Articulatory Synthesis of Speech and Singing: State of the Art and Suggestions for Future Research

Bernd J. Kröger and Peter Birkholz

Department of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and Aachen University, Aachen, Germany
bkroeger@ukaachen.de, peterbirkholz@gmx.de

Abstract. Articulatory synthesis of speech and singing aims for modeling the production process of speech and singing as human-like or natural as possible. The state of the art is described for all modules of articulatory synthesis systems, i.e. vocal tract models, acoustic models, glottis models, noise source models, and control models generating articulator movements and phonatory control information. While a lot of knowledge is available for the production and for the high quality acoustic realization of *static* spoken and sung sounds it is suggested to improve the quality of control models especially for the generation of articulatory *movements*. Thus the main problem which should be addressed for improving articulatory synthesis over the next years is the development of high quality control concepts. It is suggested to use action based control concepts and to gather control knowledge by imitating natural speech acquisition and singing acquisition scenarios. It is emphasized that teacher-learner interaction and production, perception, and comprehension of auditory as well as of visual and somatosensory information (multimodal information) should be included in the acquisition (i.e. training or learning) procedures.

Keywords: speech, singing, articulatory synthesis, articulation, vocal tract acoustics, movement control, human-human interaction, speech acquisition.

1 Introduction

Articulatory synthesis systems comprise (i) a module for the generation of vocal tract movements (control model), (ii) a module for converting this movement information into a continuous succession of vocal tract geometries (vocal tract model), and (iii) a module for the generation of acoustic signals on the basis of this articulatory information (acoustic model). It is an advantage that these systems are closely related to the natural human process of the production of speech or singing. But due to the complexity of the natural processes of speech or singing production, no current articulatory synthesis system is capable of generating high quality acoustic speech or singing signals as are generated for example by current corpus based unit selection synthesis methods.

In this paper the state of the art knowledge for articulatory synthesis of speech and singing is summarized and suggestions are made for developing a high quality system on the basis of this currently available knowledge for all levels of these systems, i.e. on the level of the control model, of the vocal tract model and of the acoustic model.

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 306–319, 2009.

2 Vocal Tract Model and Acoustic Model

2.1 Vocal Tract Models

The task of vocal tract models is to generate the complete geometrical information concerning the vocal tract (shape and position of all vocal tract organs, i.e. lips, tongue, palate, velum, pharynx, larynx, nasal cavity) and its variation over time. Shape, position, and motion of movable vocal tract organs are generated on the basis of the time functions of all vocal tract parameters defined by the model. A typical set of vocal tract parameters are: position of jaw, upper lips, lower lips, tongue tip, tongue body, velum, and larynx. Vocal tract models can be subdivided into statistical, biomechanical, and geometrical models.

Statistical models (e.g. Maeda 1988, Beautemps et al. 2001, Badin et al. 2002, Serrurier and Badin 2008) are based on large corpora of vocal tract movements measured by different techniques (MRI, EMA, or X-Ray). Articulatory information (local flesh point positions reflecting the position of a vocal tract organ or the whole shape of vocal tract organs) is extracted for each time instant (frame by frame) for the whole corpus and articulatory movement parameters are derived by statistical procedures (e.g. principal component analysis).

Biomechanical models aim to model the physiological basis of all vocal tract organs and their neuromuscular control (e.g. Wilhelms-Tricarico 1995, Dang 2004). These models mainly use finite element methods and are based on physiological knowledge about the muscular structure, tissue structure and cartilaginous or bone structure of vocal tract organs. The articulatory parameterization as well as the shaping and positioning of the vocal tract organs result from physiological knowledge.

For *geometrical models* (e.g. Mermelstein 1973, Kröger 1998, Engwall 2003, Birkholz et al. 2006) the positioning and shape of the vocal tract organs is calculated by using a set of a priori defined vocal tract parameters. The basis for the choice of distinct parameter sets depends mainly on assumptions about the variety of vocal tract configurations which the model aims to approximate. Due to the flexibility of this class of models the vocal tract shape can be fitted to articulatory data of different speakers, i.e. the model can be adapted to different speakers with different sex and age (e.g. Birkholz and Kröger 2006). Fitting the model to articulatroy data is basically done using static vocal configurations (e.g. vocalic and consonantal MRI-data, see Engwall 2003, Birkholz and Kröger 2006) but in addition, movement data at least of certain flesh points of articulators or of certain contact regions (e.g. EMA-data and EPG-data, see Engwall 2003) can be fitted in order do make sure that the model behaves correctly even in the case of articulatory movements.

2.2 Acoustic Models

The task of the acoustic models is to calculate the time varying air flow and air pressure distribution within the vocal tract and to calculate the acoustic speech signal radiated from the facial region of the model. The input information for acoustic models is lung pressure, subglottal air flow, and the geometric shape of the vocal tract tube (trachea, glottis, pharyngeal, oral, and nasal tract) for each time instant. A time-varying *tube model* is specified from the geometrical vocal tract model information, which represents the vocal tract cavities (trachea, pharynx, nasal, and oral cavity). The tube model consists of a succession of *tube sections* with constant cross sectional area. The aerodynamic and acoustic characteristics of these tube sections and their joints towards the neighboring tube sections are described by the acoustic models. Acoustic models can be subdivided into reflection type line analog models, transmission line circuit analog models, hybrid time-frequency domain models, and finite element wave propagation models.

