

3

Size Information in the Production and Perception of Communication Sounds

ROY D. PATTERSON, DAVID R.R. SMITH, RALPH VAN DINTHER,
AND THOMAS C. WALTERS

1. Introduction

This chapter is about the perception of the sounds that animals use to communicate at a distance and the information that these sounds convey about the animal as a source. Broadly speaking, these are the sounds that animals use to declare their territories and attract mates, and the focus of the chapter is the size information in these sounds and how it is perceived. The sounds produced by the sustained-tone instruments of the orchestra (brass, strings, and woodwinds) have a similar form to that of the communication sounds of animals, and they also contain information about the size of the source, that is, the specific instrument type (e.g., violin or cello) within an instrument family (e.g., strings). Animals and instruments produce their sounds in very different ways, and the comparison of these two major classes of communication sounds reveals the general principles underlying the perception of source size in communication sounds.

For humans, the most familiar communication sound is speech, and it illustrates the fact that communications sounds contain information about the size of the source. When a child and an adult say the “same” word, it is only the linguistic message that is the same. The child has a shorter vocal tract and lighter vocal cords, and as a result, the waveform carrying the message is quite different for the child. The situation is illustrated in Figure 3.1, which shows short segments of four versions of the vowel in the word “mama.” From the auditory perspective, a vowel is a “pulse-resonance” sound, that is, a stream of glottal pulses each with a resonance showing how the vocal tract responded to that pulse. From the perspective of communication, the vowel contains three important components of the information in the sound. The first component is the “*message*,” which is that the vocal tract is currently in the shape that the brain associates with the phoneme /a/. This message is contained in the shape of the resonance, which is the same in every cycle of all four waves. The second component of the information is the glottal pulse rate. In the left column of the

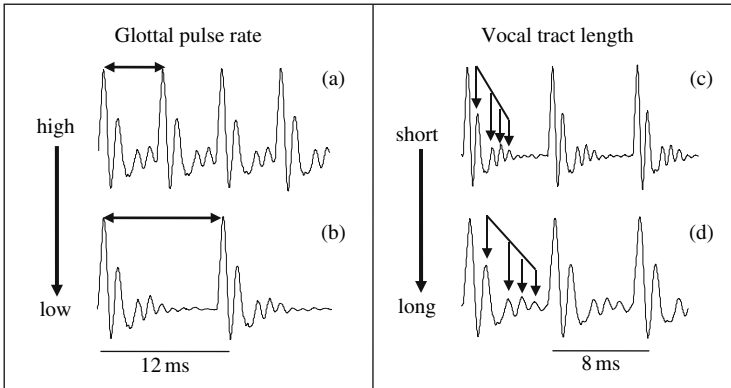


FIGURE 3.1. The internal structure of pulse-resonance sounds illustrating changes in pulse rate (a or b) and resonance rate (c or d).

figure, an adult has spoken the /a/ with a fast glottal pulse rate (a) and then a slow glottal pulse rate (b). The glottal pulse rate determines the pitch of the voice. The resonances are *identical*, since it is the same person speaking the same vowel. The third form of information is the resonance rate. In the right column, the same vowel is spoken by a child with a short vocal tract (c) and an adult with a long vocal tract (d) using the same glottal pulse rate. The glottal pulse rate and the *shape* of the resonance (the message) are the same, but the *rate* at which the resonance proceeds within the glottal cycle is faster in the upper panel. That is, the resonances of the child ring faster, in terms of both the resonance frequency and the decay rate. In summary, the stationary segments of the voiced parts of speech carry three forms of information about the sender: information about the shape of the vocal tract, its length, and the rate at which it is being excited by glottal pulses.

The components of the vocal tract (the nasal, oral, and pharyngeal passages) are tubes that connect the openings of the nose and mouth to the trachea and the esophagus. They are an integral part of the body, and they increase in length as the body grows. The decrease in pulse rate and resonance rate that occurs as humans grow up is a general property of mammalian communication sounds. Section 2 of this chapter describes the form of pulse resonance sounds, and Section 3 describes how information about source size is encoded in these sounds.

