hand is raised or lowered with respect to the reference height, then the pitch is shifted, increasing or decreasing respectively, according to the right hand's position. Pitch transposition is limited to a maximum of 12 semitones with respect to the original value. The highest pitch shift is applied when the user's right hand Y coordinate is approximately 50 cm larger than the head's one. Similarly, the lowest pitch shift possible

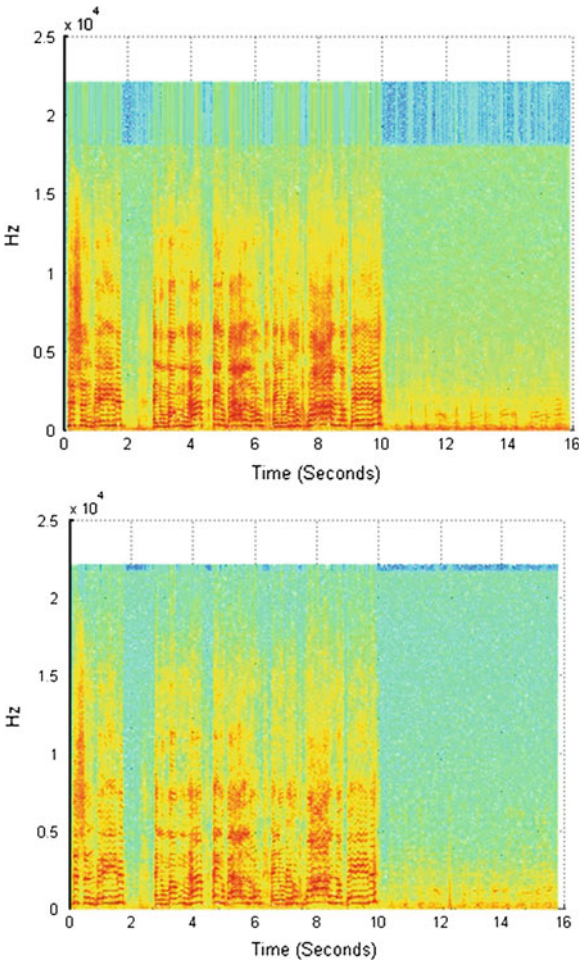


Fig. 6 Spectrogram of a vocal audio excerpt (*up*) and its pitch-shifted version (*down*). A 3 semitones shift is applied

is performed when the hand is about 50 cm smaller than the user's head height. In intermediate positions, the pitch shifting introduced is continuously distributed along the interval of semitones considered (−12, 12) proportionally to the relative height of the hand with respect to the head. Such behaviour is represented by Eq. 2, which portrays the factor F by which the audio excerpt's pitch is shifted (*relative Height* is measured in mm in this equation). Figure 8 illustrates an example of this procedure.

$$F = 2^{\frac{relative\ Height}{41{,}67 \times 12}} \tag{2}$$

Equation 2: Transcription of tracked movement into a pitch-shifting factor.

This application has been conceived from a pedagogical point of view. By correlating pitch rising and lowering with similar rising and lowering movements

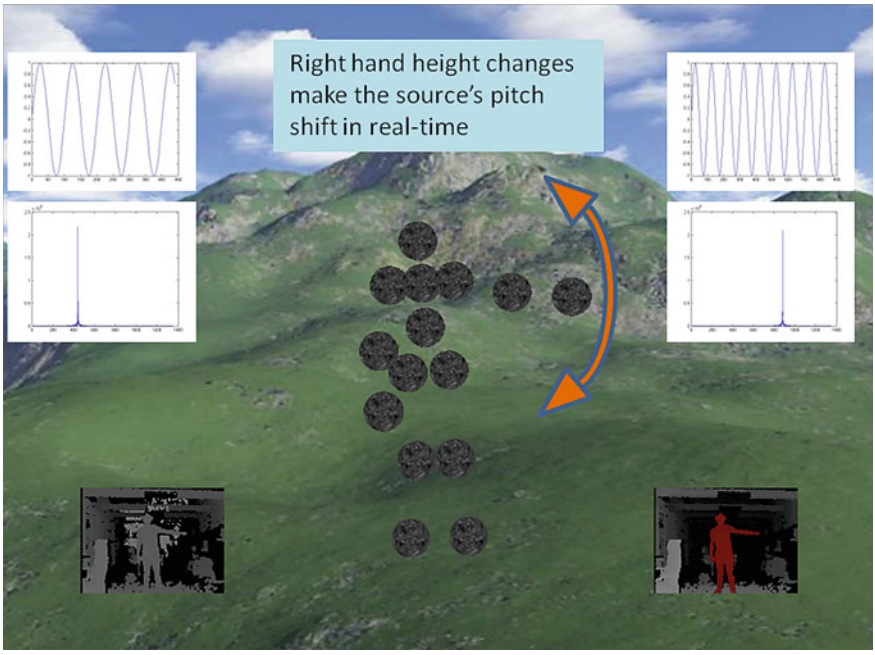


Fig. 7 Screenshot of the virtual environment developed for the pitch-shifting application