In the case of *reflection type line analog models* (e.g. Kelly and Lochbaum 1962, Liljencrants 1985, Meyer et al. 1989, Kröger 1998), forward and backward traveling partial flow or pressure waves are calculated for each vocal tract tube section in the time domain on the basis of scattering equations which reflect the impedance discontinuity at tube junctions. The calculation of pressure and flow within each tube section from trachea to mouth and nostrils and the calculation of the radiated sound wave are accomplished from the forward and backward traveling flow or pressure waves. The major shortcoming of this simulation technique is that variations of vocal tract length over time – which occur in normal speech production, e.g. within an [u]-to-[i] or [a]-to-[u] transition – can not be handled.

In the case of *transmission line circuit analog models* (e.g. Flanagan 1975, Maeda 1982, Birkholz et al. 2007), pressure and flow within each vocal tract tube section is calculated by a digital simulation of electrical circuit elements, representing the acoustic and aerodynamic properties within each vocal tract tube section. Variations of vocal tract length can be handled but, as in the case of all time domain simulation techniques, frequency dependent acoustic and aerodynamic losses (from sound radiation at nostrils and mouth, from vocal tract wall vibrations, from air frication at vocal tract walls etc.) can just be approximated. But the modeling of these loss mechanisms is essential for high quality speech synthesis since loss mechanisms for example adjust the bandwidth of formants and thus also the overall signal amplitude in different frequency regions. A very detailed discussion of acoustic and aerodynamic loss mechanisms within the vocal tract and a very detailed and gainful discussion of strategies how to approximate these loss mechanisms in time domain models is given by Liljencrants (1985). On the basis of that work it is now possible to take transmission line circuit analog models as a basic approach for high-quality articulatory synthesis.

In the case of *hybrid time-frequency domain models* (e.g. Allen and Strong 1985, Sondhi and Schroeter 1987), the frequency dependence of the acoustic simulation is modeled very detailed by calculating the acoustic transfer function for each vocal tract tube section within the frequency domain. In order to calculate flow and pressure at least at the most important locations within the vocal tract tube, time-frequency domain transformations must be calculated at these locations and for each time instant. Hence the implementation of this approach is complex and the calculation of the acoustic signal is time consuming.

In the case of *finite element wave propagation models* (e.g. El-Masri et al. 1996, Mazsuzaki and Motoki 2000) the aerodynamic and acoustic behavior of air flow and air pressure within the vocal tract is calculated very precisely by directly solving basic physical equations for the aero-acoustics of wave propagation. These models are very complex and the calculation of the acoustic signal is time consuming. This kind of models is rather appropriate for addressing general problems of transmission line wave propagation (for example for addressing the problem of noise generation within the tube) than for calculating acoustic signals in real time applications.

2.3 Glottis Models

The task of glottis models is to generate the acoustic source signal for phonation and its insertion into the vocal tract tube model. The source signal is propagated through the supraglottal cavities (pharyngeal, oral and nasal cavity) as well as through the subglottal cavities (trachea, lungs) by the acoustic model. Glottis models can be subdivided into self-oscillating models, parametric glottal area models, and parametric glottal flow models.

In the case of *self-oscillating glottis models* (Ishizaka and Flanagan 1972, Cranen and Boves 1987, Story and Titze 1995, Kröger 1998, Alipour et al. 2000, Kob 2002) the oscillation behaviour of vocal folds is calculated on the basis of general physical equations of motion leading to the waveform of the glottal area over time. Subsequently the glottal flow waveform can be calculated as a function of time on the basis of the glottal area waveform. The dynamic oscillation behaviour is generated for coupled spring mass systems representing the vocal fold dynamic behaviour and is controlled by biomechanical parameters like vocal fold tension and glottal aperture. External forces acting on the spring mass system result form the pressure distribution along the glottal constriction. Note that the term “articulatory” as is used in this paper also covers “phonatory articulation”, which comprises the adjustment of vocal fold tension and glottal aperture. Self-oscillating glottis models allow to control different speech sound qualities like qualities along the scales soft-normal-loud, pressed-modal-breathy, as well as for creaky voice and glottalizations (Kröger 1997a). Furthermore self-oscillating gottis models allow to approximate different registers in singing like chest and falsetto (Kröger 1997b) by an adjustment of the appropriate control parameters. In addition these models allow a wide variation of fundamental frequency by changing the control parameter vocal fold tension and by changing in addition other physiological parameters of the vocal folds (like overall length, mass distribution for the vocal folds, etc.) which has to be preset with respect to sex and age of a speaker or singer.