The fact that we hear the same message when children and adults say the same word suggests that the auditory system has mechanisms to adapt the analysis of speech sounds to the pulse rate and resonance rate, as part of the process that produces the size-invariant representation of the message. This suggests that there is an initial set of auditory processes that operate like a preprocessor to stabilize repeating neural patterns and segregate the pulse-rate and resonance-rate information from the information about the message. Irino and Patterson (2002) have demonstrated how these processes might work. First, the auditory system adapts the analysis to the pulse rate using “strobed temporal integration.”

Then the resulting “auditory image” is converted into a largely scale-invariant Mellin image with the aid of resonance-rate normalization. As a byproduct, the two processes produce a contour of pulse-rate information and a contour of resonance-rate information that the listener can use to estimate speaker size, and to track individual speakers in a multisource environment. Section 4 illustrates the representation of size information in the auditory system.

Section 5 describes recent psychophysical experiments that indicate that the vocal-tract-length information provided by resonance-rate normalization is perceived in terms of source size, and that it functions like a dimension of auditory perception much like pitch. Section 6 describes recent experiments designed to reveal the interaction of pulse rate and resonance rate in the perception of source size.

With regard to the topic of this book, the *auditory perception of sound sources*, the current chapter is restricted to one aspect of the perception of one class of sources, namely, the perceived size of pulse-resonance sources. We focus on this specific problem because we believe that it holds the clue to speaker normalization, that is, the ability of human listeners to recognize the message of speech *independent* of the size of the speaker. Machine recognizers are still severely handicapped in this regard. If we can characterize the transforms that the auditory system uses to perform pulse-rate and resonance-rate normalization, and integrate them with source-laterality processing (e.g., Patterson et al. 2006) and grouping by common onset (e.g., Cooke 2006), the resultant auditory preprocessor might be expected to enhance the performance of automatic speech recognition significantly. We return to the topic of speaker normalization in Section 7, where we compare our perceptual approach to speaker normalization with the more linguistic approach described by Lotto and Sullivan in Chapter 10 of this volume.

2. Communication Sounds

Pulse-resonance sounds are ubiquitous in the natural world and in the human environment. They are the basis of the calls produced by most birds, frogs, fish, and insects, as well as mammals, for messages that have to be conveyed over a distance, such as those involved in mate attraction and territorial defense (e.g., Fitch and Reby 2001). They are also conceptually very simple. The animal develops some means of producing a pulse of mechanical energy that causes structures in the body to resonate. From the signal-processing perspective, the pulse marks the start of the communication, and the resonances provide distinctive information about the shape and structure of parts of the sender’s body, and thus the species producing the sound. The pulse does not contain much information other than the fact that the communication has begun. Its purpose is to excite structures in the body of the animal that then resonate in a unique way. The resonance has less energy than the pulse but more information; it follows directly after the pulse and acts as though it were attached to it. So the location of the species-specific information is very predictable; it is tucked in behind each pulse.

In human speech, the vocal cords in the larynx at the base of the throat produce a pulse by momentarily impeding the flow of air from the lungs; this pulse of air then excites complex resonances in the vocal tract above the larynx. The mechanism is described in the next section. The mechanism is essentially the same in all mammals, and there is a similar mechanism in many birds and frogs; they both excite their air passages by momentarily interrupting the flow of air from the lungs. Fish with swim bladders often have muscles in the wall of the swim bladder (e.g., the weakfish, *Cynoscion regalis*) that produce brief mechanical pulses, referred to as “sonic twitches” (Sprague 2000), and these twitches resonate in the walls of the swim bladder in a way that makes the combination distinctive. Note that the sound-producing mechanisms in these four groups of vertebrates (fish, frogs, birds, and mammals) probably all evolved separately; the swim-bladder mechanism in the fish did not evolve into the vocal tract mechanism of the land animals, and the vocal tract mechanisms do not appear to have developed one from another. The implication is that this is convergent evolution, with nature repeatedly developing variations of the same basic solution to acoustic communication—the combination of a sharp pulse and a body resonance.

The sustained-tone instruments of the orchestra (brass, strings, and woodwinds) are also excited by nonlinear processes that produce sharp pulses that resonate in the air columns, or air cavities, of the instruments (Fletcher and Rossing 1998); so they also produce pulse resonance sounds (van Dinther and Patterson 2006). Combustion engines produce mini-explosions that resonate in the engine block; so they are also pulse-resonance sounds. They are not communication sounds in the normal sense, but they show that the world around us is full of pulse-resonance sounds, which the auditory system analyzes automatically and effortlessly.

