Chapter 2
The Perception of Family and Register in Musical Tones

Roy D. Patterson, Etienne Gaudrain, and Thomas C. Walters

2.1   Introduction

This chapter is about the sounds made by musical instruments and how we perceive them. It explains the basics of musical note perception, such as why a particular instrument plays a specific range of notes; why instruments come in families; and why we hear distinctive differences between members of a given instrument family, even when they are playing the same note. The answers to these questions might, at first, seem obvious; one could say that brass instruments all make the same kind of sound because they are all made of brass, and the different members of the family sound different because they are different sizes. But answers at this level just prompt more questions, such as: What do we mean when we say the members of a family produce the same sound? What is it that is actually the same, and what is it that is different, when different instruments within a family play the same melody on the same notes? To answer these and similar questions, we examine the relationship between the *physical* variables of musical instruments, such as the length, mass, and tension of a string, and the variables of auditory *perception*, such as pitch, timbre, and loudness. The discussion reveals that there are three *acoustic* properties of musical sounds, as they occur in the air, between the instrument and the listener, that are particularly useful in summarizing the effects of the physical properties on the musical tones they produce, and in explaining how these musical tones produce the perceptions that we hear.

The remainder of the introduction sets out the aspects of tone perception to be explained, namely, the perception of pitch, instrument family, and instrument register within a family. The second section describes the acoustic properties of tones as they pertain to music perception, and sets out some of the terminology used in the chapter. The third section explains the relationship between the physical

R.D. Patterson (✉)
Centre for the Neural Basis of Hearing, Department of Physiology,
Development and Neuroscience, University of Cambridge, Cambridge CB2 3EG, UK
e-mail: rdp1@cam.ac.uk

variables of tone production (length, mass, tension, etc.) and the acoustic variables observed in the sounds. The fourth section describes the internal representation of musical sounds in the auditory system to show how the acoustic properties of sound appear in the auditory representation of musical tones. The fifth and final section reviews the relationship between the acoustic variables of sound and the auditory variables of tone perception, and suggests how the standard definitions of pitch and timbre might be revised for use in discussions of the perception of musical tones and musical instruments.

2.1.1   Pitch, Instrument Family, and Instrument Register Within a Family

The chapter focuses on the sounds produced by the sustained-tone instruments of the orchestra and chorus, that is, the families of instruments referred to collectively as brass, strings, woodwinds, and voice. Table 2.1 shows four of the instruments in each of the families, ordered in terms of their size or their *register*. With just a small amount of training, most people can learn to identify these 16 instruments from a simple monophonic melody (van Dinther and Patterson 2006). With regard to family and register, the purpose of the chapter is to explain how auditory perception enables us to distinguish the main families and the different instruments within a family.

Imagine the sequence of tones you would hear if a trombonist, a cellist, a bassoonist, and a baritone vocalist would in turn produce the same tone, say C3 (the C below middle C on the keyboard). What is the "same" about the four tones is their pitch. What is different, and what allows us to distinguish the tones, is the distinctive timbres of the different instrument families. This is the traditional distinction between the *perceptual* variables, pitch and timbre. The pitch of a musical tone is effectively determined by the repetition rate of the sound. The sound waves produced by the sustained-tone instruments of the orchestra (brass, string, woodwind, and voice) are complex and their spectra are complex; nevertheless the tones are essentially periodic and the pitch that they produce is very closely related to the number of times that the sound wave repeats in the course of 1 s. This aspect of music perception is entirely straightforward for sustained-tone instruments. Psychoacousticians have developed models to explain how the auditory system extracts pitch from sound waves, and the models have become increasingly elaborate as they attempt to explain the pitches produced by exotic, computer-generated

Table 2.1 Sixteen common instruments illustrating four registers within each of four instrument families

Register/family	Brass	Strings	Woodwind	Voice
High	Trumpet	Violin	Soprano sax	Alto voice
Mid-high	Trombone	Viola	Alto sax	Tenor voice
Low-mid	French Horn	Cello	Tenor sax	Baritone voice
Low	Tuba	Contra bass	Baritone sax	Bass voice

waveforms, and the relative salience of these esoteric pitch perceptions. The models fall into two groups: those that follow Helmholtz (1875) and attempt to explain the perception of pitch on the basis of the frequency spectra of the sounds, and those that follow Licklider (1951) and emphasize the distribution of time intervals observed in the firing patterns that pitch-producing sounds generate in the auditory nerve. A brief overview of the debate is presented in Sect. 2.4 of this chapter; more extensive discussions are provided in a recent article by Yost (2009) and a recent chapter by de Cheveigné (2005). Despite the passion of the debate between the spectral and temporal modelers, for readers who are simply interested in the relationship between the physics of note production and perception, the pitch of the notes of the main orchestral instruments is simply the psychological correlate of the repetition rate of the waveform that the instrument produces.

With regard to timbre, the instruments of a given family have similar physical shapes, they are made of similar materials, and they are excited in similar ways, so it is not surprising that the instruments of a family produce tones with a similar sound quality, or timbre, that distinguishes the family. The categories of timbre associated with instrument families are labeled with words that describe some physical aspect of the source. So, the trumpet is a *brass* instrument, the clarinet is a *wood-wind* instrument, and the violin is a *string* instrument. The *family* aspect of timbre is largely determined by the shape of the envelope of the magnitude spectrum of the tones that the instrument produces. This aspect of musical perception is also relatively straightforward for sustained-tone instruments.

Within a family of instruments, the different members are distinguished physically by their size, and perceptually by the effect that the size of the instrument's components has on the tones they produce. There are two different aspects to instrument size, and they jointly determine our perception of the *register* of an instrument within its family. In the string family, register distinguishes the violin, viola, cello, and double bass, and the instrument names are normally used to specify the instrument's register. In the string family, as the size of the instrument increases from violin to double bass, the lengths and masses of the strings increase, and so the tones of the larger instruments have lower pitches (on average). The range of pitches that an instrument produces is one of the properties that determine the register we perceive and what instrument we hear within a family. The second aspect of instrument size is the size of the body and it also affects the register we perceive and the instrument we hear; larger bodies go with lower registers. The fact that register depends on two acoustic variables means that the perception of register is somewhat more complicated than the perception of pitch and family timbre. Nevertheless, the principles, as they pertain to the perception of musical tones, are readily comprehensible and they are a prominent topic in this chapter. To begin with, register can be regarded as the perceptual property that enables us to distinguish the size element of instruments within a family (Table 2.1) including the categorization of humans as sopranos, altos, tenors, baritones, or basses. Note that children, when they begin to sing, are sopranos and they progress down in pitch to their eventual range as they grow up.

In summary, the main purpose of this chapter is to describe how the *physical* variables of tone generation are related to the *acoustic* variables of tones as sounds in the air, and how these acoustic variables are related to the *perception* of melodic pitch, family timbre, and register within an instrument family. There is a secondary aspect of register, associated with the perception of individual instruments, that allows us to distinguish the upper and lower notes by the sound of the tones themselves, and we refer to the tones as coming from the upper or lower "register" of a particular instrument, or voice. We return to this secondary aspect of tone perception later in the chapter, once the acoustic properties of sound, and their primary role in perception, have been set out.

2.2   Pulse-Resonance Sounds and Acoustic Scale

The tones that one hears in the natural environment are typically "pulse-resonance" sounds (Patterson et al. 2008), for example, the calls that mammals, birds, frogs, and fish use to declare their territories or attract mates (e.g., Fitch and Reby 2001). The vowels of speech and the sustained tones of orchestral instruments are also pulse-resonance sounds. So they are the normal tones that one hears every day in the man-made environment and in the natural world.

2.2.1   Origin of Pulse-Resonance Sounds

The production of a pulse-resonance sound is conceptually simple. The animal just has to develop some means of producing an acoustic pulse that will, then, resonate in one or more of the structures in the animal's body. Once the basic mechanism arises in response to the need for communication, evolution can refine the sound with successive modifications to make it more distinctive and efficient. In present-day animals, the pulse generating mechanism typically produces a stream of pulses that occur regularly in time, and in models of tone production, the mechanism that produces the stream of pulses is referred to as "the source" of the sound. The resonances in the animal's body are collectively referred to as a "the filter," and in most animals, the filters have evolved to give the animal's call a distinctive timbre. The stream of pulses with their resonances forms a tone, and these tones provide the basis for animal communication. They also broadcast the species of the caller.

In almost all mammals, the source mechanism is the vocal folds in the larynx at the base of the throat; they produce pulses by momentarily impeding the flow of air from the lungs. The pulses of air then excite resonant cavities in the airway between the larynx and the lips, and this filter of resonant cavities is referred to as the vocal tract. A short segment of a synthetic /a/ that sounds like the vowel in "car" is presented in Fig. 2.1a. The wave shows that the sound is periodic and each cycle contains an acoustic pulse followed by a decaying resonance with a complex shape. A vowel is normally on the order of 100–300 ms in duration, so the complete
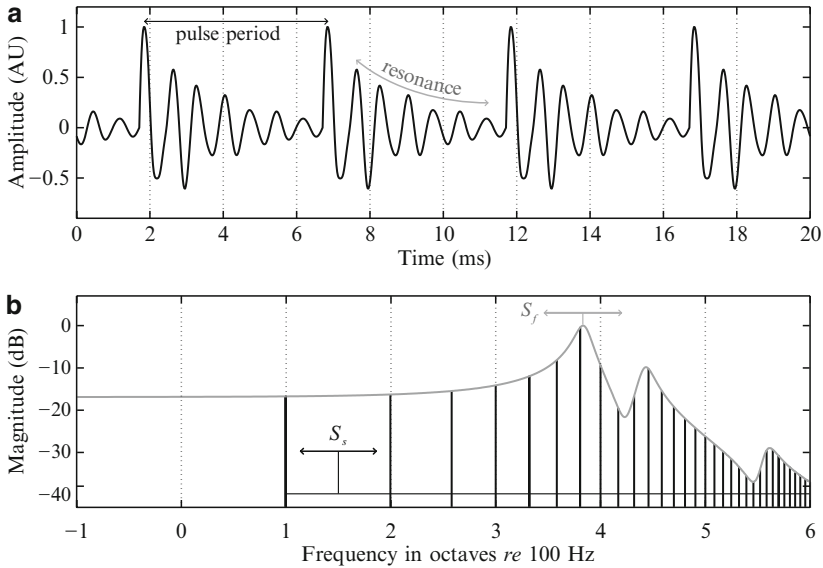
Fig. 2.1 The waveform and magnitude spectrum of a child's vowel /a/. (**a**) The waveform, which is a plot of acoustic pressure as a function of time, shows a repeating pattern that starts with a pulse. The repetition period, or pulse period, is shown by the black arrow. Each pulse is followed by a resonance that decays in time, as shown by the gray arrow. (**b**) The long-term magnitude spectrum, that is, the distribution of energy across frequency, is composed of harmonics represented by the vertical black lines that form the *fine-structure* of the spectrum. The frequency axis is logarithmic and scaled in number of octaves *re* 100 Hz. The position of the fine-structure, that is, the position of the set of harmonics taken as a unit, is the acoustic scale of the source S_s. This quantity is related to the pulse period shown on the waveform. The spectral envelope, shown in gray, depicts how the resonators in the vocal tract filter the pulses. Its shape determines the vowel type. Its position on the log-frequency axis is the acoustic scale of the filter, S_f

waveform for the /a/ in "car" would contain 20–60 of the pulse-resonance cycles shown in Fig. 2.1a. The waveform repeats every 5 ms so the "repetition rate" of the tone is 200 cycles per second (cps), and this value is used to specify its pitch.