In the case of *parametric glottal area models* (e.g. Titze 1989, Cranen and Schroeter 1996), the time function of the glottal area waveform is predefined while glottal flow and glottal pressure results from the insertion of the glottal area model into the acoustic-aerodynamic model. An advantage of this kind of models is that fundamental frequency is a direct control parameter, while this is not the case in self-oscillating glottis models. But it should be noted that these models need a detailed description of the glottal area waveform for the control of different sound qualities like e.g. breathy, normal, pressed. In the case of self-oscillating glottis models the glottal area waveform directly results from underlying articulatory settings like more or less glottal ab-/adduction.

In the case of *parametric glottal flow models* (e.g. Fant et al. 1985) the time function of glottal flow is directly parameterized and inserted into the acoustic model. But this may prohibit the simulation of important acoustic effects like the modification of the glottal flow waveform resulting from acoustic interactions with the supralaryngeal vocal tract. Furthermore, as was mentioned in the case of parametric glottal area models, the control of voice quality in the case of these models can only be done successfully if a detailed knowledge of the flow waveform parameters is available (e.g. from inverse filtering procedures, Fant 1993).

2.4 Noise Source Models

The task of noise source models is to generate and to insert noise source signals into the acoustic transmission line model. Noise signals result from turbulent air flow, mainly occurring downstream in front of a vocal tract constriction in the case of a high value of volume flow. Noise source models can be subdivided into parametric and generic noise source models.

In the case of *parametric noise source models* (e.g. Mawass et al. 2000, Birkholz et al. 2007), turbulent noise is inserted into the vocal tract tube if a defined aerodynamic situation occurs: sufficient narrow constriction and sufficient high air flow. High quality acoustic signals can be generated in all cases of noise generation within the vocal tract (i.e. for glottal aspiration, for plosive noise bursts, and for fricative noise) by optimizing the frequency contour of the inserted noise, by optimizing the parameters controlling the noise amplitude, and by optimizing the location for insertion of the noise source within the acoustic transmission line model with respect to the location of the constriction.

In the case of *generic noise source models* (e.g. Sinder 1999) the occurrence of turbulences and its acoustic consequences are calculated by solving basic physical aero-acoustic equations. But these procedures are very time consuming and thus these kind of models are rather used for the elucidation of the basic physical aero-acoustic principles of noise source generation than for generating noise in an articulatory synthesizer for speech and singing aiming for real time applications.

3 Control Models for Speech and Singing

It is the task of control models to generate the complete vocal tract control information for each time instant of signal production. The control model specifies the whole set of vocal tract control parameters defined by the appropriate vocal tract model for each time instant during the production of an utterance or song. Thus the control model generates the complete set of movements for all vocal tract organs (articulatory movements) during the whole time course of the utterance or song. Control models can be subdivided into segmental and action based control models for speech. A first control model for singing has been established using an action based concept.

The input specification for *segmental control models* (Kröger 1992 and 1998) is a phonological description of the utterance. A succession of time intervals is specified from this input information for the production of each syllable and subsequently for the production of each sound. Time instants (labels) are specified within these sound production time intervals at which specific spatial targets (spatial goals) have to be reached by specific vocal tract organs. Coarticulation results form articulatory underspecification (Kröger 1992, 1998, and 2003) since targets are specified exclusively for those vocal tract organs which are involved in an obligatory way in the production of the intended sound; e.g. the lips in the case of a labial sound ([p], [b], [f], [v], [m], …), the tongue tip in the case of an apical sound ([t], [d], [s], [z], [n], …), or the tongue body in the case of a dorsal sound ([k], [g], [x], [ŋ], …). Lips and tongue body position have to be specified in the case of all vowels since each vowel requires a specific formation of the whole vocal tract tube. The position of the velum has to be specified for all sounds with respect to the fact whether they are nasals or nasalized vowels (low position of the velum), oral sonorants (high position of the velum), or obstruents (high position of the velum and the velopharyngeal port is tightly closed). For obstruents (e.g. plosives and fricatives) it is necessary to ensure an air pressure build up within the pharyngeal and oral cavity of the vocal tract. The rest position of the vocal folds (vocal fold aperture) has to be specified for each sound with respect to the fact whether the sound is voiced (vocal folds abducted), voiceless (vocal folds adducted), or a glottal stop (vocal folds tightly adducted).

The input specification of *action based control models* for speech (Saltzman and Munhall 1989, Browman and Goldstein 1992, Kröger 1993, Saltzman and Byrd 2000, Goldstein et al. 2006, Birkholz et al. 2006, Kröger and Birkholz 2007; also called gestural control models) is the phonological description of the utterance including a prosodic annotation. This information is converted into a *discrete score of vocal tract actions.* The vocal tract actions (or gestures) are associated with each other in a specific way (Fig 1, top). Subsequently this action score is formulated in a quantitative way as *high level phonetic score of vocal tract actions* which can be interpreted from a cognitive and sensorimotor viewpoint as a *high level representation of the motor plan* of the utterance (cp. Kröger et al. 2008). Starting with the qualitative discrete description (Fig. 1 top: discrete action score) a first quantitative version of the vocal tract action score or motor plan of a syllable or word is calculated by the control model (Fig. 1 bottom). Here, the temporal duration and time course of degree of realization for each vocal tract action (including onset and offset intervals) and a specification of the vocal tract action goal in the high level somatosensory domain (not displayed in Fig. 1) is specified. Subsequently the motor plan can be executed which leads to concrete movement trajectories for all articulators (i.e. a *lower level phonetic articulatory description* which is not displayed in Fig. 1). This lower level phonetic articulatory description is obtained by using concepts of inverse dynamics, if more than one articulator is involved in the execution of a vocal tract action (e.g. lower jaw, upper and lower lips for a labial closing action, cp. Saltzman and Munhall 1989).