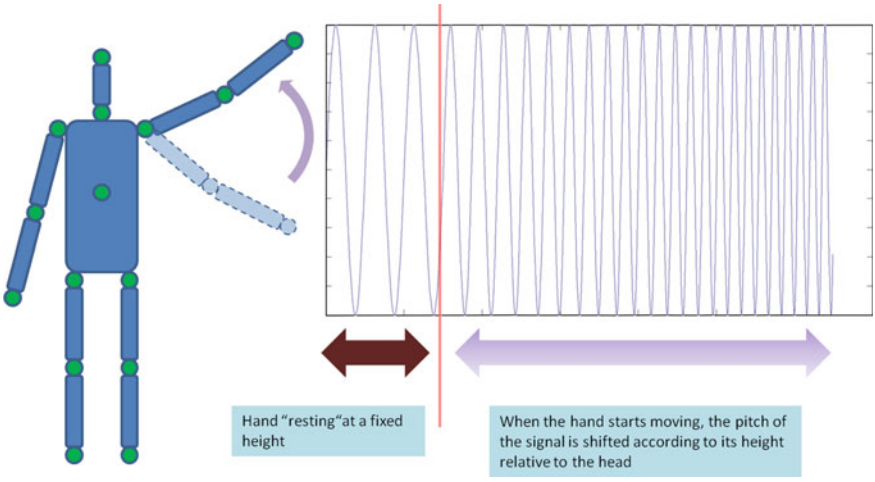


Fig. 8 Example of pitch-shifting of a single tone. When the hand moves from the blurred position to the end position, the pitch shift factor changes according to the height of the hand at each instant

of the hand, it becomes easier to understand the musical concept of pitch, especially for children.

The idea of coupling body motion with the generation of sound effects can be a useful tool for music learning, as previously indicated in the introduction, making the experience more enjoyable and explorative, and, at the same time, acting as an additional motivator towards the learning process.

3.2 Kinect-Battery: Playing Drums with Kinect

In the previously presented application, motion tracking was used in combination with music signal processing to achieve real-time pitch transposition. In this second application, we present a system that is capable of emulating a simple drum-like instrument by capturing user hand motion.

A first approach to the implementation of a simple air-drum would be to geometrically define a virtual drum in a given position in space. Since the *Kinect* 3D sensor actually provides coordinates for each position joint within its field of view, it could be possible to define the position and dimensions of a virtual drum within that space. Nevertheless, several usability drawbacks can be found in this approach. First of all, the most glaring problem when the virtual drum is located at a fixed position is that the drum would be "invisible" in the real world. Thus, unless the user is wearing a head-mounted display device, the user will need to keep watching an external display to find the drum. This can be easily overcome by requiring the user to stay at a particular position. Another issue is that the drum should be resized according to the height of the user, to ensure that children and adults alike can use it comfortably.

While these are definitely minor issues, they can pose an unnecessary inconvenience for user comfort. Therefore, an immediate refinement to the previous approach would be to virtually attach the virtual drum to the user; this can be achieved by placing the drum in a fixed relative position with respect to the user. This modification removes the previous issues and does not restrain user movements.

Figure 9 shows a screenshot of an application running this implementation. The framework of the application is the same one used in the pitch-shifting application. The virtual drum is represented by a simple geometric shape. Whenever the user moves either hand downwards and "hits" the upper top of the drum, a drum sound is played (see Fig. 10 for a schematic representation of this model). It is important to notice that the virtual drum is obviously intangible in the physical world. Therefore, the user might force the system to keep playing a sound by constantly waving the hand in short up-and-down moves near the top of the drum, creating a very unrealistic and unnatural sound output. To prevent this, both hands have to be raised a certain height before being capable of triggering a new sound when "hitting" the top surface of the virtual drum. The speed of the downwards movement is also taken into account to modify the intensity of the sound played