There are also many examples of communication sounds that consist of a single pulse with a single resonance: Gorillas beat their chests with cupped hands, elephants stomp on the ground, and blue whales boom. Chickens and lemurs cluck every few seconds as they search for food in leaf litter. Humans clap their hands to attract attention. The percussive instruments such as xylophones, woodblocks, and drums also produce single-cycle pulse-resonance sounds. One important class of these percussive sounds is the “struck bars and plates,” which are described in Chapter 2 of this volume, by Bob Lutfi. These percussive sources produce very different sounds from those of animals and sustained-tone instruments because the resonance occurs within the material of the bar, or plate, rather than in an air column, or air cavity, in an animal or instrument. The materials of the bars and plates (typically metal or wood) are dense and stiff, and so the resonances ring much longer in these sources. Nevertheless, they are pulse-resonance sources, and the principles of sound production and perceptual normalization are similar to those for the sustained tones produced by speech and musical sources.

The variety of these pulse-resonance sounds, and the fact that humans distinguish them, is illustrated by the many words in our language that specify

transient sounds; words such as click, crack, bang, thump, and word pairs such as ding/dong, clip/clop, tick/tock. In many cases, a plosive consonant and a vowel are used to imitate some property of the pulse-resonance sound.

Finally, it should be noted that in the world today, most animals produce their communication sounds in pulse-resonance “syllables,” that is, streams of regularly timed pulses, each of which carries a copy of the resonance to the listener. The syllables are on the order of 200–800 ms in duration, with a pulse rate in the region 10–500 Hz. The pulse rate rises a little at the onset of the sound, remains fairly steady during the central portion of the sound, and drops off with amplitude during the offset of the sound, which is typically longer and more gradual than the onset. A selection of four of these animal syllables is presented in Figure 3.2; they are the calls of (1) a Mongolar drummer, or Jamaica weakfish (*Cynoscion jamaicensis*), (2) a North American bullfrog (*Lithobates catesbeiana*), (3) a macaque (*Macaca mulatta*), and (4) a human adult saying /ma/.¹ The notes of sustained-tone instruments are like animal syllables with fixed pulse rates and comparatively flat temporal envelopes. Both of these classes of communication sound are completely different from the sounds of inanimate sources such as wind and rain, which are forms of noise. In the natural world,

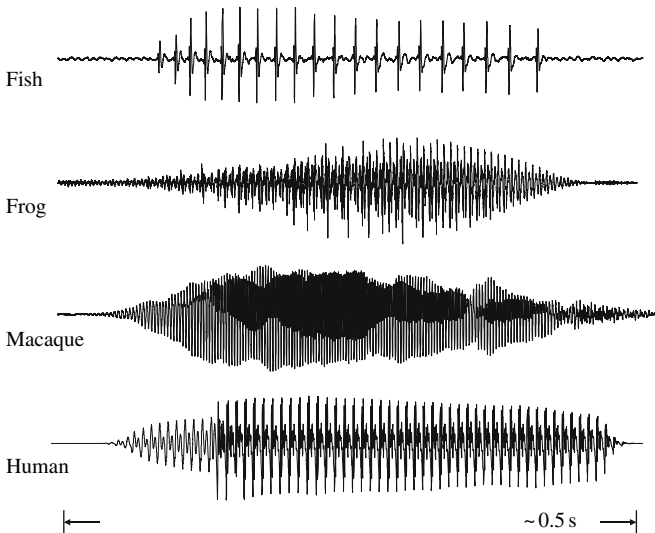


FIGURE 3.2. Communication calls from four animals (a fish, a frog, a macaque, and a human) illustrating that they all use “pulse-resonance” sounds for communication, and the duration of these animal syllables is on the order of half a second.

¹The Mongolar drummer call is available at <http://www.fishecology.org/soniferous>. The bullfrog and macaque calls were kindly provided by Mark Bee and Asif Ghazanfar, respectively. Many of the sounds presented in this chapter can be downloaded from the CNBH website: <http://www.pdn.cam.ac.uk/cnbh/>.

the detection of a pulse-resonance sound in syllable form immediately signals the presence of an animate source in the local environment.