The advantage of action based control models in comparison to segmental control models is that the dynamics and kinematics of articulatory *movements* can be controlled explicitly. This is a very important feature since sensorimotor movement control is an essential task in all kinds of human behavior or movement generation. Furthermore coarticulation directly results from temporal overlap of actions (i.e. coproduction, Browman and Goldstein 1992) in action based models. Moreover it can be shown that the variability of segmental phonetic surface descriptions of an utterance – as occurs in many languages if speaking rate is increased – directly results from increase in overlap in time and decrease in duration and decrease in degree of realization of vocal tract actions (Kröger 1993).

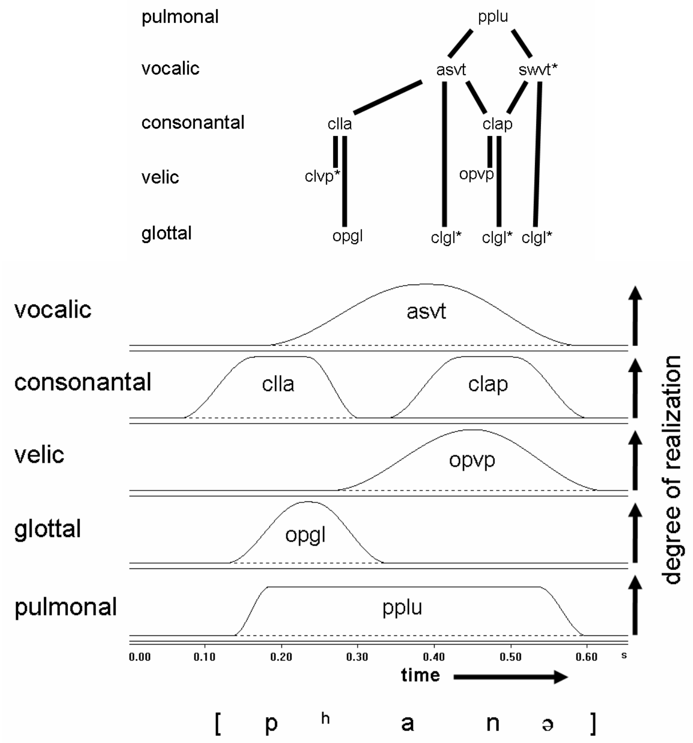

Fig. 1. Discrete score of speech actions (top) and high level phonetic speech action score (bottom) for the German word "Panne" ("glitch") using the action based control model of Birkholz et al. (2006). Top: Discrete specification of actions (4-letter-abbreviations) and their associations (lines between these abbreviations). Bottom: High level phonetic action specification, displaying (i) duration of activation time intervals for each vocal tract action and (ii) degree of realization for each vocal tract action. Action symbols (cp. Kröger and Birkholz 2007): vocal tract short /a/-shaping action (asvt), labial full closing action (clla), apical full closing action (clap), velopharyngeal opening action (opvp), glottal opening action (opgl), positive lung pressure action (pplu). In the case of full closing actions the maximum degree of realization indicates the temporal interval of full closure (see the consonantal full closing action clla and clap). Note that the activation interval for each action also includes onset and offset time intervals of the action. Default actions (default gestures) are marked by an asterisk (i.e. vocal tract schwa-shaping action swvt, velopharyngeal closing action clvp, glottal closing action clgl). Contours for degree of realization of actions are not shown for default actions within the action activation score. Note that action realization patterns not directly reflect the resulting time functions of articulator movements (i.e. lower level phonetic descriptions). These time functions are not displayed here.

A first control model for the articulation and phonation during singing has been established within an action based framework (Birkholz 2007). Here, beside the control of the articulation of speech sounds, in addition a concept is added which in parallel controls the production of notes (or tones) and which synchronizes the speech sound articulation – i.e. the *sound actions* – with respect to the musical time measure and notes – i.e. with respect to the *tone actions* (Fig 2.).