These are the main characteristics of pulse-resonance sounds as they appear in the time domain. Many birds and frogs also excite resonances in their air passages by momentarily interrupting the flow of air from the lungs, although the details of the source and filter mechanisms are somewhat different. Fish do not have air passages but many of them have swim bladders that resonate and function as the filter. The bladder is excited by muscles in the wall of the swim bladder (e.g., the weakfish, *Cynoscion regali*) that produce brief mechanical pulses referred to as "sonic twitches." This muscle source produces twitches in regularly timed streams (Sprague 2000). A brief introduction to the pulse-resonance sounds produced by animals is presented in Patterson et al. (2008).

Pulse-resonance tones are very different from environmental noises such as wind in the trees or waves on the beach, or man-made noises like extractor fans, jet engines,

or the boiling of a kettle. Noises arise from turbulent systems where the source vibrates randomly. Noise waveforms are not periodic and so they do not produce salient pitch perceptions. The filtering is incidental and evolution is not involved in tuning the filter to make the sound distinctive or improve communication. One continuous noise sounds much like another when they have the same loudness. Perceptually, pulse-resonance tones, with their pronounced pitch and distinctive timbre, tend to capture the listener's attention, whereas continuous noises are commonly ignored.

Returning to musical sounds, the sustained tones that singers produce when the voice is used as an instrument are vowels, and so the singing voice produces pulse-resonance tones. The instruments of the brass, string, and woodwind families also produce pulse-resonance tones (van Dinther and Patterson 2006). Each of the families has a source mechanism that produces regular streams of pulses that are filtered by resonances in the instrument's body (Fletcher and Rossing 1998). Several examples are presented in Sect. 2.3. The remainder of this section describes the acoustic properties of pulse-resonance tones as they appear in the magnitude spectra of the sounds, and how the properties do, *or do not*, vary with the size of the instrument or singer.

2.2.2 Acoustic Properties of Pulse-Resonance Sounds

The set of vertical lines in Fig. 2.1b shows the long-term *magnitude spectrum* of the vowel, that is, the distribution of energy across frequency, averaged over 100 ms, or more, of time. The frequency axis is logarithmic in this case, similar to the place, or "tonotopic," dimension of the cochlea. The vertical lines show that the energy is restricted to frequencies that are integer multiples of a single, *fundamental* frequency, designated F0. The fundamental of this harmonic series, and the frequency spacing between the harmonics (Fig. 2.1b), are the spectral representation of the repetition rate of the sound, which is the inverse of the period observed in the waveform (Fig. 2.1a). In this example, all three of these acoustic variables have the value 200 cps. The dashed line connecting the tops of the harmonics in the lower panel shows the *spectral envelope* of the vowel.

The soft shouldered peaks that appear in the spectral envelopes of speech sounds are referred to as formants. Individual formants are normally designated by the frequency of the peak in the envelope, but the concept of a formant actually includes the shape and width of the envelope in the region of the peak, as well as the peak frequency. The shape that the set of formants collectively impart to the envelope in the spectral domain (Fig. 2.1b), is related to the shape of the damped resonance following each glottal pulse in the time domain (upper panel). The resonators in the bodies of musical instruments do not produce such distinctive formants as the resonances of the vocal tract, but the principles are the same for all pulse-resonance sounds. The shape of the spectral envelope corresponds to the shape of the resonance in the waveform, and this shapes determine the

distinctive sound quality, or timbre, of an instrument family. The set of harmonics that constitute the magnitude spectrum of a sound will be collectively referred to as the *fine-structure* of the spectrum to distinguish the magnitude spectrum (solid vertical lines) from its envelope (gray line).

Now consider the changes that occur in the tones of a specific instrument family as the size of the instrument increases. For example, consider what happens to vowel sounds as children grow into adulthood. When children begin to speak they are about 0.85 m tall, and their height increases by about a factor of two as they mature. In humans (and other animals), the source and filter are components of the body and both the source and the filter increase in size as young mature into adults. With regard to the source in humans, the glottal pulse rate (GPR) decreases by about an octave as the child grows up and the vocal cords become longer and more massive. The decrease in GPR is greater than an octave for males and less than an octave for females, but even for females, it is a large change. With regard to the filter, vocal tract length increases in proportion to height (Fitch and Giedd 1999; Turner et al. 2009, their Fig. 4), and as a result, the formant frequencies of children's vowels decrease by about an octave as they mature (Lee et al. 1999; Turner et al. 2009). The effects of growth on the fine-structure and envelope of the spectrum of a vowel are quite simple to characterize, provided the spectrum is plotted on a logarithmic frequency scale. In this case, the set of harmonics that define the fine structure of the spectrum (the vertical lines in Fig. 2.1b) moves, *as a unit*, toward the origin as the child matures into an adult. In speech, the pattern of formants that defines a given vowel type remains largely *unchanged* as people grow up (Peterson and Barney 1952; Lee et al. 1999; Turner et al. 2009). In other words, for a given vowel, the shape of the spectral envelope does not change as a child matures; rather, the spectral envelope just shifts slowly toward the origin, moving about an octave in total as a child matures into an adult. Thus, in the current example, the vowel remains an /a/, and does not change to an /e/, an /o/ or an /u/, as a child matures into an adult.

The "position of the spectral envelope of a sound on a logarithmic frequency axis" is a property of a sound as it occurs in the air (Cohen 1993). For a pulse-resonance tone, this property is the *acoustic scale* of the filter that defines the resonances, and in the case of the human voice, it is closely related to vocal tract length (a physical variable). The "position of the fine-structure of the spectrum on a logarithmic frequency axis" is also a property of a sound as it occurs in the air. For a pulse-resonance tone, it is the *acoustic scale* of the source, and in the case of the human voice, it is closely related to glottal pulse rate (a physical variable). The two acoustic scale variables are very useful for summarizing the effects of physical variables such as mass and length on the perceptions produced by instruments, and, as a result, they play a prominent role in the remainder of the chapter. For brevity, "the scale (S) of the source (s)" is designated S_s, and "the scale (S) of the filter (f)" is designated S_f. Turner et al. (2009) have recently reanalyzed several large databases of spoken vowels and shown that almost all of the variability in formant frequency data that is not vowel-type information is S_f information. To reduce confusion between the two acoustic scale variables, S_s and S_f, we use cycles per

second (cps) for the units of the scale of the source, S_s, and kiloHertz (kHz) for the scale of the filter, S_f, since the position of the spectral envelope and the unit for the frequency dimension of the magnitude spectrum is Hertz.

In summary, the important distinctions for the remainder of the chapter are as follows:

1. The pulse rate of the source is a physical variable (e.g., GPR). It determines the repetition rate of the wave, which is known as the acoustic scale of the source, S_s. Repetition rate and S_s are both acoustic variables, and they in turn, determine the pitch of a pulse-resonance tone Pitch is a perceptual variable.
2. The size of a resonator in the body of a person or an instrument is a physical variable (such as length or volume). It determines the rate at which the resonance oscillates in the waveform (van Dinther and Patterson 2006), and it determines the *position* of the spectral envelope along the frequency axis of the magnitude spectrum. It is known as the acoustic scale of the filter, S_f, and it is an acoustic variable that affects the perception of source size and the perception of register within an instrument family.
3. The shape of the spectral envelope determines the instrument family aspect of timbre.
4. Register is the term used to describe the joint action of the acoustic variables, S_s and S_f, on the perception of musical tones and instruments. The values of S_s and S_f reflect the physical sizes of the source and filter in the instrument, respectively, and so the perception of register is closely related to the perception of instrument size, or singer size. The vocal terms *soprano*, *alto*, *tenor*, and *bass* are commonly used to specify register within families, as in tenor sax or bass fiddle.
5. Finally, note that the voice differs from other instruments with respect to timbre, in one important regard. When vowel type changes, say, from /a/ to /i/, the shape of the envelope changes. The shape does not change with the size of the singer from child to adult, whereas the acoustic scale values, S_s and S_f, do. So, different vowels are like different instrument families in the perception of musical tones. One useful, and reasonable, way to think of vowels is that they form a cluster of instrument families (unified by the fact that they are perceived to come from humans) and that the differing timbres of the members of this family are somehow more similar to each other than they are to the timbres of other musical instrument families.

2.2.3   Terminology

2.2.3.1   Source

There are many meanings of the word *source* in the description of sounds and how they are produced. To avoid confusion when reading this chapter, focus on what the *source* is a source *of*. So, when listening to an orchestra, one specific musician, and

the instrument they are playing, jointly form the source of one of the streams of musical tones that the orchestra is producing. In contrast, the "source" of the *energy* in these tones is the arm of the musician, in the case of string instruments, and the diaphragm of the singer in the case of a vocalist. The "source" in a *source-filter system* is a mechanism in between the source of the energy and the complete instrument in combination with the musician. In the source-filter description of tone production, the word "source" means the mechanism that produces the stream of abrupt amplitude changes, or pulses, which subsequently excite the set of resonances in the body of the instrument, or the vocal tract of the singer. It is a specialized meaning of the word "source," but it is straightforward and it is the only use of the word "source" in this chapter.

2.2.3.2   Noise

Throughout the current chapter, we use the word "noise" as an acoustic term that refers to the fact that the waveform is aperiodic and the amplitude varies randomly with time. These sounds are typically heard as background sounds and do not draw your attention. There is, of course, another use of the word "noise" that can occur in a musical context. For example, when there are competing sounds in an environment, perhaps a Mozart symphony on the radio and a rock concert on television, an individual listener might say, "Turn off that noise!," referring to the source which is interfering with the source they are trying to hear. The current chapter is not concerned with multisource environments and so the latter use of "noise" does not arise in this chapter.

2.2.3.3   Scale

In the phrase "acoustic scale" the word *scale* is being used in the mathematical sense, rather than the musical sense. In mathematics "a scale factor" is a number that tells you how big one value is relative to another. A musical scale is a set of frequency intervals within an octave. There is a connection between the two uses of scale inasmuch as the intervals of a musical scale (such as a fifth) are defined by specific scale factors (~1.5 in the case of a fifth), but acoustic scale refers to a single value rather than a set of musical scale values.

2.3   The Pulse-Resonance Tones of Musical Instruments

This section describes how the sustained-tone instruments of the orchestra produce their tones, and the relationship between the physical properties of the instrument on the one hand, and the three main acoustic properties of these sounds on the other hand.

2.3.1 The Source of Excitation and the Acoustic Scale Variable, S_s

In general terms, the "source" in these instruments is a highly nonlinear, resonant system that produces a temporally regular stream of acoustic pulses. The mechanism is conceptually similar for the voice, brass instruments, and woodwind instruments; in these instruments, the source momentarily interrupts the flow of air from the lungs, and it does so regularly in time. The individual mechanisms are, however, quite diverse. For example, the source is the vocal folds in the case of the voice; whereas, in brass instruments, it is the lips coupled to the main tube via the mouthpiece; and in the woodwinds, it is the lips coupled to the main tube via the reed. In string instruments, the mechanism is completely different; it is the bow coupled to a string. Despite the diversity of mechanisms, all of the sources produce streams of very precise acoustic pulses (brass and woodwinds), or abrupt changes in amplitude (strings) that function in a similar way. As a result, the sound waves produced by sustained-tone instruments are all pulse-resonance sounds. (In Fourier terms, the overtones of the pulse rate are locked to the pulse times both in frequency *and phase* up to fairly high harmonic numbers.)

The acoustic scale of the source of excitation is termed the *source scale* or S_s; it is effectively the repetition rate of the wave as it occurs in the air between the instrument and the listener. S_s is determined by physical properties of the instrument, like length and mass, which are not themselves acoustic variables. S_s largely determines the pitch we hear, but S_s is not itself an auditory variable. It is an intervening, acoustic variable that describes a property of the sound in the air, and it should be distinguished from pitch, which is the auditory variable of perception. The relationship between S_s and the physical variables of the instrument will be illustrated by comparing how S_s is determined in the vocal tract and in string instruments.