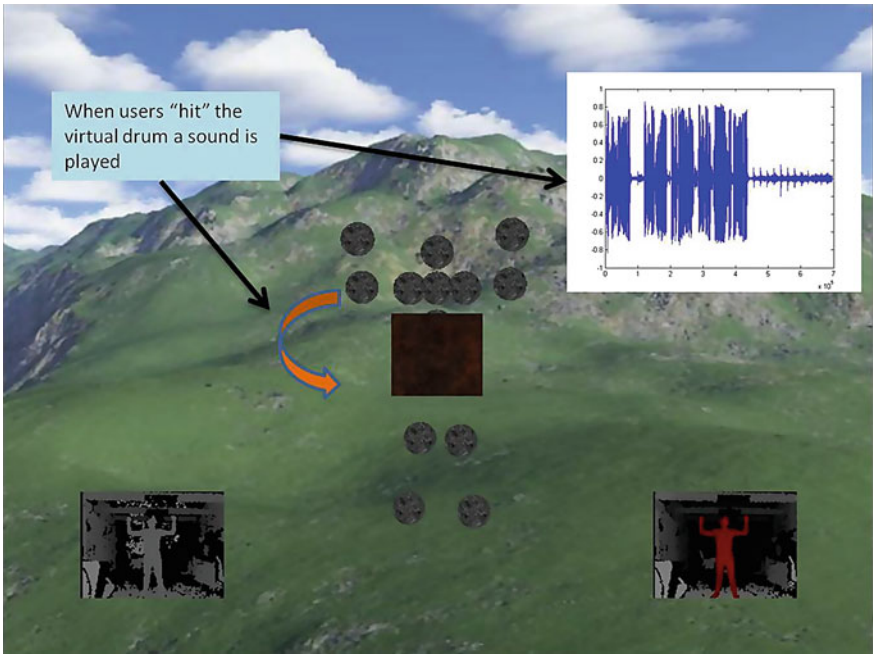


Fig. 9 Screenshot of the virtual environment for the drum simulator

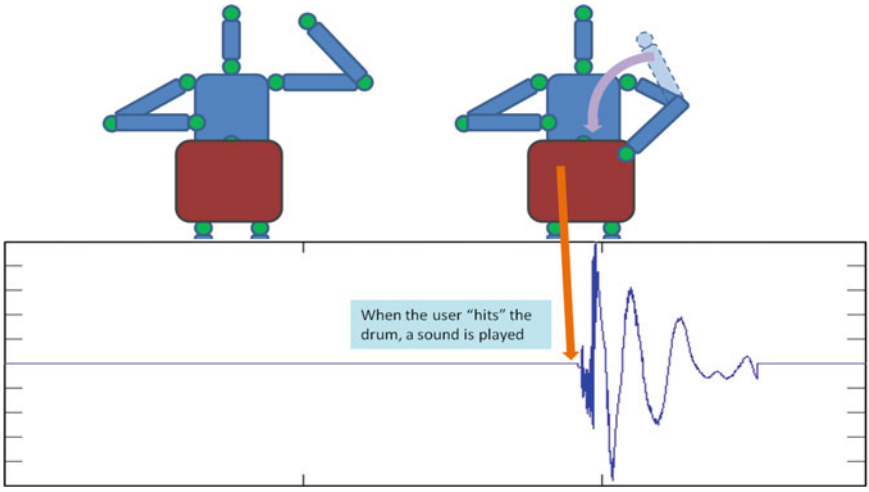


Fig. 10 Illustration of the "hit-drum-to-play" model

accordingly. A drum kit functionality can be easily implemented by creating several virtual drums distributed around the user.

A second approach to the implementation of the air-drum would rely strictly on user's motion to infer when the user "intends" to hit the drum. More specifically, the system would follow the movement of the hands, and whenever a drum-hitting gesture was detected, it would play the corresponding sound. In this second case, there would be no virtual drum, as the triggering of the sound would not depend on the exact position of user's hands but on the identification of drum-hitting gestures.

There are several reasons why this approach constitutes a more interesting approach. First of all, it removes the need for "invisible" drums. Furthermore, noisy samples of the hands' position can be more problematic when using the first approach, since it is more reliant on the precision of the user when "hitting" the drum.