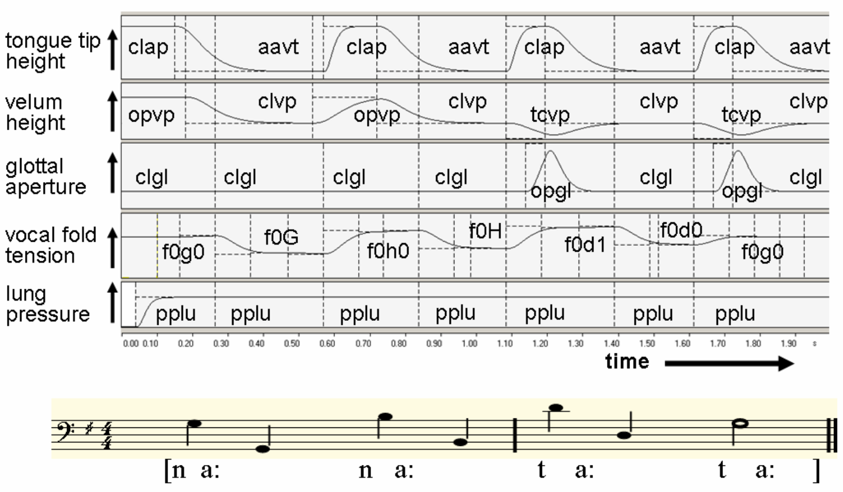

Fig. 2. Lower level phonetic action score for selected articulatory parameters (see on top: tongue tip height, velum height, glottal aperture, vocal fold tension, and lung pressure) displaying action specifications (onset time and target locations) and the resulting articulatory movement trajectories for a short singing phrase (for notes and text, see bottom). The tonal range used is that of a baritone singer. Sound action symbols (cp. Kröger and Birkholz 2007): vocal tract long /a/-shaping action (aavt), apical closing action (clap), velopharyngeal opening action (opvp), velopharyngeal closing action (clvp), velopharyngeal tight closing action (tcvp), glottal closing action (clgl), glottal opening action (opgl), positive lung pressure action (pplu). In this case of singing in addition tone actions for adjusting a specific fundamental frequency occur. These tone actions mainly adjust vocal fold tension. For example: g0 fundamental frequency action (f0g0). Occurring notes (or tones) from low to high: G, H, d0, g0, h0, d1 form different tone actions.

As is exemplified by the description of the action based control concept for speech and for singing by both examples (i.e. by the word "Panne", Fig. 1 and by the short singing phrase, Fig. 2) it is very essential to differentiate at least three levels of control: the *discrete and abstract functional (sometimes called phonological) level of action realization*, and two quantitative behavioral levels, the *quantitative higher behavioral or phonetic level of action realization*, quantitatively describing the degree of realization of a vocal tract action over time and the *quantitative lower behavioral or phonetic level of action realization*, quantitatively describing the spatiotemporal trajectories at least for the end articulators which are essential for realizing the goal of an action. Discrete tone action specification is that, what is exactly given by the notes in Fig.2 (bottom), i.e. musical tone pitch, tone length and discrete levels of tone dynamics (e.g. fortissimo, forte, mezzo forte, piano, pianissimo). High level phonetic tone action specification in addition includes further specifications like a detailed quantitative description of length, pitch, vibrato etc. and the lower level phonetic specification of a tone action includes a specification of how a tone is realized by a specific setting of articulatory (and phonatory) parameters like lung pressure and vocal fold tension.

4 Towards an Acting and Comprehending Articulatory Synthesis System Comprising Multimodal Interaction

4.1 The Importance of Interaction

The main limiting factor for high quality articulatory synthesis for speech as well as for singing is the problem how to get the *behavioral knowledge* for specifying quantitatively the *functional description* of actions. Main questions are: How can we get a quantitative description of articulatory and phonatory behavior in normal speech? How is articulatory and phonatory behavior modified in speech, if a syllable is more or less stressed, if a syllable is stressed emphatically, if the emotional state of a speaker changes (for example from normal to fear, to happiness, or to sadness), or if health conditions change? How is articulatory and phonatory behavior modified, if a speaker addresses an audience (i.e. gives a talk to an audience) or if a speaker is in a one-to-one communication situation (colloquial, informal or casual speech)? And how is articulatory and phonatory behavior modified if a singer performs different singing styles or if a singer increases from an amateur to a professional singer?

Many of these questions are not easy to solve. But there are a lot of arguments which indicate that it is important to embed synthesis models in a broader framework including aspects of human-human interaction. Human-human interaction is essential during learning phases for speaking or singing. Thus it can be assumed that synthesis systems are able to sound more natural if behavioral or control knowledge is collected during scenarios which are similar to natural speech acquisition scenarios (e.g. Bailly 1997, Guenther 2006, Guenther et al. 2006, Kröger and Birkholz 2007, Kröger et al. 2008). Therefore it is important to integrate articulatory synthesis models into a broader communication framework. Learning or acquiring of speech and singing implies communication with partners who are capable of commenting or judging the production trials of the synthesis model.

In the case of word learning (lexical learning) as well as in the case of learning tones or short spoken or sung phrases, the synthesis model (learner) and the expert (teacher) who judges the trials or items produced by the model, provide the basis for word acquisition (Fig. 3). The synthesis model hears an item (e.g. the word "dolly") produced by the expert and has an initial hypothesis how to reproduce (to imitate) the item. The expert listens to the items produced by the model (to the imitation items) and gives feedback. The expert can ignore, reject, or question ("what do your mean?") the item if the production is not understandable or not produced in a way which allows a classification as normal or at least understandable. Thereupon the synthesis model has the chance to modify the motor plan and to produce a further or corrected realization (version) of the item. This leads to a training loop (Fig. 3). If a vesion of an item is judged as acceptable by the expert, the expert will award this version of the item (e.g. "How wonderful! Yes that is a dolly"). This commendation causes the synthesis system (learner) to store the actual motor plan of the actual item as an acceptable motor representation of this word.