2.3.1.1 The Source of Excitation in the Human Voice

The vocal folds produce glottal pulses in bursts and, although the vocal folds are rather complicated structures, the effects of the physical variables on the rate of pulses can be described using the expression for a tense string. The glottal pulse rate, GPR, is largely determined by the length, L, mass, M, and tension, T, of the vocal folds, and the form of the relationship is

$$GPR \propto \sqrt{\frac{T}{ML}} \qquad (2.1)$$

Two of these physical variables are determined by the size of the person – the length and mass of the vocal folds. Both of these variables increase as a child

grows up, and both of these terms are in the *denominator* on the right-hand side of the equation, so as the child *increases* in height the pitch of the voice *decreases*. The average GPR for small children is about 260 cps, both for males and females. For females, GPR just decreases with height throughout life dropping to, on average, about 160 cps in adult women. For males, GPR decreases with height until puberty at which point the vocal folds suddenly increase in mass and the GPR drops to, on average, about 120 cps. So the length and mass of the vocal folds are a major determinant of vocal register, that is, whether a singer is a soprano, alto, tenor, baritone, or bass.

To produce a melody, a singer varies the tension of his or her vocal folds. So learning to sing in tune is largely a matter of learning to control the tension of the vocal folds – holding the tension fixed during sustained notes and changing it abruptly between notes. Tension is in the *numerator* of the mathematical expression (2.1), and so as a singer *increases* the tension, he or she *increases* the GPR. There is considerable overlap in the note ranges of the soprano, alto, tenor, baritone, and bass voices; in fact, the highest note of a bass is typically a note or two above the lowest note of a soprano. The effect of all three of these variables (T, M, and L) on GPR is constrained by the fact that the GPR value is related to the square root of these variables. So, for example, a singer has to change the tension of the voice by a factor of four to produce a one-octave change that would double the GPR.

In summary, for a specific individual, the size of the vocal folds (length and mass) determines the individual's long-term average GPR, and it determines the S_s component of the register of their voice. The tension of the vocal folds is varied to produce a melody. So, the long-term average S_s value, calculated over a sequence of musical phrases, reveals the register of the singer's voice; short-term deviations of S_s from the longer-term average, in discrete steps with regular timing, are the hallmarks of vocal melody.

2.3.1.2   The Source of Excitation in the String Family

The excitation mechanism in stringed instruments is the string pushed by the bow. As the musician draws the bow across a string, the string is pushed or pulled away from its resting position until the tension becomes too great, at which point, it snaps back, producing an abrupt, unidirectional change in amplitude. The direction is opposite to the direction that the bow is moving. The result is, nevertheless, a pulse-resonance sound inasmuch as the harmonics are locked in phase, and the internal representation of the sound has a pulse-resonance form in any given frequency band. Although the bow-string system is rather complicated physically (McIntyre et al. 1983), the relationship between pulse rate, PR, and the main physical variables is the same as for the vocal folds, namely,

$$\mathrm{PR} \propto \sqrt{\frac{T}{ML}} \qquad (2.2)$$

In this case, however, *T*, *M*, and *L* refer to the tension, mass, and length of the string, rather than to the corresponding properties of the vocal folds. The two physical variables associated with the size of the source (the length and mass of the string) are the most important excitation variables in this family of instruments and they each have *two* roles to play. Consider first the pulse rates of the open-strings on these instruments: Both the mass and length variables are in the denominator on the right-hand side of the equation, so *increases* in size, be they length or mass, lead to *decreases* in pulse rate. For a given member of the family (violin, viola, cello, or contra bass), the length of the four strings is fixed, and as the size of a family member increases, the string length gets longer in discrete steps. As a result, string length plays an important role in determining register within the string family. The mass of the string increases with its length, so it also contributes to the register we perceive. Mass also plays an important role in determining the range of notes that an individual instrument can play; the mass is varied across the four strings to extend the range beyond that which can be provided on any one string. Finally, the musician varies the length of individual strings to produce the different tones within that string's range.

Instrument makers are very adept at using mass and length to vary the pulse rate of notes within a family. If a musician depresses the lightest string on the largest instrument (the contra bass) at a point near the bridge on the neck, the pulse rate of the note will actually be a little higher than the pulse rate of the open-string note of the heaviest string on the smallest member of the family (the violin). In both cases, the notes are just below middle C on the keyboard.

2.3.1.3 Excitation Mechanisms of the Woodwind and Brass Instrument Families

The excitation of woodwind and brass instruments is described in terms of fluid mechanical "valves" that momentarily close the flow of air through the instrument. The closure causes a sharp acoustic pulse that resonates in the tube beyond the mouthpiece. For woodwind instruments, the valve is the reed in conjunction with the lips. For brass instruments, the source is not clearly localized within the instrument. The source of *energy* is the stream of air produced by the player who controls the pressure with the tension of the lips. The source of excitation is pulsatile because the mouthpiece is coupled to the tube between the mouthpiece and the bell (i.e., the body of the instrument), and the tube can only resonate at certain frequencies. Thus, the pulses originate from the lips, but the pulse rate is determined by the effective length of the tube, and this functional tube length is varied by the valves (or the slide) to control the pulse rate of the note.

Despite the complexities of excitation, these two families of instruments produce pulse-resonance sounds in which the acoustic scale of the source S_s controls the repetition rate of the note, and thus contributes to define the instrument's register within its family. The pulsatile nature of the excitation generated by these systems, and the temporal regularity of the pulse stream, mean that the dominant components

of the spectrum are strictly harmonic and they are phase locked (Fletcher and Rossing 1998). Fletcher (1978) provides a mathematical basis for understanding the origin of the phase locking, which is referred to as mode locking in musical instrument theory. Detailed descriptions of the mechanisms are provided in Benade (1976), Fletcher (1978), and McIntyre et al. (1983); a brief overview is provided in van Dinther and Patterson (2006).

2.3.1.4   Summary of the Role of S_s in Determining Melody and Register Within a Family

Comparison of the excitation mechanisms for the different instrument families shows that these mechanisms are similar, inasmuch as they all produce regular streams of pulses and the pulse rate is affected in the same way by the size of the components in the source. As a result, pulse rate decreases as instrument size increases in all of these instrument families. At the same time, the method whereby the pulse rate is varied to produce a melody is fundamentally different: the variable that controls pulse rate in the voice is the tension of the vocal folds, and the singer *increases* the tension to *increase* the pulse rate; whereas the variable that controls pulse rate in string instruments is string length, and the musician *decreases* the length to *increase* the pulse rate. The brass and woodwind instruments are like the strings, inasmuch as the pulse rate is varied to produce a melody by varying the length of part of the instrument; brass and woodwind instruments are different from the strings inasmuch as the length in this case is tube length rather than string length.

Although different instrument families employ very different mechanisms to produce acoustic pulses (and it is important for musicians to understand something of these mechanisms in order to play their instruments properly), all of these instruments nevertheless produce pulse-resonance tones, and the melody information in music is a sequence of pulse-rate values that specify the momentary acoustic scale of the source of excitation. Although the relationship between the physical variables involved in instrument excitation and the repetition rate of a given note is complex, the relationship between the acoustic-scale variable, S_s, which summarizes the action of the source, and the pitch we perceive is straightforward.

2.3.2   The Filtering of the Excitation Pulses and the Acoustic Scale of the Filter, S_f

The "filter" in musical instruments is a set of resonators that increase in size with register within an instrument family, and together the resonators determine the acoustic scale of the filter, S_f. Each of the pulses produced by the excitation mechanism of a sustained-tone instrument is filtered by body resonances within the instrument. In the time domain, it is these resonators in the body of the instrument that produce the resonances that appear attached to each pulse in the waveform (e.g., Fig. 2.1a).

In the frequency domain (e.g., Fig. 2.1b), the body resonances produce the distinctive shape of the envelope of the magnitude spectrum, and consequently, they determine the timbre of the family. In the case of the voice, the dominant resonances are associated with the larger cavities of the vocal tract (Chiba and Kajiyama 1941; Fant 1960). The tongue makes a constriction in the vocal tract that divides it into a mouth cavity and a throat cavity. These cavities resonate like tubes and/or bottles and they introduce formant peaks into the vowel spectrum (Fig. 2.1b). The tongue position is varied to produce the different vowels. This changes the relative sizes of the cavities, and thus, the relative positions of the formants in the spectrum (Chiba and Kajiyama 1941; Fant 1960). For stringed instruments, the most important resonances are associated with the plates of the body (wood resonances), the body cavities (air resonances), and the bridge (structural resonances) (Benade 1976). For brass and woodwind instruments, the prominent resonances are associated with the shape of the mouthpiece, which acts like a Helmholtz resonator, and the shape of the bell, which determines the efficiency with which the spectral components radiate into the air (Benade and Lutgen 1988). Woodwind instruments are like brass instruments, but the materials are different. So, just as there are many source mechanisms for generating the pulse stream, there are many systems of body resonances that lead in turn to many distinctive spectral envelopes.

Within a family of instruments, the most prominent distinction between the members of the family is the size of the body of the instrument, and the primary effect of instrument size on the perception of register within a family is straightforward (van Dinther and Patterson 2006): If the size of an instrument is changed while keeping its shape the same, the result is a proportionate change in S_f, the acoustic scale of the filter mechanism in the body of the instrument. That is, if the three spatial dimensions of an instrument are increased by a factor, a, keeping the materials of the instrument the same, the natural resonances decrease in frequency by a factor of $1/a$. The shape of the spectral envelope is preserved under this transformation, and so, if the spectral envelope is plotted on a log-frequency axis, the envelope shifts as a unit toward the origin, without changing shape, and the change in S_f will be the logarithm of the relative size of the two instruments: $\log(1/a)$. This uniform scaling relationship is called "the general law of similarity of acoustic systems" (Fletcher and Rossing 1998), and it is used to produce much of the difference in S_f between the tones produced by different instruments within a family. Numerical examples illustrating how the spatial dimensions of an instrument affect its resonances are provided by van Dinther and Patterson (2006).

Comparison of the filter systems of the different instrument families shows that the spectral envelope is affected in the same way by changes in the size of the filter-system components; specifically, the resonant frequencies *decrease* as body size *increases* and so the spectral envelope shifts toward the origin as the sizes of the components increase. So size affects the filter system in the same way as it affects the excitation mechanism. It is another example of the fact that bigger things vibrate more slowly. The wood-plate and bridge resonances of the string-family filter system are complex, and they are fundamentally different from the bell and mouthpiece resonances of the brass-family filter system, which are also complex.

Despite the complexity of the relationship between the physical variables involved in body filtering and the shape of the resultant spectral envelope, the relationship between the acoustic properties and the perception of the notes is fairly straightforward. The shape of the spectral envelope determines the family aspect of timbre; the acoustic scale of the filter, S_f, determines the register we perceive, and thus, which instrument within the family. In all of these instrument families, the register decreases from soprano to bass as instrument size increases and the spectral envelope shifts toward the origin.

2.3.3   Constraints on the Acoustic-Scale Variables in Orchestral Instruments

In Sects. 2.3.1 and 2.3.2, the relationship between the physical variables involved in the production of musical tones, and the acoustic scale of the source, S_s, and the filter, S_f, was presented in theoretical terms without reference to the practicalities of constructing and playing instruments. In the real world, it turns out that it is not possible to simply scale the spatial dimensions of instruments to achieve registers ranging from soprano to bass in most instrument families; the bass member would be too large and/or the soprano member too small. This section reviews the spatial scaling problem, and describes how the instrument makers produce tones with a wide range of acoustic scale values without using excessively large or small instruments.