The second reason why resorting to a drum-hitting recognition model can be a more enticing prospect is that the 3D sensing technology used to track user movement has some limitations in terms of accuracy, delay, etc. In particular, the delay between user movements and user's avatar transcription of the motion tracked could be as high as 200 ms. This is a hurdle that cannot be easily overcome with the former approach. Notwithstanding, it is feasible to look for features in the user tracked motion that can offer some insight on whether the user intends to perform a drum-hitting gesture or not. Thus, these features can be used to compensate for the delay introduced by the application. Moreover, by this way it is possible to predict when the user is going to perform a drum-hitting gesture according to a reference of previously performed gestures and the corresponding features extracted.

There are several ways in which a drum-hitting motion can be characterized. In the following lines, we present some possible features that can be used for the recognition of drum-hitting-like gestures in a system with meaningful delay. To illustrate these ideas, we will consider that the user performs drum-hitting gestures in a somewhat exaggerated way. Note that short moves are very hard to track with the 3D sensing technology considered.

- **Descending Acceleration**. A feature that is clearly linked to a downwards drum-hitting gesture is that of velocity and acceleration. More specifically, a negative acceleration peak along the Y axis will be expected to be found. At the start of the downwards movement, the acceleration over the Y axis should have a negative value; at the time when the drum would be actually hit, velocity should ideally become zero for that instant (so acceleration should have become positive instants before). Finally, since it would not be natural to end the movement exactly at the drum-hitting point, it stands to assume that in a proper drum-hitting gesture, the hand would ascend again after performing a beat. Thus, at the "end" of the gesture, the acceleration over the Y axis should become positive. This is a simple feature that can be used (along with a certain threshold and some constraints to the X, Y and Z coordinates to avoid false positives) to discern whether the user actually performs a drum-hitting motion or not. Furthermore, it is feasible to use this very same feature to predict whether the user intends to perform such a gesture: if the user makes a downwards movement with a sufficiently large value of speed/acceleration, it is very likely

that a drum-hitting gesture is to be performed. Thus, by setting a certain threshold on the measured acceleration/velocity of the user hand, it is possible to predict when the user intends to "play the drums". Therefore, the system can give a response before the actual motion is fully detected, helping to mitigate the effects of the delay of the tracking device (see Fig. 11).

- **Linear Prediction of Position**. Another way to address the problem of the delay is to use a linear predictor to estimate the evolution of user's motion according to the behaviour observed in the samples already captured. This would allow the system to actually predict the position of the user's hand and identify potential drum-hitting gestures prior to the actual completion of such gestures. Such identification would be accomplished, for example, by fitting the samples to a function that characterizes the gesture with sufficient accuracy. In this case, it is also important to account for the fact that different users might tend to perform their gestures at different velocities, so the information in the database should cover a wide enough assortment of gestures performed at different speeds or be properly parameterized to prevent this.

These features can be used by themselves to detect whether a drum-hitting gesture has been performed or not. Despite this, it is more likely to get better results in the gesture-recognition task when combining the outputs given by these features. This can be achieved by training a Bayesian classifier or by resorting to other machine learning methods.

It is important to notice that the data sampled should be appropriately normalized in order to minimize the dependence of the motion data tracked on the actual size of each user. Even for users similarly sized, the proximity to the 3D sensing device will also affect the actual range of values that will be tracked.

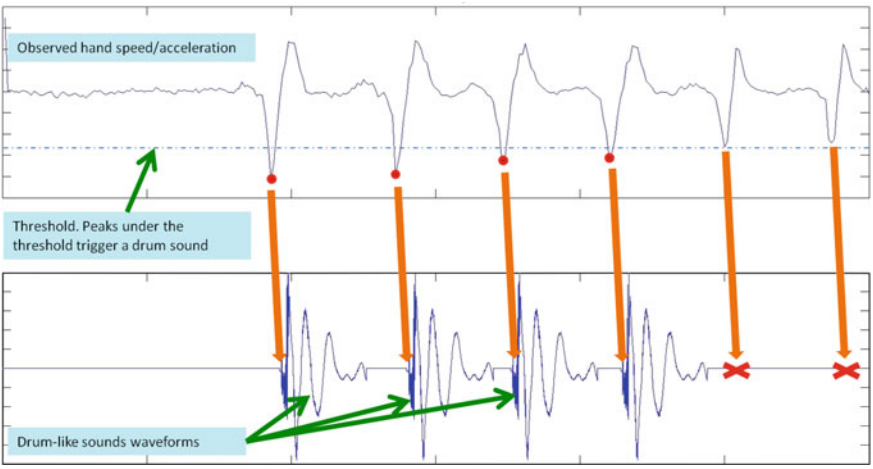


Fig. 11 Example of gesture detection using descending velocity (or acceleration): if there is a peak of downward speed faster than a given threshold, a drumbeat is played

One simple way to overcome this limitation is to use the torso-to-head distance as a reference to normalize the data tracked. This simple reference removes the dependence of the data sampled on the proximity to the sensor and mitigates the effects of user size variability in the size of potential users.