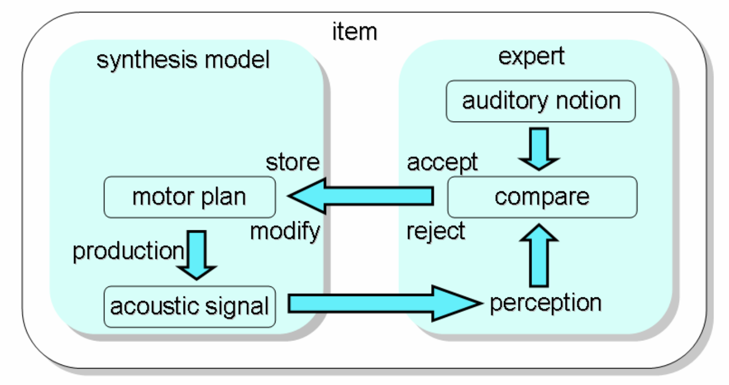

Fig. 3. The communication model for speech or singing acquisition. The model comprises a synthesis control model (or learner) and an expert (or teacher). The functioning of the communication model is described in the text.

Speech and non-specialized singing acquisition takes place during the first years of lifetime while choir singing training or solo-singing training occurs later during lifetime. But the human-human interaction scenario or learner-teacher interaction scenario described above holds for all these training situations. The term “item” introduced in Fig. 3 holds for speaking as well as for singing items.

4.2 The Importance of Action, Perception, and Comprehension

Beside this learner-teacher scenario a second scenario occurs during speech acquisition where the synthesis model by itself trains, judges, and learns acceptable realizations of spoken or sung items (e.g. words, single tones, spoken or sung phrases). This implies that the synthesis model comprises a perception and comparison component (see for example the feedback loop in models of speech production, as described by Guenther 2006). In this case the synthesis model starts by storing an auditory representation of a spoken or sung item which often has been produced by communication partners and tries actively to reproduce (to imitate) this item by himself (*action*). The model compares the perceptual representations of these self-productions with the stored auditory representation of that item (*perception*). A motor plan is stored as motor representation of this item if the auditory difference between already stored and the self-produced item falls below a certain acceptance limit. Thus in the context of Fig. 3 the synthesis model comprises (includes) the expert, since the model already includes a representation of the ideal training result and since the model comprises a perception and comparison module.

At least *comprehension*, i.e. the association between the function (concept or meaning) and the behavioral sensorimotor representation of a word, tone, or a spoken or sung phrase is a precondition for these learning (or action-perception) scenarios described above. Functional features represent the intention of a speech or singing act and thus are basic for each communication behavior.

4.3 The Importance of Multimodality

Firstly, during acquisition of speech and singing, a synthesis or production model not just stores an *auditory representation* and a *motor representation* of each speech or sung item. As a result from many imitation or self-production trials the model also stores a *somatosensory representation* of the item as well. Thus after successful training of a word or phrase, the model has stored an auditory, a somatosensory, and a motor plan representation of each item.

Secondly, human-human communication or interaction scenarios which are important for learning to speak or to sing, normally profit also from visual information, i.e. information which occurs if both communication partners gaze each other. Thus also *visual perception* should be included in speech and singing acquisition scenarios: The model or learner also stores a representation of the lips, jaw, chest and body movements for a trained spoken or sung item. It is well known for speech that even persons without hearing losses profit from visual facial information, for example if the auditory information is not fully available (e.g. in the case of disturbing noise, Schwartz et al. 2004). That is a proof for the fact that mental representations of facial speech movements (i.e. jaw and lip movements and in some cases also movements of the front part of the tongue if visible) exist.

These facts underline the need of multimodal data and an audiovisual learner-teacher scenario as an interactive framework for gathering knowledge for a control model for articulatory synthesis if high quality synthesis of speech or singing is aimed for. Beside speech production also in the case of singing production it becomes directly apparent that audiovisual learner-teacher scenarios are extremely needful. That can be audiovisual feedback from the mother or from caretakers in the case of acquisition of singing, audiovisual feedback of other chorister or of the conductor in the case of choir singing, or audiovisual feedback of the singing teacher in the case of solo-singing training.