   The spatial scaling problem arises from the desire to simultaneously satisfy three design criteria for families of sustained-tone instruments: The first criterion is that the instruments should produce notes that are heard to have a strong musical pitch, whose clarity and salience provide for effortless communication of melodies and their variations. This places an important constraint on the relationship between the acoustic scale variables, S_s and S_f. The instrument's filter system must resonate at frequencies corresponding to the first 10 harmonics of the pulse rate of each note that the instrument is intended to play; that is, the instrument must emit significant amounts of acoustic energy in the range from the pulse rate of each note to three octaves above that pulse rate. This is necessary because the pitch of notes where the energy is carried by harmonics above about the tenth is not sufficiently salient to support accurate perception of novel melodies (Krumbholz et al. 2000; Pressnitzer et al. 2001). The second criterion is that the members of each instrument family should, together, produce notes that cover a significant portion of the musical scale, which for the keyboard encompasses about seven octaves from, say, 27.5–3520 cps. When combined with the first criterion, the second criterion effectively requires that the instruments of a given family have matched S_s and S_f values for all of the registers in the range from soprano to bass. This is a very demanding constraint, particularly when combined with the third criterion, which is that the instruments should be playable and portable. This last, practical constraint places limitations on the sizes of instruments which, in turn, means that the desired range of notes cannot be achieved by simply scaling instrument size in accordance with the law of acoustic similarity.

There are problems for the instrument maker at both ends of the register range. For example, in the string family, there is a limit to how short the neck can be on the smallest member of the family (the violin) if the contact points where the string is pressed onto the neck are to be far enough apart for a musician to play the notes of a melody accurately and quickly. And at the other end of the range, if the instrument maker attempts to scale up the soprano version of the family to provide the bass member, the instruments become too large to play *and* too large to carry. Hutchins (1967, 1980) described the problems encountered when you try to construct a family of eight stringed instruments covering the entire range of orchestral registers based on the properties of the violin. The double bass member of the family would have to be six times the size of the violin, if simple scaling of instrument dimensions were to be used to provide a shift of six octaves in the spectral envelope. The length of a violin is about 0.6 m, so the double bass in this hypothetical family would have to be 3.6 m tall. The lower notes on the strings of such a double bass would not be reachable for most musicians and the instrument would not be portable. So, the problem is this: Although instrument makers can scale the dimensions of instruments to achieve much of the desired change in S_s and S_f, it is not possible to use the scaling of spatial dimensions, on its own, to provide the full range of registers in each family, and at the same time, ensure that the pitch of each note is sufficiently strong to support accurate melody perception.

So how do instrument makers solve this problem, and how do they construct families of instruments that produce tones with salient pitches over the full range of registers from soprano to bass – instruments that are, at the same time, playable and portable? The first criterion of instrument production is immutable; the instrument must produce energy in the first three octaves of the pulse rate if the note is to have a well defined pitch. The third criterion is essential; the instruments have to be playable and portable. So how do the instrument makers provide a wide range of notes on instruments with manageable sizes? This is where the knowledge and craft of the instrument maker come to the fore. What is required is not that the soprano instruments be excessively small and the bass instruments be excessively large; what matters is that the instruments produce tones with a wide range of S_s and S_f values, and that the S_s and S_f values are coordinated throughout the range. So what the instrument makers have done is find ways of extending the range of S_s and S_f values beyond what is practical with spatial-dimension scaling, by adjusting other physical properties of the instruments such as the mass of the strings, the thickness of the plates or the depth of the volume of the air cavity. They scale the physical dimensions of the family so that the largest member is portable and the smallest member is playable, and then they adjust other physical properties of the instrument to achieve the desired acoustic scale values for the source mechanism and the filter system (e.g., Schelleng 1963).

Consider the case of the source scale in the string family: The strings on the larger members such as the cello and contra bass are not as long as the law of acoustic similarity would require because it would make the instruments unwieldy. The instrument makers increase the linear mass of the strings (the mass per meter)

by winding metal coils around the string. This increased mass causes the strings to vibrate more slowly as illustrated by Eq. (2.2). The instrument makers use a change in mass to obtain the lower ranges of notes on the lower strings of any given member of the family.

With regard to the filter scale in the string family: The filter systems of the larger members of the family are not as large as the law of acoustic similarity would require, because it would make the instruments too heavy and too large. The instrument makers adapt the characteristics of the instruments to preserve the sound quality while making them usable at the same time. The main resonance is driven by the cavity mode of the body which functions like a Helmoltz resonator. The volume of the instrument as well as the surface area of the *f*-holes are the key parameters. The open strings of the cello are tuned to pulse rates three times lower than those of the violin. However, the plates of the cello's body are only 2.1 times larger than those of the violin (Schelleng 1963), while the rib height of the cello is about four times that of the violin (Fletcher and Rossing 1998). Thus the volume of the cello is 17 times larger than that of the violin; this is equivalent to uniform spatial scaling by a factor of 2.6. To lower the body resonances to the desired values, the instrument makers vary the mass, thickness and arching of the body plates. Specifically, the body plate of the cello is made proportionally thinner than that of the violin which lowers the body resonance frequency (e.g., Molin et al. 1988).

Having established that the acoustic scale variables are balanced in the sustained-tone instruments of the orchestra, we can return to the secondary aspect of register, associated with the perception of tones from a single instrument, that is, the within-instrument register. Register, in this sense, is "a part of an [instrument's range] having a distinctive tonal quality" (Kennedy 1985, p. 585). So we speak of the chest and head registers of an individual's voice, or the upper and lower register of an instrument's range. In acoustic scale terms, the perception of register within an instrument's range, is a perceptual distinction concerning the relative values of S_s and S_f. When the S_s values of a succession of notes are high relative to the S_f of the singer or the instrument, we perceive that the person is singing, or the instrument is playing, in the upper register, and vice verse.

Finally, note that that the range of tones covered by the registers of the voice, from soprano to bass, is only about four octaves in total (from about C6 down to slightly over C2). The range of the string-family instruments (taken together) covers almost seven octaves (from just under C8 to just over C1). The singing teacher can help a vocalist strengthen tones toward the ends of their natural range, but they cannot stretch the vocal tract length or add significant mass to the vocal folds.

In summary:

1. Although the physics of the source mechanisms that excite the sustained-tone instruments is complicated, and the physics varies markedly from family to family, the acoustic scale of the source, S_s, provides a convenient summary of the action of the source as it pertains to tone perception. The source determines the repetition rate of the wave, or the position of the fine structure of the magnitude spectrum (on

a log frequency axis), and this, in turn, determines the pitch of the tone, and contributes to the perception of an instrument's register within its family.

2. Although the physics of the resonance mechanisms that filter the source waves is complicated, and the physics varies markedly from family to family, the acoustic scale of the filter, S_f, provides a convenient summary of the action of the filter with regard to its contribution to the perception of an instrument's register within its family.

3. Within a family, when source size is increased to increase the acoustic scale of the tones and lower the pitch, the acoustic scale of the filter has to be increased to maintain the distinctive timbre of the family, and to ensure that the tones continue to produce a strong pitch. At the same time, the increase in filter scale contributes to the lowering of the perception of the register of the instrument within its family.

4. Within a family, it is not possible to produce tones whose pitches span the entire range of the keyboard simply by varying the spatial dimensions of the source and the filter. To achieve the desired acoustic scale values, and the appropriate balance between the acoustic scale values, the instrument maker has to vary other physical properties like the mass of the strings and the stiffness of the plates.

2.4 The Auditory Representation of Pulse-Resonance Sounds and Acoustic Scale

This section presents a brief description of a time-domain model of auditory perception to show how the auditory system constructs our internal representation of musical tones, and to illustrate how the acoustic scale variables appear in this representation of sound. The internal representation is referred to as an auditory image and the stages of the auditory model are intended to simulate all of the auditory processing required to transform a sound into our initial perception of that sound (Patterson et al. 1992, 1995). The processes are analogous to those that the visual system uses to convert light entering the eye into an initial visual image of that light. Although the algorithms used to simulate the construction of the auditory image are straightforward in signal processing terms, auditory models are not commonly used to explain the perception of tones in music and speech research. The most common representation of sound in these research communities is the spectrogram, which is a temporally ordered sequence of magnitude spectra. The spectrogram is a linear-time, *linear*-frequency representation of sound, and it is normally plotted with time on the abscissa (*x*-axis) and frequency on the ordinate (*y*- axis) so that time progresses from left to right as the sound progresses. This section begins with a comparison of two auditory images (shown in Fig. 2.2) that illustrate the essentials of the auditory image as it pertains to the perception of musical tones, and how this representation of sound differs from the spectrogram.

2.4.1 Auditory Images

There are now a number of time-domain models of auditory processing that attempt to simulate the neural response to complex sounds like musical notes at a succession of stages in the auditory pathway, and which produce representations of sound that might be regarded as auditory images (e.g., Slaney and Lyon 1990; Meddis and Hewitt 1991; see de Cheveigné 2005 for a review). In these models, the auditory image is typically constructed in four stages that respectively simulate the operation of (1) the outer and middle ear, (2) the basilar partition, (3) the inner hair cells along the basilar partition, and (4) the temporal integration mechanism in the mid-brain. The Auditory Image Model (AIM; Patterson et al. 1992, 1995) is used to illustrate the construction of auditory images and the form of acoustic scale information in the image, as we currently understand it. What differs from one time-domain model to another is the degree to which they attempt to simulate the details of auditory processing in each stage, and the theoretical bases for the mechanisms chosen to represent these auditory processes. The differences are not particularly important for present purposes because the section is just intended to illustrate the general form of the internal representation of sound and the form of the acoustic scale variables in the internal representation.

The auditory image of a baritone singing the vowel /a/ on the note G2 is presented in Fig. 2.2a, and for comparison, the auditory image of a French horn playing the same note is shown in Fig. 2.2b; the figure is reproduced from van Dinther and Patterson (2006), which provides a more detailed description of the image construction process. The auditory images are the large, two-dimensional, "waterfall" plots; the dimensions of the auditory image are time-interval on the abscissa (from 1 to 35 ms *increasing* toward the left) and frequency on the ordinate (from 0.1 to 6.0 kHz). The properties of the auditory image are introduced with
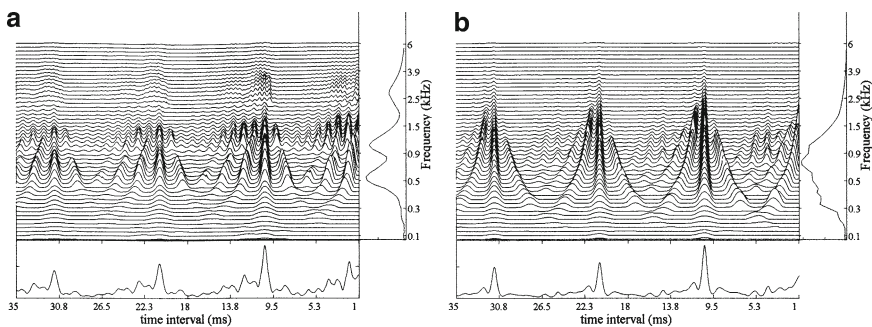

Fig. 2.2 Auditory images of the note G2 (198 cps) as sung by a baritone (**a**) and as played by a French horn (**b**). The waterfall plot represents the strobe-stabilized neural activity as a function of time interval since the last strobe time in each frequency channel (see text and Fig. 2.3 for details). The lower profile on each panel is the summary temporal profile. The peaks in this profile show the repetition rate of the sound. The height of the peaks relative to the baseline represents pitch strength (From van Dinther and Patterson 2006)

reference to the four stages of processing used to construct them, and the aspect, or aspects of the auditory image that each stage of processing imparts to the image. The vertical profiles to the right of each image, and the horizontal profile below each image, are introduced once the description of the auditory image itself is complete. (There are multipanel figures in van Dinther and Patterson (2006) that show how the auditory images in Fig. 2.2 were constructed, and how they change as the acoustic scale of the source and the acoustic scale of the filter vary.)