Another issue, especially with regard to avoiding false positive detection, is that certain constraints should be applied to gesture detection features for those cases in which it is obvious that no significative gestures can be performed. Examples of such constraints can refer to the mean velocity of the movement or the coordinates of the location of the hand at the beginning of the motion.

One final aspect to consider is that of a drum-kit simulation. We previously stated that, with the first approach discussed (virtual battery simulation in 3D space), it would be difficult to implement a drum-kit due to the precision necessary to "hit" the space confined by the invisible virtual drums. In this second approach, a problem is the delay introduced by the system in the detection of a gesture. Since the gesture recognition measures proposed do not depend on the actual position on the XZ plane, a simple way to discriminate between different drums would be to use the coordinates on the X and Z axes whenever a gesture is detected to select the corresponding drum sound to play. While this partially brings back some issues regarding precision when "hitting" an invisible 3D space, this solution is by far more robust in this case. First of all, discriminating between positions in the XZ plane is far simpler than detecting when the user "hits" a specific 3D volume. Secondly, and more important, since a sound is played when a gesture has been detected, as long as a gesture is correctly performed, a sound will play. Precision errors might result in having the wrong drum sound played, but the system would respond. On the contrary, with the first implementation, "missing the drum" implies that no sound would be played, which would be more frustrating to the user. Furthermore, with the second approach, if the user wants to virtually play drums, he/she will need to perform gestures that are reasonably similar to the actual gestures a musician would make when playing the drums.

Like the pitch-shifting application presented in the previous section, this application can also be seen from a pedagogical point of view. In particular, it can be used to teach rhythmic patterns in basic music stages. But perhaps the most enticing application of such a system would be as an accessible way for novice users to musical expression and performance. Albeit limited, the proposed application helps to lower barriers for lay people in musical terms, and allows for a reasonable simulation of a drum-like instrument. Also, it removes most of the hurdles related to instrument maintenance and overuse, as well as offering an explorative and intuitive gateway to the world of musical performance.

3.3 Other Interaction Paradigms

So far we have presented in detail some aspects regarding specific applications for real-time integration of human motion and musical responses. However, there are

many other enticing and appealing applications and interaction paradigms that allow for the integration of motion cues and music signal processing. In this section, some examples of interaction paradigms that can be used to achieve new forms of musical expression will be briefly presented.

3.3.1 The Virtual Director: An Example of Tempo Modification with Hand Motion

The title of music director is normally used in many symphony orchestras to designate the principal conductor and artistic leader of the orchestra. The role of the director is to oversee the overall musical performance, supervising and guiding the musical performance of the orchestra. The director guides the musicians with the motion of his hands and arms. Thus, a simple approximation to a virtual director application can be easily achieved with a body motion tracking system, as shown in Fig. 12.

More concretely, it is relatively easy to use skeletal joint data to discriminate over a given range of tempo. For example, to discriminate among *andante*, *adagio* and *presto*, the space around the user could be divided into three regions according to height: a lower region (for example, from slightly over the waist and downwards), an upper region (from the shoulders upwards), and a middle region (approximately between shoulders and waist). Therefore, if the user moves his/her hands in the lower region, tempo would be set to *adagio*, *andante* in the middle region, and *presto* for the upper region. Thus, by detecting that the user is performing a waving motion similar to the archetypical waving motion used by music directors, it is feasible to implement a basic version of a music director simulator. More complex implementations could be achieved by extending the range of tempo considered, or by analyzing the waving speed instead of or in addition to the height of the hands, as shown in Fig. 13.