5 Discussion

In this paper the state of the art for articulatory synthesis of speech and singing is discussed. One of the major findings is that the current vocal tract models and acoustics models including glottis and noise source models are capable of producing high quality acoustic output. This can be demonstrated at best for isolated *static* vowels and consonants (at least fricatives, laterals, nasals) in speech or singing (cp. Mawass et al. 2000, Birkholz 2007). The main problem is the generation of *control information* for the vocal tract *movements*, i.e. to generate natural *behavioral information*. In this paper it is argued for training these systems during learning to speak or to sing in a similar way as training and learning occurs during natural speech and natural singing acquisition. It is argued for developing *acting, perceiving, and comprehending virtual humans* in the context of a *multimodal learner-teacher interaction scenarios*. Therefore the development of a high quality articulatory-acoustic production or synthesis model always comprises high quality multimodal perception and comprehension tools including auditory, somatosensory, and visual perception of external as well as of feedback signals for developing high quality control modules for production. Thus it is feasible to take into account the close relationship between production and perception as is claimed in recent neurocognitive production-perception models (Hickok and Poeppel 2007).

Due to the learner-teacher scenario introduced in this paper it must be emphasized that action behavior (i.e. motor plans for speech and singing items) can be learned mainly from auditory or audiovisual information. It is not necessary to train articulatory production models for speech and singing by accessing a huge articulatory data base. If an articulatory-acoustic model and an action based control model is used, the production model comprises physiological constraints, which allow the generation of motor plans from auditory or audiovisual imitation processes without any other data (i.e. no articulatory data are needed). Motor plans can be optimized by a few trail and error loops as described above as the learner-teacher interaction scenario.

Last but not least, since the articulatory synthesis model already generates natural lip and jaw movements for speech and singing, it would be advantageous to incorporate the articulatory synthesis model within a complete facial model and moreover within a complete full length whole body model (avatar or virtual agent). And in addition the modeling of speech accompanying gesturing (eyebrow movements, head movements, hand and arm movements, whole body movements at least for the upper part of the body) could be added. Speech accompanying gestures can be described qualitatively and quantitatively in the same framework as vocal tract actions: There exist functional and behavioral levels for the description of actions, and action planning and action execution can be separated in a similar way. A further benefit from incorporating speech accompanying actions (or gestures) is that these actions facilitate the transfer of intentions and of meanings within communication and interaction scenarios. And it can easily been shown that speech accompanying gesturing increases the impression of the overall quality level of the speech or singing synthesis.

Acknowledgments. This work was supported in part by the German Research Council DFG grant Kr 1439/13-1.

References

Alipour, F., Berry, D.A., Titze, I.R.: A finite-element model of vocal-fold vibration. Journal of the Acoustical Society of America 108, 3003–3012 (2000)

Allen, D.R., Strong, W.J.: A model for synthesis of natural sounding vowels. Journal of the Acoustical Society of America 78, 58–69 (1985)

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Savariaux, C.: Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. Journal of Phonetics 30, 533–553 (2002)

Bailly, G.: Learning to speak: sensory-motor control of speech movements. Speech Communication 22, 251–267 (1997)

Beautemps, D., Badin, P., Bailly, G.: Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. Journal of the Acoustical Society of America 109, 2165–2180 (2001)

Birkholz, P.: Articulatory synthesis of singing. In: Bloothooft, G. (ed.) Synthesis of Singing Challenge. Antwerp, Belgium (2007), http://www.let.uu.nl/~Gerrit.Bloothooft/personal/SSC/index.htm

Birkholz, P., Kröger, B.J.: Vocal tract model adaptation using magnetic resonance imaging. In: Proceedings of the 7th International Seminar on Speech Production, pp. 493–500. Belo Horizonte, Brazil (2006)

Birkholz, P., Jackèl, D., Kröger, B.J.: Construction and control of a three-dimensional vocal tract model. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006) Toulouse, France, pp. 873–876 (2006)

Birkholz, P., Jackèl, D., Kröger, B.J.: Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech, and Language Processing 15, 1218–1225 (2007)

Browman, C.P., Goldstein, L.: Articulatory phonology: An overview. Phonetica 49, 155–180 (1992)

Cranen, B., Boves, L.: On subglottal formant analysis. Journal of the Acoustical Society of America 81, 734–746 (1987)

Cranen, B., Schroeter, J.: Physiologically motivated modeling of the voice source in articulatory analysis / synthesis. Speech Communication 19, 1–19 (1996)

Dang, J., Honda, K.: Construction and control of a physiological articulatory model. Journal of the Acoustical Society of America 115, 853–870 (2004)

El-Masri, S., Pelorson, X., Saguet, P., Badin, P.: Vocal tract acoustics using the transmission line matrix (TLM) method. In: Procedings of ICSPL, Philadelphia, pp. 953–956 (1996)

Engwall, O.: Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. Speech Communication 41, 303–329 (2003)

Fant, G.: Some problems in voice source analysis. Speech Communication 13, 7–22 (1993)

Fant, G., Liljencrants, J., Lin, Q.: A four-parameter model of glottal flow. Speech Transmission Laboratory - Quarterly Progress and Status Report 4/1985. Royal Institute of Technology, Stockholm, pp. 1–13 (1985)

Flanagan, J.L., Ishizaka, K., Shipley, K.L.: Synthesis of speech from a dynamic model of the vocal cords and vocal tract. The Bell System Technical Journal 54, 485–506 (1975)