The first stage of processing simulates the effect of the outer and middle ear on incoming sound as it travels from the air through to the cochlea. It is these structures that determine the lower and upper frequency limits for human hearing in young normal listeners. Accordingly, from the perspective of music perception, the first stage determines the range of frequencies that young people normally hear, which is from about 0.1–12 kHz. The vertical dimension of the auditory image is the frequency dimension and so the first stage of processing determines the upper and lower bounds of the auditory image and how activity dies away as it approaches the edges of the image. In AIM, the weighting function is based on the loudness model of Glasberg and Moore (2002). In the case of speech and music, there is very little energy in the region above about 6 kHz, and what is there has very little effect on our perception of musical tones and speech sounds, so the plot of the auditory image is normally limited to 6 kHz as in the images presented in Fig. 2.2.

The second stage of processing simulates the spectral analysis performed in the cochlea by the basilar membrane in conjunction with the outer hair cells and the tectorial membrane; these structures are collectively referred to as the "basilar partition." The spectral analysis creates the tonotopic dimension along the basilar partition, and it creates the acoustic frequency dimension of auditory perception shown as the vertical dimension in the auditory image. In AIM, as in most time-domain models of perception, the spectral analysis is simulated with a bank of "auditory filters." Each filter creates a "frequency channel" in the auditory image; that is, the filter passes acoustic energy in a small frequency region about its "center frequency," and outside this "pass-band," the filter progressively attenuates acoustic energy as the frequency of that energy diverges from the centre frequency of the filter. This is the essence of an auditory filter. The width of the pass-band of the auditory filter increases with its centre frequency, and the spacing of the filters along the frequency dimension increases with center frequency. As a result, the tonotopic dimension of the cochlea is a quasi-logarithmic frequency axis as shown in the auditory images of Fig. 2.2. In the current version of AIM, the auditory filter is the compressive, gammachirp auditory filter (Irino and Patterson 2001; Patterson et al. 2003).

Each of the lines in the auditory image shows the recent history of activity in a specific frequency channel; the vertical position of the low-level activity in the channel shows the center frequency of each filter. The activity in adjacent channels is correlated and, as a result, the set of filter outputs gives the visual impression of a surface in auditory image space. The surface is AIM's simulation of the internal representation of sound that is assumed to be the basis of one's initial perception of a sound. The tones of music produce distinctive structures in the auditory image as illustrated in Fig. 2.2; the structures are referred to as "auditory figures" because

they stand out like figures when presented in background noise (Patterson et al. 1992). The tonotopic dimension of the auditory image is similar to the frequency dimension in Fig. 2.1b insofar as it is quasi-logarithmic; it differs from the strictly logarithmic frequency dimension of Fig. 2.1b inasmuch as the density of channels decreases somewhat below about 0.5 kHz (e.g., Moore and Glasberg 1983).

In the current version of AIM, the auditory filter is the *compressive*, *gammachirp* auditory filter (Irino and Patterson 2001; Patterson et al. 2003). For readers with an interest in the details, the gammachirp auditory filter is a development of the gammatone auditory filter (Patterson et al. 1995; Unoki et al. 2006). The gammatone filter is essentially symmetric and it is linear, that is, it does not change shape with stimulus level. The gammachirp auditory filter is asymmetric and the asymmetry varies with stimulus level, as dictated by human masking data (Unoki et al. 2006). In the dynamic version of this gammachirp filter (Irino and Patterson 2006), a form of fast-acting compression is incorporated into the auditory filter itself. The compression responds to level changes *within* the individual cycles of pulse-resonance sounds and, as a result, the filter restricts the amplitude of the pulse and amplifies the resonance relative to the pulse in each cycle (see Irino and Patterson 2006, their Figs. 7 and 9).

The third stage of processing simulates neural transduction, that is, the conversion of basilar partition motion into neural activity in the cochlea at the input to the auditory nerve. In AIM, neural transduction is assumed to take place separately in each frequency channel. Specifically, the amplitude versus time wave that flows out of each auditory filter is (1) half-wave rectified (that is, the negative values are set to 0) and (2) low-pass filtered to simulate the upper limit on the firing rate of auditory nerve fibers. The result is a simulation of the aggregate firing of all of the primary auditory nerve fibers associated with that region of the basilar membrane (Patterson 1994a); this function is referred to as a neural activity pattern (NAP). The rapidly oscillating function in Fig. 2.3 shows the NAP flowing from a single auditory filter in response to an /a/ vowel with a GPR of 116 cps and a period of 8.6 ms. The auditory filter is centered just above 1.0 kHz, so the individual cycles of the NAP are just under 1 ms in duration. Each cycle of the vowel produces a distinct cycle of activity in the NAP. There is one of these NAP functions for each of the filters in the filterbank, and together they simulate the response of the cochlea to the vowel.

The fourth stage simulates auditory temporal integration and it converts the set of NAP functions flowing from the auditory filterbank into AIM's simulation of our auditory image of a sound, that is, the neural representation that forms the basis of what we perceive when presented with a sound. In auditory models, this fourth stage of processing is currently hypothetical, in the sense that we do not know precisely how or where it is performed. The reason why perceptual models require a fourth stage is because the time scale of level variation in the NAP functions is not compatible with our perception of sounds; it is clear that there must be some form of temporal integration in the system prior to the neural representation that is the basis of perception. Consider the NAP function in Fig. 2.3: It shows the response to a little over three cycles of the vowel (a total duration of only 0.03 s), so a 1-s segment of the vowel with 116 cycles would be about 30 times the length of the
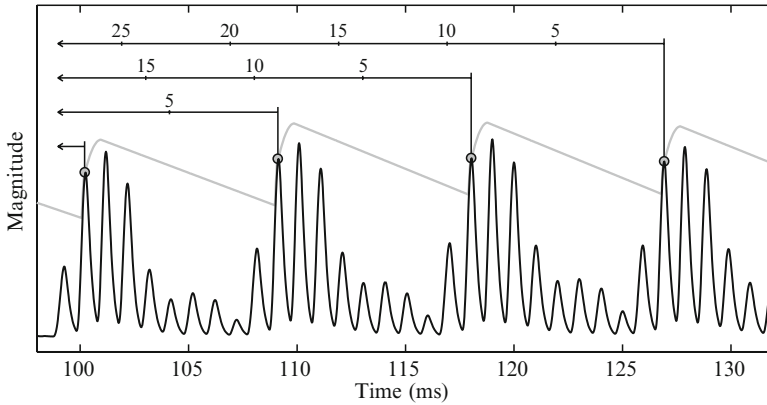
Fig. 2.3 Detail of the neural activity pattern produced by an /a/ vowel at the output of the auditory filter centered at 1018 Hz. The gray line shows the adaptive threshold used to calculate the grey dots, which show the strobe points. The vertical lines and backward arrows show how time intervals are calculated from each of the strobe points backwards in time to earlier points in the pattern, and generate the NAP segment that is added into the corresponding channel of the auditory image to produce the stabilized version presented in Fig. 2.2

segment shown in Fig. 2.3. If Fig. 2.3 were a real-time display (like the neural representation that we perceive), these 30 cycles of the NAP would flow very rapidly from right to left across the display in the course of 1 s, and it would just be a blur. So if the NAP functions were the basis of perception, we would not be able to use the fine-grain temporal information in the NAP functions. However, perceptual research on pitch and timbre indicates that at least some of the fine-grain, time-interval information in the NAP functions is preserved in the auditory image (e.g., Patterson 1994a, b; Yost et al. 1998; Krumbholz et al. 2003). This means that the temporal integration process used to construct the auditory image cannot be simulated by a running temporal average process, like that used to construct the spectrogram because an averaging process would blur the temporal fine structure within the averaging window (Patterson et al. 1992, 1995).

Patterson et al. (1992) argued that it is the fine-structure of periodic sounds that is preserved rather than the fine-structure of aperiodic sounds (e.g., noises), and they showed that the fine-structure of periodic sounds could be preserved by a form of "strobed temporal integration" controlled by an adaptive threshold. The adaptive threshold for the vowel NAP in Fig. 2.3 is shown by the line with gray dots above the NAP function. It is a form of temporal envelope which emphasizes where the individual cycles of the NAP function start (the dots). These strobe points are used to direct the temporal integration process as indicated by the vertical lines and horizontal arrows above each strobe point. As the start of each new cycle of the NAP function is identified (the dots), a section of the NAP function from the strobe point back to 35 ms before the strobe point (the horizontal lines), is copied and added as a unit into the corresponding channel of the auditory image. In the process the strobe *time* in the NAP function is subtracted from absolute time in the NAP and so, in the

auditory image, the activity associated with any given strobe extends from 0 ms in the auditory image (Fig. 2.2), backwards for 35 ms. As the activity in successive cycles is very similar for pulse-resonance sounds, successive cycles sum to produce a stabilized representation of the pattern in the NAP.

The set of all image channels (one for each filterbank channel) is AIM's representation of our internal auditory image, and the auditory images in Fig. 2.2 were constructed in this way (Patterson et al. 1992, 1995; Patterson 1994a, b). The image decays fairly slowly with respect to the rate of cycles in pulse-resonance sounds (specifically with a half life of 30 ms). So a stabilized version of the neural pattern within the cycle of the sound builds up in the auditory image when the sound comes on and stays there as long as the sound is stationary. When the sound goes off, it decays away to nothing in about 100 ms.

More detailed descriptions of auditory image construction are presented in Patterson et al. (1995), van Dinther and Patterson (2006), and Ives and Patterson (2008). The auditory image is similar in form to the autocorrelogram (Slaney and Lyon 1990; Meddis and Hewitt 1991; Yost 1996) but the construction of the auditory image is more efficient and it preserves the temporal asymmetry of pulse-resonance sounds. The similarities and differences between auditory images and autocorrelograms are described in Patterson and Irino (1998).

2.4.2   The Spectral Profile and S_f

While the processing of pulse-resonance sounds up to the level of our initial perception of them may seem complicated, the relationship between the acoustic properties of these sounds, as observed in their waves and log-frequency spectra, and the features that appear in the auditory images of pulse-resonance sounds is relatively straightforward. In Fig. 2.2, the spectral profile to the right of each auditory image is the average of the activity in the image across time intervals; it simulates the tonotopic distribution of activity observed in the cochlea and in neural centers of the auditory pathway up to auditory cortex. The frequency axis is quasi-logarithmic like the tonotopic dimension of the cochlea (Moore and Glasberg 1983). The spectral profile of the auditory image is similar to the "excitation pattern" described by, for example, Zwicker (1974) and Glasberg and Moore (1990), inasmuch as they all simulate the distribution of activity along the tonotopic axis in the auditory system with a compressed measure of magnitude.

The three peaks in the spectral profile of the /a/ (G2) of the baritone in Fig. 2.2a show the formants of this vowel. Note, that the profile from AIM is similar to the envelope of the magnitude spectrum of the child's vowel, shown in Fig. 2.1b, except that the pattern in Fig. 2.2a is shifted toward the origin with respect to that in Fig. 2.1b because in Fig. 2.2a, the singer is an adult.

The spectral profile of the auditory image is similar in form to the envelope of the magnitude spectrum. Both are *covariant* representations of family and register information (van Dinther and Patterson 2006); the family information is contained in the

shape of the envelope, and the register information is in the position of the envelope, S_f, along the frequency axis. Comparison of the spectral profiles of the auditory images in Fig. 2.2a and b show that, whereas the spectral envelope of the voice is characterized by three distinct peaks, or formants, the envelope of the French horn is characterized by one broad region of activity.

2.4.3   The Time-Interval Profile and S_s

The resolution of the auditory filter, at the sound levels where we normally listen to music, is not sufficient to define individual harmonics of pulse-resonance sounds beyond the first three or four harmonics (e.g., Ives and Patterson 2008). As a result, the fine structure of the magnitude spectrum and S_s are not readily apparent in the spectral profile of the auditory image for musical sounds. However, the S_s information is present in the auditory image, in the form of the vertical ridge in the 10-ms region of the image. The ridge shows that there is a concentration of activity at the period of the tone in most channels of the auditory images in Fig. 2.2a and b. Thus, the acoustic scale of the source is readily observed in this simulation of the neural representation of sound, even though the construction of the auditory image includes a temporal integration process with a half life of 30 ms. This is because strobed temporal integration preserves the temporal fine structure of periodic components of sounds like the sustained parts of vowels and musical notes.