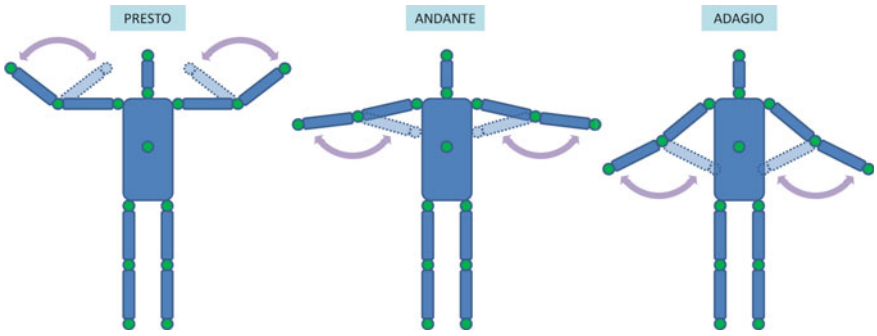


Fig. 12 Schematic representation for a simple implementation of a virtual director by means of a height-based system for tempo selection

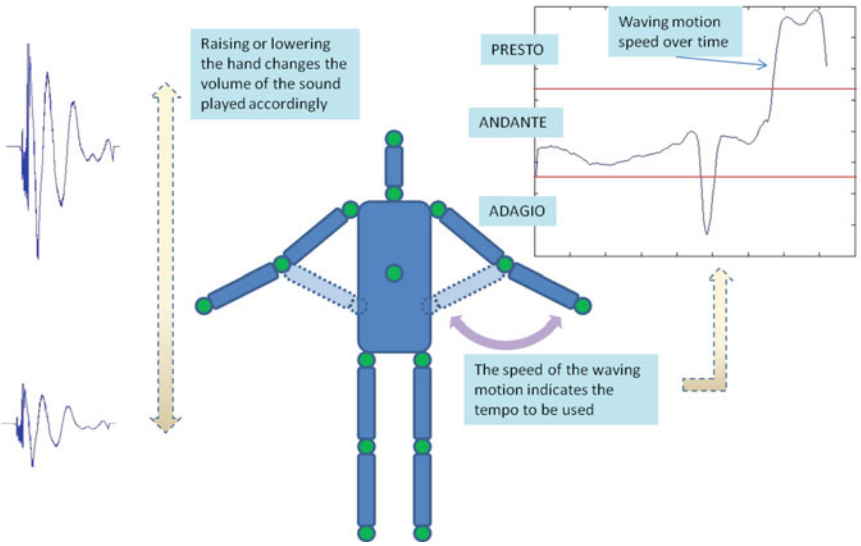


Fig. 13 An alternative interaction paradigm for a virtual director

3.3.2 Instrument Simulation

Previously, we have discussed the implementation of a simple drum-like instrument, but a similar approach can be followed for the implementation of other types of instruments. The xylophone and other percussion instruments could be easily simulated using a system similar to the one previously proposed. Obviously, the skeletal model obtained from *Kinect's* depth map would not be useful to implement an "air-piano". However, by processing the depth and RGB images simultaneously, it would be possible to find the position of each finger to implement such application. Of course, the computational cost would be larger than in the previous examples. Also, in this case a visual reference in the real world would be necessary since users cannot be expected to correctly hit the keys of an invisible piano accurately.

3.3.3 Advanced Pitch-Shifting/Correction

An evolution of the previously presented pitch-shifting application could combine the idea of pitch shifting by motion with an onset detector. In particular, this application could follow the next scheme: the user records a performance of a musical piece. Then, this sample is processed and segmented into the notes actually played. Afterwards, the application shows the timeline of the song divided by the note-segments detected (for example, using a pentagram, as shown in Fig. 14). Then, the user could modify the pitch of each identified note-segment

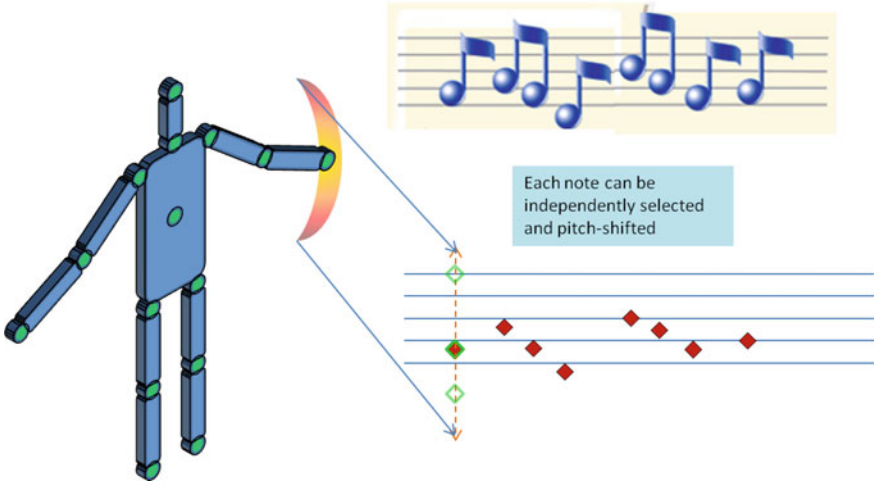


Fig. 14 Portrayal of an advanced pitch-shifter

along the pentagram, for example using the left hand to select the notes to be pitch-shifted, and the right hand to raise or lower the pitch. Again, this can be a useful application for pedagogical purposes, especially in teaching the basics of music theory.