Goldstein, L., Byrd, D., Saltzman, E.: The role of vocal tract action units in understanding the evolution of phonology. In: Arbib, M.A. (ed.) Action to Language via the Mirror Neuron System, pp. 215–249. Cambridge University Press, Cambridge (2006)

Guenther, F.H.: Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39, 350–365 (2006)

Guenther, F.H., Ghosh, S.S., Tourville, J.A.: Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language 96, 280–301 (2006)

Hickok, G., Poeppel, D.: Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences 4, 131–138 (2007)

Ishizaka, K., Flanagan, J.L.: Synthesis of voiced sounds from a two-mass model of the vocal cords. The Bell System Technical Journal 51, 1233–1268 (1972)

Kelly, J.L., Lochbaum, C.C.: Speech synthesis. In: Flanagan, J.L., Rabiner, L.R. (eds.) Speech Synthesis, Dowden, Hutchinson & Ross, Stoudsburg, pp. 127–130 (1962)

Kob, M.: Physical modeling of the singing voice. Unpublished doctoral thesis. RWTH Aachen University, Aachen (2002)

Kröger, B.J.: Minimal rules for articulatory speech synthesis. In: Vandewalle, J., Boite, R., Moonen, M., Oosterlinck, A. (eds.) Signal Processing VI: Theories and Applications, pp. 331–334. Elesevier, Amsterdam (1992)

Kröger, B.J.: A gestural production model and its application to reduction in German. Phonetica 50, 213–233 (1993)

Kröger, B.J.: Zur artikulatorischen Realisierung von Phonationstypen mittels eines selbstschwingenden Glottismodells. Sprache-Stimme-Gehör 21, 102–105 (1997a)

Kröger, B.J.: On the quantitative relationship between subglottal pressure, vocal cord tension, and glottal adduction in singing. Proceedings of the Institute of Acoustics 19, 479–484 (1997b) (ISMA 1997)

Kröger, B.J.: Ein phonetisches Modell der Sprachproduktion. Niemeyer Verlag, Tübingen (1998)

Kröger, B.J.: Ein visuelles Modell der Artikulation. Laryngo-Rhino-Otologie 82, 402–407 (2003)

Kröger, B.J., Birkholz, P.: A gesture-based concept for speech movement control in articulatory speech synthesis. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 174–189. Springer, Heidelberg (2007)

Kröger, B.J., Lowit, A., Schnitker, R.: The Organization of a Neurocomputational Control Model for Articulatory Speech Synthesis. In: Esposito, A., Bourbakis, N., Avouris, N., Hatzilygeroudis, I. (eds.) Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction. LNCS (LNAI), vol. 5042, pp. 121–135. Springer, Berlin (2008)

Liljencrants, J.: Speech Synthesis with a Reflection-Type Line Analog. Dissertation, Royal Institute of Technology, Stockholm (1985)

Maeda, S.: A digital simulation of the vocal-tract system. Speech Communication 1, 199–229 (1982)

Maeda, S.: An articulatory model based on statistical analysis. Journal of the Acoustical Society of America 84 (supl.1), 146 (1988)

Matsuzaki, H., Motoki, K.: FEM analysis of 3D vocal tract model with asymmetrical shape. In: Proceedings of the 5th Seminar on Speech Production, pp. 329–332. Seeon, Germany (2000)

Mawass, K., Badin, P., Bailly, G.: Synthesis of French Fricatives by Audio-Video to Articulatory Inversion. Acta Acustica 86, 136–146 (2000)

Mermelstein, P.: Articulatory model for the study of speech production. Journal of the Acoustical Society of America 53, 1070–1082 (1973)

Meyer, P., Wilhelms, R., Strube, H.W.: A quasiarticulatory speech synthesizer for German language running in real time. Journal of the Acoustical Society of America 86, 523–540 (1989)

Saltzman, E.L., Munhall, K.G.: A dynamic approach to gestural patterning in speech production. Ecological Psychology 1, 333–382 (1989)

Saltzman, E., Byrd, D.: Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. Human Movement Science 19, 499–526 (2000)

Schwartz, J.L., Berthommier, F., Savariaux, C.: Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93 B69- 78, B69–B78 (2004)

Serrurier, A., Badin, P.: A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. Journal of the Acoustical Society of America 123, 2335–2355 (2008)

Sinder, D.J.: Speech synthesis using an aeroacoustic fricative model. PhD thesis, Rutgers University, New Jersey (1999)

Sondhi, M.M., Schroeter, J.: A hybrid time-frequency domain articulatory speech synthesizer. IEEE Transactions on Acoustics, Speech, and Signal Processing 35, 955–967 (1987)

Story, B.H., Titze, I.R.: Voice simulation with a body cover model of the vocal folds. Journal of the Acoustical Society of America 97, 1249–1260 (1995)

Titze, I.R.: A four-parameter model of the glottis and vocal fold contact area. Speech Communication 8, 191–201 (1989)

Wilhelms-Tricarico, R.: Physiological modelling of speech production: Methods for modelling soft-tissue articulators. Journal of the Acoustical Society of America 97, 3085–3098 (1995)