Moreover, the temporal information associated with the acoustic scale of the source is enhanced in the time-interval profile of the auditory image. This profile appears below the auditory image and shows the activity averaged across filter channels. In this time-interval profile, the position of the largest peak (in the region to the left of 1.25 ms) provides an accurate estimate of the period of the sound (for G2, 10.2 ms). Moreover, the height of the peak, relative to the level of the background at the foot of the peak, provides a good measure of the salience of the pitch percept (Yost et al. 1996; Patterson et al. 2000; Ives and Patterson 2008). Thus, in time-domain models involving auditory images, the most obvious correlate of the acoustic scale of the source, S_s, in an instrument is a concentration of time intervals at a particular value in the temporal profile. This form of S_s information is more like the time between peaks in the sound wave (Fig. 2.2a) rather than the position of the fine structure in the magnitude spectrum of the sound (Fig. 2.2b).

2.4.4   Summary of Auditory Image Construction and the Acoustic Scale Information in the Image

In auditory models of perception, the auditory image that simulates the neural substrate of perception is typically constructed in four stages: A spectral weighting function, similar to the audiogram in form, simulates the middle-ear filtering that

limits sensitivity to very high and very low frequencies. An auditory filterbank simulates the spectral analysis performed in the cochlea. Neural transduction is simulated with half-wave rectification and low-pass filtering. A sophisticated form of temporal integration stabilizes the repeating neural patterns produced by pulse-resonance sounds and completes the construction of the auditory image.

The main vertical ridge in the auditory image, and the corresponding peak in the time-interval profile, are the auditory model's representation of the acoustic scale of the source, S_s. They move left to longer time intervals as the pulse rate of the sound decreases, and to the right to shorter time intervals as the pulse rate increases. When this S_s marker stands out clearly in the time-interval profile well above the background activity, the sound is effectively periodic and the tone is heard to have a strong pitch. When the scale of the filter, S_f, changes, the complex pattern in the auditory image simply moves up or down in frequency *without changing shape*. Similarly, the distribution of activity in the spectral profile of the image moves up or down without changing shape.

2.5 The Acoustic Properties of Pulse-Resonance Sounds and the Auditory Variables of Perception

The final section of this chapter reviews the relationship between the acoustic properties of sound and three variables of auditory perception, loudness, pitch, and timbre, to illustrate how they relate to the variables of music perception described in the sections above, namely, melody, instrument family and register within a family. The American National Standards Institute (ANSI) has provided official definitions of loudness, pitch, and timbre, and these definitions are widely quoted. This section begins with the definitions as they appear in ANSI (1994), as they might have been expected to specify just those relationships between physical and perceptual variables that we require to explain the perception of musical notes. The definitions are:

12.03 Loudness. That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud.

12.01 Pitch. That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch depends primarily on the frequency content of the sound stimulus, but it also depends on the sound pressure and the waveform of the stimulus. Note: The pitch of a sound may be described by the frequency or frequency level of that simple tone having a specified sound pressure level that is judged by listeners to produce the same pitch.

12.09 Timbre. That attribute of auditory sensation that enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are dissimilar. Note: Timbre depends primarily upon the frequency spectrum, although it also depends upon the sound pressure and the temporal characteristics of the sound.

These definitions are useful, inasmuch as they illustrate the desire to relate properties of perception to physical properties of sound, and they illustrate what is regarded by auditory scientists as a principled way of proceeding with this task. Unfortunately, the definitions focus on the perceptual properties without, in the end, specifying the relationship of each to the corresponding, acoustic, or physical variables, other than to say that both pitch and timbre *depend primarily on the frequency content of the sound*. Although true, this is not very helpful because it does not say which aspect of the frequency information is associated with pitch and which aspect is associated with timbre. The discussion of acoustic scale in Sect. 2.2 suggests that, for musical sounds at least, we can be more specific about the relationship between the acoustic properties of sound and the perceptions associated with musical notes and instruments. In particular, S_s, the position of the fine structure of the magnitude spectrum, largely determines the pitch of a musical note, and a melody is an ordered sequence of S_s values. The shape of the spectral envelope is closely associated with the perception of instrument family, or the family aspect of timbre. So it is envelope shape that supports the general distinction between, for example, brass and string instruments. And, S_f, the position of the envelope of the magnitude spectrum, combines with S_s to determine the register of the instrument within a family. The acoustic scale variables S_s and S_f are also prime determinants of our perception of the size of an instrument or the height of a singer. In this final section of the chapter, we review the relationship between these acoustic properties of sound and the traditional auditory variables, pitch and timbre, with a view to developing a more useful description of the mapping between the acoustic and auditory variables as they pertain to music perception.

2.5.1 Effect of Source Size on Pitch and Timbre

Consider the definitions of pitch and timbre, and the question of how we perceive the physical changes that take place in a vowel as a child grows up, or how we perceive the physical changes that take place in a musical note as it is played on successively larger members of an instrument family, for example, when a trumpet, trombone, and tuba play C3, one after another. The logic of the ANSI definition of timbre is not entirely clear, but it would appear to involve a process of elimination, in which variables of auditory perception that *do not* affect timbre are identified and separated from the remaining variables, which by default are part of timbre. The perceptual variables of particular interest are duration, loudness and pitch.

Duration is the variable that is most obviously separable from timbre, and it illustrates the logic underlying the definition of timbre (although there is not actually a standard definition of the perception of duration). If a singer holds a note for a longer rather than a shorter period, it produces a discriminable change in the sound but it is not a change in timbre. Duration has no effect on the magnitude spectrum of a sound, once the duration is well beyond that of the temporal window used to produce the magnitude spectrum. The sustained notes of music are typically longer

than 200 ms in duration, and the window used to produce the magnitude spectrum is usually less than 100 ms, so duration is unlikely to play a significant role in family timbre or register timbre. In general, then, the perceptual change associated with a change in the duration of a sustained note is separable from changes in the timbre of the note.

Loudness is also largely separable from timbre. If we turn up the volume control when playing a recording, the change will be perceived predominantly as an increase in loudness. The pitch of any given vowel and the timbre of that vowel will be essentially unaffected by the manipulation. The increase in the intensity of the sound produces a change in the magnitude spectrum of the vowel – both the fine structure and the envelope shift vertically upwards – but there is no change in the frequencies of the components of the fine structure and there is no change in the relative amplitudes of the harmonics. Nor is there any change in the shape of the spectral envelope. So, loudness is also separable from timbre.

Thus, acoustic variables that do not affect either the shape of the envelope of the magnitude spectrum or the frequencies of the spectral components do not affect the timbre of the sound. The question is: "What happens when a simple shift is applied to the position of the fine structure, or to the position of the envelope, of a sound (on a log-frequency axis), that is, when we change S_s, S_f, or both?" The current definition of timbre suggests that a change in S_s, which is heard as a change in pitch, does not affect the timbre of the sound, whereas a change in S_f, which is heard as a change in speaker size or instrument size, does affect the timbre of the sound. This is where the current definition of timbre becomes problematic, that is, when it treats the two aspects of acoustic scale differently with regard to their role in the perception of timbre.

Note, in passing, that shifting the position of the fine structure of the magnitude spectrum, while holding the envelope fixed, produces large changes in the relative magnitudes of the harmonics as they move through the region of a formant peak. So the relative magnitudes of the components in the spectrum can change substantially without producing a change in timbre, by the current definition. Note, also, that shifting the envelope of the magnitude spectrum while holding the position of the fine-structure constant produces similar changes in the relative magnitudes of the component frequencies as they move through formant regions. Such shifts do not change the timbre category of a musical sound (the family timbre); they change the apparent size of the source, and if the change is large enough they change the perceived register of the instrument, which, of course, is a timbre change, by the current definition.

2.5.2 Acoustic Scale "Melodies" and the Perception of Pitch and Timbre

The discussion focuses on a set of four melodies designed to emphasize the role of the acoustic scale variables in the perception of vocal pitch and timbre. The novel

Fig. 2.4 Musical notation for four short melodies. The black notes show the acoustic scale of the source, S_s, and thus, the melodic pitch during the course of the musical sequence. The gray, flipped, notes represent the acoustic scale of the filter, S_f, on a musical scale. The original speaker's voice defines the note E (the bottom line on the staves) for both acoustic scales

aspect of the melodies is that, in some cases, the acoustic scale of the filter, S_f, varies over the course of the melody, either on its own, or in conjunction with changes in S_s. The scale of the filter is normally fixed when an instrument plays a melody. A form of musical notation for the melodies is presented in Fig. 2.4; it shows that the melodies all have four bars containing a total of eight notes. The melodies are in ¾ time, with the fourth and eight notes extended to give the sequence a musical feel. The melodies have a "phonological text," that is, the notes are sung as syllables (*pi*, *pe*, *ko*, *kuuu*; *ni*, *ne*, *mo*, *muuu*), which emphasizes the human quality of the voice. As the timbre changes from vowel to vowel, it engages the phonological system and allows us to distinguish the role of envelope shape in melody perception, from the role of S_s and the role of S_f. The phonological text is the same for all four melodies.

The syllables were originally sung by an adult male (author R. P.) who has an average GPR of about 120 cps and a vocal tract length of about 16.5 cm. STRAIGHT (Kawahara and Irino 2004) was used to vary the scale of the source, S_s and the scale of the filter, S_f, for each of the syllables, to simulate changes in the GPR and VTL of the singer. The matrix of tones used to produce the melodies is shown in Fig. 2.5. The abscissa of the matrix (*x*-axis) is the acoustic scale of the source, S_s, and it was varied to produce an octave of notes using the diatonic major scale of Western music. The ordinate of the matrix (*y*-axis) is the acoustic scale of the filter, S_f, and it was varied to simulate voices with an octave range of vocal tract

lengths ranging from about 10–20 cm. As with the S_s dimension, the specific values of S_f were determined by the diatonic major scale of Western music. In other words, the S_f ratio between any two notes has the same numerical value as the corresponding S_s ratio, and the values of the S_f ratios are indicated in musical notation by the note names, C, D, E, etc. The manipulation of S_f effectively extends the domain of notes from a diatonic musical scale to a diatonic musical plane as shown in Fig. 2.5.