3.3.4 Real-Time Motion-Based Composition

Motion-based interfaces can provide the means for real-time expressive rendering of musical pieces, lowering the barriers of music theory for a naïve user. This can be easily achieved, for example by allowing the user to create music by selecting notes and placing them on a score (in this case, the motion interface becomes mainly an additional compelling element). Also, the previously commented pitch-shifting application can be considered a real-time composition-by-motion scheme. A more interesting system could be based on combining motion detection with an automatic score generator.

4 Conclusions

In this chapter, a brief overview of the capabilities of advanced human computer interaction for the design and development of musical interactive applications has been drawn. More specifically this chapter has focused on the combination of human motion with music signal processing techniques in order to achieve an assortment of effects or applications. After summarizing the main advantages and

disadvantages of the different technological alternatives for the implementation of human motion-tracking systems, we have turned our attention to the Microsoft *Kinect 3D* sensor given its availability, low-cost, non-intrusiveness and high performance features.

We have shown details of two example applications that make use of a 3D sensor depth camera to implement a real-time motion-based application for pitch shifting and a drum simulator. However, as evidenced by the bibliography and the alternative examples shown in the previous sections, the interaction possibilities are nearly endless.

In view of the work developed, the main advantage of the applications created or described is their high level of accessibility, since no prior musical knowledge is required for the user to actually interact with the sounds played. As previously stated, this makes for an innovative and explorative way to better understand the concepts behind music, sound, or sound effects. Taking into account this observation, it is easy to get to the conclusion that this kind of applications can be potentially useful for music learning purposes, diminishing some of the hurdles concerning music theory by means of a more "tangible" visualization of theoretical concepts.

Furthermore it is expected that this type of added interactivity will also make the experience more compelling and fulfilling. This should encourage the application of advanced technologies and devices for music related tasks, including entertainment and learning.

Acknowledgments This work was supported by the Ministerio de Economía y Competitividad of the Spanish Government under Project TIN2010-21089-C03-02 and Project IPT-2011-0885-430000 and by the Ministerio de Industria, Energía y Turismo of the Spanish Government under project TSI-090100-2011-25.

References

Antle, A. N., Droumeva, M., & Corness, G. (2008). Playing with the sound maker: Do embodied metaphors help children learn? *Proceedings of the 7th International Conference on Interaction Design and Children (IDC '08)* (pp. 178–185).

AudioSurf. http://www.audio-surf.com/, release Feb 2008. Last accessed 14 June 2012.

Bakker, S., van den Hoven, E., & Antle A. N. (2011). MoSo tangibles: Evaluating embodied learning. *Proceedings of the fifth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '11)* (pp. 85–92).

Baratoff, G., & Blanksteen, S. (1993). Tracking devices. Technical Report, Human Interface Technology Laboratory. http://www.hitl.washington.edu/scivw/EVE/I.D.1.b.TrackingDevices.html. Last accessed 15 May 2012.

Barbancho, A. M., Barbancho, I., Tardon, L. J., & Urdiales C. (2009). Automatic edition of songs for guitar hero/frets on fire. *IEEE International Conference on Multimedia and Expo (ICME 2009)* (pp. 1186–1189). June–July 2009, New York.

Beatles Rock Band. The. http://www.thebeatlesrockband.com/, release Sept 2009. Last accessed 14 June 2012.

Castellano, G., Bresin, R., CamurriI, A., & Volpe, G. (2007). Expressive control of music and visual media by full-body movement. *Proceedings of the 7th International Conference on New Interfaces for Musical Expression (NIME'07)* (pp. 390–391).

Craft, K. (2009). Guitar hero III passes $1b. http://www.edge-online.com/news/activision-guitar-hero-iii-passes-1b. Last accessed 14 June 2012.

De Dreu, M. J., van der Wilk, A. S. D., Poppe, E., Kwakkel, G., & van Wegen, E. E. H. (2012). Rehabilitation, exercise therapy and music in patients with Parkinson's disease: A meta-analysis of the effects of music-based movement therapy on walking ability, balance and quality of life. *Parkinsonism and Related Disorders, 18*(1), 114–119.