The arrows in Fig. 2.5 show the sequences of notes in each melody. This alternative notation for the melodies illustrates the interaction of the acoustic scale variables. Returning to Fig. 2.4, for each melody, the black notes show the progression of intervals for S_s (or GPR) as each melody proceeds, and the grey notes show the progression of intervals for S_f (or VTL) as the melody proceeds. The sound files for the melodies are available at http://www.acousticscale.org/link/SHAR2010Demo. The shaded note (E, E) on the S_s–S_f plane provides the anchor for the notation; it has the same GPR and VTL values as the original syllables.
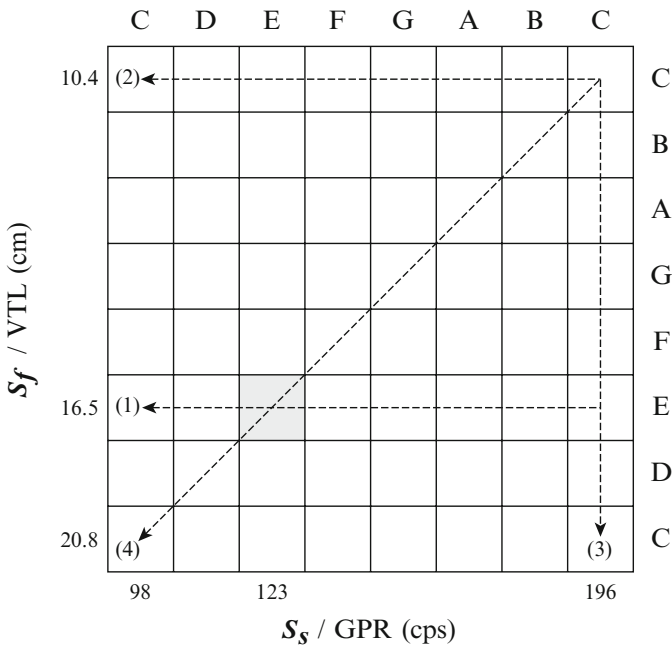

Fig. 2.5 The S_s-S_f plane, or GPR–VTL plane. The abscissa is the acoustic scale of the source S_s, increasing from left to right over an octave. The ordinate is the acoustic scale of the filter S_f, doubling from top to bottom. The plane is partitioned into squares that represent the musical intervals. The square associated with the original speaker is highlighted in gray. The dashed lines show the progression of notes in the four melodies of Fig. 2.4

2.5.2.1 Melody 1

The first melody simulates the normal situation wherein a singer with a fixed vocal tract length (VTL) varies the tension of the vocal cords to vary S_s in accordance with the black notes in Staff (1) of Fig. 2.4. The gray notes (for S_f) do not vary in this melody, indicating that the VTL of the singer is fixed. The VTL is relatively long, so the singer is heard to be an adult male. The pitch of the voice drops by an octave over the course of the melody from about 200 cps, which is well above the original pitch, down to about 100 cps, which is a few notes below the original pitch. This descending melody is within the normal range for a tenor, and the melody sounds natural. As the melody proceeds, the fine-structure of the spectrum, S_s, shifts, as a unit, with each change in GPR, and over the course of the melody, it shifts an octave towards the origin. The ANSI definition of timbre implies that these relatively large S_s changes, which produce large pitch changes, do *not* produce timbre changes, and this seems entirely compatible with what we hear in this melody. So, this melody illustrates the commonly held belief, embodied in the ANSI definitions, that pitch is largely separable from timbre, much as duration and loudness are.

2.5.2.2 Melody 2

Problems arise when we extend the example and synthesize a version of the same melody but with a singer that has a much shorter vocal tract, like that of a small child (Fig. 2.4, Staff [2]). There is no problem at the start of the melody; it just sounds like a child singing the melody. The starting pitch is low for the voice of a small child but not impossibly so. As the melody proceeds, however, the pitch decreases by a full octave, which takes it beyond the normal range for a child. As a result, in the latter part of the melody, we hear the voice quality change and, by the end of the melody, the child comes to sound rather more like a dwarf. The ANSI definition of timbre does not provide any basis for understanding the voice quality change from a child to a dwarf; within the traditional framework, the changes that we hear as the melody proceeds are just pitch changes. But traditionally, voice quality changes associated with a change in speaker changed are regarded as timbre changes. This is the first form of problem with the standard definition of timbre – changes that are nominally pitch changes producing what would normally be classified as a timbre change.

2.5.2.3 Melody 3

In the next example (Fig. 2.4, Staff [3]), the roles of the acoustic-scale variables, S_s and S_f, are reversed. The position of the fine structure, S_s, is held fixed while the position of the envelope, S_f, shifts by an octave toward the origin. The change in S_f

simulates a doubling of the VTL, from about 10–20 cm, which would normally be associated with a doubling of height. The S_f ratios between successive notes of the melody have the same numerical values as the S_s ratios of the first two melodies. As melody 3 proceeds and the envelope shifts down by an octave, the child seems to get ever larger, the voice comes to sound something like that of a counter tenor, that is, a tall person with an inordinately high pitch. The ANSI definition of timbre does not say anything specific about how changes in the position of the spectral envelope affect timbre or voice quality; the acoustic scale variable, S_f, was not recognized when these standards were written. Nevertheless, the definition gives the impression that any change in the spectrum that produces an audible change in the perception of the sound, without producing a change in duration, loudness or pitch, produces a change in timbre. Experiments with scaled vowels and syllables show that the just noticeable change in S_f is about 7% for vowels (Smith et al. 2005) and 5% for syllables (Ives et al. 2005), so the intervals in the melody would be expected to produce perceptible S_f changes. Because traditionally, voice quality changes are thought to be timbre changes, the fact that the singer at the start of the melody (a child) is different from the singer at the end of the melody (a counter tenor) seems compatible with the definition of timbre; the singer changes and the timbre changes. However, we are left with the problem that large changes in S_s and S_f both seem to produce changes in voice quality, but whereas the perceptual changes associated with large shifts of the fine-structure along the log-frequency axis are not timbre changes, the perceptual changes associated with large shifts of the envelope along the same log-frequency axis are timbre changes, according to the ANSI definition. They both produce changes in the relative amplitudes of the spectral components, but neither changes the shape of the envelope and neither form of shift alters the phonological values of the individual syllables.

2.5.2.4   Melody 4

The problems involved in attempting to unify the perception of voice quality with the definition of timbre become more complex when we consider melodies where both S_s and S_f change as the melody proceeds. Consider the melody produced by covarying S_s and S_f to produce the notes along the diagonal of the S_s–S_f plane (Fig. 2.5). The musical notation for the melody is shown in Fig. 2.4, Staff (4). This melody is perceived to descend an octave as the sequence proceeds, and there is a progressive increase in the perceived size of the singer from a child to an adult male (with one momentary reversal at the start of the second phrase). It is as if we had a set of singers varying in age from 4–18 in a row on stage, and we had them each sing their assigned syllable in order, and in time, to produce the melody. This melody, in combination with the others, makes it clear that there is an entire plane of singers with different vocal qualities defined by different combinations of the acoustic scale variables, S_s and S_f. The realization that there is a whole plane of voice qualities makes it clear just how difficult it would be to produce a clean

definition of timbre that excludes one of the acoustic scale variables, S_s, and not the other, S_f. If changes in voice quality are changes in timbre, then changes in pitch (S_s) can produce changes in timbre. This would seem to undermine the utility of the current definitions of pitch and timbre.

2.5.3   The Concept of Acoustic Scale and the Definitions of Pitch and Timbre

2.5.3.1   The "Second Dimension of Pitch" Hypothesis

At first glance, there would appear to be a fairly simple way to solve the problem; we could designate the perceptual dimension associated with the acoustic scale of the filter, S_f, to be a second dimension of pitch. Then, this second dimension of pitch could be excluded from the definition of timbre along with the first dimension of pitch. For the singing voice, manipulation of S_f on its own would sound like the change in perception that occurs over the course of melody 3, where S_s is fixed on the upper C and S_f decreases by a factor of two over the course of the melody. This does, however, lead to several problems. First, semitone changes in the scale of the filter, S_f, are barely large enough to hear differences in the associated perception so this second dimension of pitch would not support accurate perception of novel melodies, in the way that the first dimension of pitch does (e.g., Pressnitzer et al. 2001; Ives and Patterson 2008). The salience of changes in S_f is more like the salience of the weak S_s pitch that arises when the energy in a tone is restricted to high, unresolved harmonics, and pitch discrimination requires S_s changes of four semitones, or more. The second dimension of pitch would, in some sense, satisfy the ANSI definition of pitch which is not concerned with melodies, and which only requires that the attribute of auditory sensation can be used to order notes on a scale extending from low to high. It seems reasonable to say that the tones at the start of melody 3 sound "higher" than the tones at the end of the melody, which would support the "second dimension of pitch" hypothesis.

The "second dimension of pitch" hypothesis also leads to another problem. To determine the pitch of a sound, it is traditional to match the pitch of that sound to the pitch of either a sinusoid or a click train, that is, to a perception that is based on the scale of the source, S_s. Moreover, it seems likely that if listeners were asked to pitch match each of the notes in melody 3, among a larger set of sounds that diverted attention from the orderly progression of S_f in the melody, they would probably match all of the tones with the same sinusoid or the same click train, and the pitch of the matching stimulus (bound to an S_s value) would be the upper C. This would leave us with the problem that the second dimension of pitch, based on S_f, changes the perception of the sound but it does not change the pitch to which the sound is matched (its S_s value). So the "pitch" change associated with a change in S_f would have to be segregated from a normal pitch change and given a separate definition. It would also require changes in the ANSI definitions of pitch and timbre because currently, a change in perception (such as that associated with changes in S_f) that

does not produce a change in S_s pitch (or loudness, or duration) is a change in timbre. In short, the "second dimension of pitch" hypothesis would appear to lead us back to the position that changes in S_f produce changes in the timbre of the sound.

The "second dimension of pitch" hypothesis also implies that if we play a random sequence of notes on the musical plane of Fig. 2.5, the voice quality changes that we hear are all pitch changes, and they involve no change in timbre. This seems unreasonable when the acoustic scale changes are sufficiently large to produce a clear change in the perception of *who* is singing.

Finally, there is the problem that many people hear the perceptual change in melody 3 as a change in speaker size, and they hear a more pronounced change in speaker size when changes in S_f are combined with changes in S_s, as in melody 4. To ignore the perception of speaker size, is another problem inherent in the "second dimension of pitch" hypothesis; source size is an important aspect of perception, and pretending that changes in the perception of source size are just pitch changes seems like a fundamental mistake for a model of perception.

2.5.3.2   The Scale of the Filter, S_f, as a Dimension of Timbre

Rather than co-opting the acoustic scale of the filter, S_f, to be a second dimension of pitch, it would seem more reasonable to think of it as an internal dimension of timbre – a dimension of timbre that for voices is associated with vocal register, singer gender, and singer size. This, however, leads to a different problem which is, in some sense, the inverse of the "second dimension of pitch" problem. Once it is recognized that shifting the position of the fine structure of the spectrum is inherently similar to shifting the position of the envelope of the spectrum, and that the two position variables are different aspects of the same property of sound (acoustic scale), then it seems unreasonable to have one of these variables, S_f, within the realm of timbre and the other, S_s, outside the realm of timbre. For example, consider the issue of voice quality; both of the acoustic scale dimensions affect voice quality and they interact in the production of a specific voice quality (e.g., man, woman, child, dwarf, counter tenor). Moreover, the scale of the source, S_s, affects the perception of the singer's size, in a way that is similar to the perceptual effect of the scale of the filter, S_f (Smith and Patterson 2005). Thus, if we define the scale of the filter, S_f, to be a dimension of timbre, then we need to consider that the scale of the source, S_s, may also need to be a dimension of timbre. After all, large changes in S_s affect voice quality which is normally considered to be an aspect of timbre.

2.5.4   The Independence of Spectral Envelope Shape

There is one further aspect of the perception of these melodies that should be emphasized, which is that neither of the acoustic scale manipulations causes a

change in the perception of the phonology of the syllables; we always hear *pi*, *pe*, *ko*, *kuuu*; *ni*, *ne*, *mo*, *muuu*, independent of the VTL and GPR values of the singers. That is, the changes in timbre that give rise to the perception of a sequence of syllables are unaffected by changes in S_s and S_f, even when these scale changes are large (Ives et al. 2005; Smith et al. 2005). The changes in timbre that define the phonology are associated with changes in the shape of the envelope, as opposed to the position of the spectral envelope or the position of the spectral fine structure. Changes in the shape of the envelope produce changes in vowel type in speech and changes in instrument family in music. Changing the position of the envelope and changing the position of the fine structure both produce substantial changes in the relative amplitudes of the components of the magnitude spectrum, but they do not change the timbre category of these sounds; that is, they do not change the vowel type in speech or the instrument family in music.