FretsOnFire. http://fretsonfire.sourceforge.net/, debut release Aug 2006. Last accessed 14 June 2012.

Gleicher, M., & Ferrier, N. (2002). Evaluating video-based motion capture. *Proceedings of Computer Animation 2002 (CA'02)* (pp. 75–80).

Gower, L., & McDowall, J. (2012). Interactive music video games and children's musical development. *British Journal of Music Education, 29*, 91–105.

Grollmisch, S., Dittmar, C., & Gatzsche, G. (2009). Concept, implementation and evaluation of an improvisation based music video game. *IEEE Consumer Electronics Society's Games Innovation Conference* (pp. 210–212). London.

Guitar Hero. http://hub.guitarhero.com/, last release Feb 2011. Last accessed 14 June 2012.

Holland, S., Bouwer, A., Dalgleish, M., & Hurtig, T. (2010). Feeling the beat where it counts: Fostering multi-limb rhythm skills with the haptic drum kit. *Proceedings of the Fourth International Conference on Tangible, Embedded, and Embodied Interaction (TEI'10)* (pp. 21–28).

Jorda, S. (2010). The reactable: Tangible and tabletop music performance. *Proceedings of the 28th of the International Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA'10)* (pp. 2989–2994).

Karimi, R., Nanopoulos, A., & Schmidt-Thieme, L. (2012). RFID-enhanced museum for interactive experience. *Multimedia for Cultural Heritage, Communications in Computer and Information Science, 247*(3), 192–205.

Khoo, E. T., Merritt, T., Fei, V. L., Liu, W., Rahaman, H., Prasad, J., & Marsh, T. (2008). Body music: Physical exploration of music theory. *Proceedings of the 2008 ACM SIGGRAPH Symposium on Video Games (Sandbox '08)* (pp. 35–42).

Levin, G., & Lieberman, Z. (2004). In-situ speech visualization in real-time interactive installation and performance. *Proceedings of the 3rd International Symposium on Non-Photorealistic Animation and Rendering* (pp. 7–14).

Mancini, M., Bresin, R., & Pelachaud, C. (2007). A virtual head driven by music expressivity. *IEEE Transactions on Audio, Speech and Language Processing, 15*(6), 1833–1841.

Migneco, R., Doll, T., Scott, J., Hahn, H., Diefenbach, P., & Kim, Y. (2009). An audio processing library for game development in flash. *IEEE Consumer Electronics Society's Games Innovation Conference* (pp. 201–209). London.

Mikestar. http://www.mikestar.com/de/skore/start, debut release Oct 2009. Last accessed 14 June 2012.

Perry, L. D. S., Smith, C. M., & Yang, S. (1997). An investigation of current virtual reality interfaces. *ACM Crossroads Student Magazine, 3*(3), 23–28.

PrimeSense (2011a). OpenNI: User guide. Available: http://www.openni.org/Documentation.aspx. Last accessed 15 May 2012.

PrimeSense (2011b). Prime sensor NITE 1.3 algorithms notes. Available: pr.cs.cornell.edu/humanactivities/data/NITE.pdf. Last accessed 15 May 2012.

Rock Band. http://www.rockband.com/, debut Wii version release Sept 2008. Last accessed 14 June 2012.

Shivappa, S. T., Trivedi, M. M., & Rao, B. D. (2010). Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE, 98*(10), 1692–1715.

SingStar. http://www.singstargame.com/, debut release May 2004. Last accessed 14 June 2012.

Smisek, J. (2011). 3D with kinect. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 1154–1160).

Taylor, R., Torres, D., & Boulanger, P. (2005). Using music to interact with a virtual character. *International Conference on New Interfaces for Musical Expression* (pp. 220–223).

UltraStar. http://www.ultrastar.pl/about, release Oct 2010. Last accessed 14 June 2012.

Villaroman, N., Rowe, D., & Swan, B. (2011). Teaching natural user interaction using OpenNI and the Microsoft Kinect sensor. *Proceedings of the 2011 Conference on Information Technology Education (SIGITE '11)*. ACM (pp. 227–232).

Wang, C. Y., & Lai, A. F. (2011). Development of a mobile rhythm learning system based on digital game-based learning companion. *Lecture Notes in Computer Science, 6872*, 92–100.

Welch, G., & Foxlin, E. (2002). Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Computer Graphics and Applications, 22*(6), 24–38.

Zölzer, U. (Ed.). (2002). *DAFX digital audio effects*. Wiley: New York.