2.5.5 Summary

The ANSI definitions of pitch and timbre are not much help in understanding the perception of musical tones, in the sense of understanding what gives rise to the perception of melody, instrument family, and register within a family. The ANSI definitions simply associate both pitch and timbre with unspecified aspects of the frequency content of a sound. In music and speech research, it is traditional to segregate one aspect of the frequency information, namely, F0 (the repetition rate of the sound), from the remainder of the information which is represented by the spectrogram. F0 is then associated with the pitch of the instrument or the pitch of the voice, in the same way that we have associated the scale of the source, S_s, with pitch. Thus, in music and speech research there is, at least, the segregation of the main determinant of pitch from the distribution of frequency information across the acoustic frequency dimension. The difference between these approaches and the acoustic-scale approach presented in this chapter is illustrated in Fig. 2.6. The upper row shows how the frequency information is (or is not) divided up in each case, and
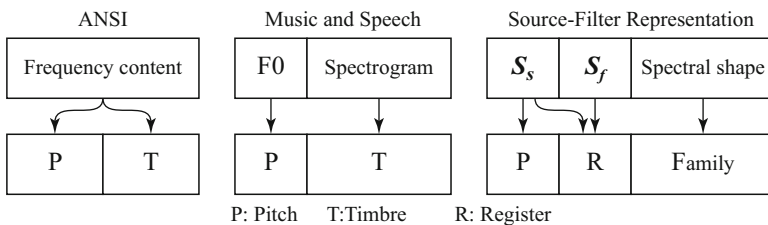

ANSI		Music and Speech		Source-Filter Representation		
Frequency content		F0	Spectrogram	S_s	S_f	Spectral shape
P	T	P	T	P	R	Family

P: Pitch   T:Timbre   R: Register

Fig. 2.6 The relationship between the acoustic variables (upper row) and the psychological variables related to the perception of a musical tone (lower row) as defined in the ANSI standard (ANSI 1994), in music and speech research, and using the acoustic scale variables

the lower row shows the components of auditory perception; the arrows indicate the associations between the components of the frequency information and the components of perception. In the first column, which corresponds to the ANSI definition, there is only one arrow associating all of the frequency content, indiscriminately, with both pitch and timbre. The second column, corresponding to music and speech research, shows how F0 is segregated from the spectrogram and associated with pitch.

The third column shows how the scale of the source, S_s, and the scale of the filter, S_f, are segregated from the shape of the envelope of the magnitude spectrum in the current approach. The scale of the source is directly related to musical pitch and melody. The shape of the envelope is directly related to the family aspect of timbre, and for the human voice, this is further subdivided into different vowel types. These aspects of the mapping from acoustic properties to perceptual variables are straightforward. The mapping between acoustic properties and register within a family is a little more complicated; both of the acoustic scale variables contribute to the perception of *register*. Both of the acoustic scale variables also contribute to the perception of instrument size and singer size, which are related perceptions in different contexts. It is also the case that the relative magnitude of the acoustic scale variables contributes to our perception of whether a specific instrument is a good, or bad, example of its class. Although the division of frequency information into three components, and the mapping from these components to the perception of musical tones, is somewhat more complicated than in traditional descriptions, it is not excessively complicated, and it does provide for a much better understanding of how the physical properties of instruments, and the acoustic properties of sound relate, to the auditory perceptions that musical tones produce.

2.6 Conclusions

Recent research on the role of acoustic scale in the perception of sound suggests that the frequency information observed in the magnitude spectrum of a sound is segregated by the auditory system into three parts: the spectral envelope shape, the acoustic scale of the source, S_s, and the acoustic scale of the filter, S_f. The spectral envelope shape determines the basic timbre category of a sound, which in music is the instrument family, and in the singing voice expands to produce the different vowel types. These timbre categories are largely independent of the acoustic scale variables, S_s and S_f. In speech, these two acoustic scale variables jointly determine much of the static voice quality of the speaker, and thus our perception of a speaker's sex and size (e.g., Smith and Patterson 2005). This suggests that it would be useful to distinguish between the "what" and "who" of timbre in speech, that is, what is being said, and who is saying it. With regard to the timbre of musical tones, the distinction between envelope shape and the acoustic scale variables provides an

explanation for the distinction between family timbre (envelope shape) and register timbre (S_s and S_f). In both speech and music, S_s exhibits a limited degree of independence from timbre inasmuch as (1) variation of GPR to produce prosodic distinctions does not change the perception of who is speaking, and (2) variation of the pulse rate in musical instruments to produce a melody does not change the perception of the instrument that is playing. There are, however, limits to the independence; large changes in pulse rate produce changes in the perception of who is speaking or which member of an instrument family is playing.

References

ANSI (1994) American National Standard Acoustical Terminology, ANSI S1.1–1994 (R1999). New York: American National Standard Institute.

Benade AH (1976) Fundamentals of Musical Acoustics. Oxford: Oxford University Press.

Benade AH, Lutgen SJ (1988) The saxophone spectrum. J Acoust Soc Am 83:1900–1907.

Chiba T, Kajiyama M (1941) The Vowel, its Nature and Structure. Tokyo: Tokyo-Kaiseikan.

Cohen L (1993) The scale representation. IEEE Trans Sig Proc 41:3275–3292.

de Cheveigné A (2005) Pitch perception models. In: Plack CJ, Oxenham AJ, Fay RR, Popper AN (eds), Pitch: Neural Coding and Perception. New York: Springer, pp. 169–233.

Fant G (1960) Acoustic Theory Of Speech Production. The Hague: Mouton De Gruyter.

Fitch WT, Giedd J (1999) Morphology and development of the human vocal tract: A study using magnetic resonance imaging. J Acoust Soc Am 106:1511–1522.

Fitch WT, Reby D (2001) The descended larynx is not uniquely human. Proc R Soc Lond Ser B 268:1669–1675.

Fletcher NH (1978) Mode locking in nonlinearly excited inharmonic musical oscillators. J Acoust Soc Am 64:1566–1569.

Fletcher NH, Rossing TD (1998) The Physics of Musical Instruments. New York: Springer.

Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47:103–138.

Glasberg BR, Moore BCJ (2002) A model of loudness applicable to time-varying sounds. J Audio Eng Soc 50:331–342.

Helmholtz HLF (1875) On the Sensations of Tone as a Physiological Basis for the Theory of Music. London: Longmans, Green.

Hutchins CM (1967) Founding a family of fiddles. Phys Today 20:23–37.

Hutchins CM (1980) The new violin family. In: Benade AH (ed), Sound Generation in Winds, Strings, Computers. Stockholm: The Royal Swedish Academy of Music, pp. 182–203.

Irino T, Patterson RD (2001) A compressive gammachirp auditory filter for both physiological and psychophysical data. J Acoust Soc Am 109:2008–2022.

Irino T, Patterson RD (2006) A dynamic compressive gammachirp auditory filterbank. IEEE Trans Audio Speech Lang Processing 14:2222–2232.

Ives DT, Patterson RD (2008) Pitch strength decreases as F0 and harmonic resolution increase in complex tones composed exclusively of high harmonics. J Acoust Soc Am 123:2670–2679.

Ives DT, Smith DRR, Patterson RD (2005) Discrimination of speaker size from syllable phrases. J Acoust Soc Am 118:3186–3822.

Kawahara H, Irino T (2004) Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In: Divenyi PL (ed), Speech Separation by Humans and Machines. Kluwer Academic, pp. 167–180.

Kennedy M (1985) The Oxford Dictionary of Music. Oxford: Oxford University Press.

Krumbholz K, Patterson RD, Pressnitzer D (2000) The lower limit of pitch as determined by rate discrimination. J Acoust Soc Am 108:1170–1180.

Krumbholz K, Patterson RD, Nobbe A, Fastl H (2003) Microsecond temporal resolution in monaural hearing without spectral cues? J Acoust Soc Am 113:2790–2800.

Lee S, Potamianos A, Narayanan S (1999) Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J Acoust Soc Am 105:1455–1468.

Licklider JCR (1951) A duplex theory of pitch perception. Experientia 7:128–133.

McIntyre ME, Schumacher RT, Woodhouse J (1983) On the oscillations of musical instruments. J Acoust Soc Am 74:1325–1345.

Meddis R, Hewitt M (1991) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. J Acoust Soc Am 89:2866–2882.

Molin NE, Lindgren L-E, Jansson EV (1988) Parameters of violin plates and their influence on the plate modes. J Acoust Soc Am 83:281–291.

Moore BCJ, Glasberg BR (1983) Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J Acoust Soc Am 74:750–753.

Patterson RD (1994a) The sound of a sinusoid: Spectral models. J Acoust Soc Am 96:1409–1418.

Patterson RD (1994b) The sound of a sinusoid: Time-interval models. J Acoust Soc Am 96:1419–1428.

Patterson RD, Irino T (1998) Modeling temporal asymmetry in the auditory system. J Acoust Soc Am 104:2967–2979.

Patterson RD, Robinson K, Holdsworth J, McKeown D, Zhang C, Allerhand M (1992) Complex sounds and auditory images. In: Cazals Y, Demany L, Horner K (eds), Auditory Physiology and Perception. Oxford: Pergamon Press, pp. 429–446.

Patterson RD, Allerhand MH, Giguère C (1995) Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. J Acoust Soc Am 98:1890–1894.

Patterson RD, Yost WA, Handel S, Datta AJ (2000) The perceptual tone/noise ratio of merged iterated rippled noises. J Acoust Soc Am 107:1578–1588.

Patterson RD, Unoki M, Irino T (2003) Extending the domain of center frequencies for the compressive gammachirp auditory filter. J Acoust Soc Am 114:1529–1542.

Patterson RD, Smith DDR, van Dinther R, Walters TC (2008) Size Information in the production and perception of communication sounds. In: Yost WA, Popper AN, Fay RR (eds), Auditory Perception of Sound Sources. New York: Springer, pp. 43–75.

Peterson GE, Barney HL (1952) Control methods used in a study of the vowels. J Acoust Soc Am 24:175–184.

Pressnitzer D, Patterson RD, Krumbholtz K (2001) The lower limit of melodic pitch. J Acoust Soc Am 109:2074–2084.

Schelleng JC (1963) The violin as a circuit. J Acoust Soc Am 35:326–338.

Slaney M, Lyon RF (1990) A perceptual pitch detector. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, New Mexico, pp. 357–360.

Smith DRR, Patterson RD (2005) The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J Acoust Soc Am 118:3177–3186.

Smith DRR, Patterson RD, Turner RE, Kawahara H, Irino T (2005) The processing and perception of size information in speech sounds. J Acoust Soc Am 117:305–318.

Sprague MW (2000) The single sonic muscle twitch model for the sound-production mechanism in the weakfish, Cynoscion regalis. J Acoust Soc Am 108:2430–2437.

Turner RE, Walters TC, Monaghan JJM, Patterson RD (2009) A statistical formant-pattern model for estimating vocal-tract length from formant frequency data. J Acoust Soc Am 125:2374–2386.

Unoki M, Irino T, Glasberg B, Moore BC, Patterson RD (2006) Comparison of the roex and gammachirp filters as representations of the auditory filter. J Acoust Soc Am 120:1474–1492.

van Dinther R, Patterson RD (2006) Perception of acoustic scale and size in musical instrument sounds. J Acoust Soc Am 120:2158–2176.

Yost WA (1996) Pitch of iterated rippled noise. J Acoust Soc Am 100:511–518.

Yost WA (2009) Pitch perception. Atten Percept Psychophys 71:1701–1715.

Yost WA, Patterson RD, Sheft S (1996) A time domain description for the pitch strength of iterated rippled noise. J Acoust Soc Am 99:1066–1078.

Yost WA, Patterson RD, Sheft S (1998) The role of the envelope in processing iterated rippled noise. J Acoust Soc Am 104:2349–2361.

Zwicker E (1974) On the psychophysical equivalent of tuning curves. In: Zwicker E, Terhardt E (eds), Facts and Models in Hearing. New-York: Springer-Verlag, pp. 132–140.