3. Size Information in Communication Sounds

3.1 *The Effect of Source Size in Vocal Sounds*

In general, as a mammal matures and becomes larger, there is a consistent and predictable decrease in both the resonance rate and the pulse rate of its communication sounds, primarily because they are produced by structures that increase in size as the animal grows. The vibration of the vocal tract of a mammal is often modeled in terms of the standing waves that arise in a tube closed at one end (Chiba and Kajiyama 1942; Fant 1970). The resonances of the vocal tract are referred to as formants, and for present purposes, the relationship between resonance rate and vocal tract length can be taken to be

$$F1_{\text{voice}} = \frac{c}{4L_{\text{tract}}}, \quad (3.1)$$

where c is the speed of sound in air (340 m/s) and L_{tract} is the length of the vocal tract, which can be as long as 17 cm in tall men. So the frequency of the first formant for men is on the order of 500 Hz = $\frac{340}{4 \times 0.17}$. The point to note is that the size variable, L_{tract} , is in the denominator on the right-hand side of the equation, which means that as a child grows up into an adult and the length of the vocal tract increases, the resonance rate of the first formant *decreases*. This is a general principle of formants and of mammalian communication sounds.

The vocal cords produce glottal pulses in bursts, and the vibration of the vocal cords can be modeled by the equation for the vibration of a tense string, although the vocal cords are actually rather complicated structures. The glottal pulse rate, $F0_{\text{voice}}$, is the fundamental mode of vibration of the vocal cords, and the relationship between glottal pulse rate and the properties of a tense string is

$$F0_{\text{voice}} = \sqrt{\frac{T_{\text{cords}}}{M_{\text{cords}}(4L_{\text{cords}})}} \quad (3.2)$$

where T_{cords} , M_{cords} , and L_{cords} are the tension, mass, and length of the vocal cords. In this case, there are two physical variables associated with size: They are the length of the vocal cords and their mass; both increase as a child grows up. The point to note is that both the mass and length terms are in the denominator on the right-hand side of the equation, and they combine multiplicatively, so an increase in size, be it length or mass, leads to a decrease in glottal pulse rate in either case. The average $F0_{\text{voice}}$ for children is about 260 Hz, and it decreases progressively to about 120 Hz in adult men. The reduction in pulse rate with increasing size is also a general principle of mammalian communication sounds.

Thus, when we encounter a new species of mammal, we do not need to learn about the relationship between call and size. If the syllables of one individual have a consistently lower pulse rate and a consistently lower resonance rate than the syllables of a second individual, then we can predict with reasonable confidence that the first individual is larger without ever having seen a member of the species.

Speakers can also vary the tension of the vocal cords and change the pitch of the voice voluntarily. They do this to make prosodic distinctions in speech; for example, in many European languages, speakers raise the pitch of the voice at the end of an utterance to indicate that it is a question. This is also how singers change their pitch to produce a melody. The voluntary variation of tension makes the use of pulse rate as a size cue somewhat complicated. But basically, for a given speaker, the long-term average value of the voice pitch over a sequence of utterances is size information rather than speech information, whereas the short-term changes in pitch over the course of an utterance are speech information (prosody) rather than size information.

Finally, note that in pulse-resonance sounds, the frequency of the resonance is always greater than the pulse rate; this is one of the defining characteristics of the sounds used by mammals for communication.

3.2 *The Effect of Source Size in Musical Instrument Sounds*

The instruments of the orchestra are grouped into “families” (brass, strings, woodwinds, percussion). The members of a family (e.g., trumpet, French horn, and tuba) have similar construction, and they produce similar sounds; they differ primarily in their size. The mechanisms whereby sustained-tone instruments (the brass, string, and woodwind families) produce their notes are quite different from one another, and quite different from the way mammals produce syllables. Nevertheless, the excitation in sustained-tone instruments is a regular stream of pulses (Fletcher and Rossing 1998), each of which excites the body resonances of the instrument. As a result, sustained-tone instruments produce pulse-resonance sounds (van Dinther and Patterson 2006), and the sounds reflect the size of the source both in their pulse rate and their resonance rate, albeit in rather different ways than for the voice.

The French horn illustrates the form of the size information. It is a tube closed at one end like the vocal tract, and so the equation that relates fundamental frequency to tube length is the same as the one used to specify the frequency of the first formant of the voice,

$$F0_{\text{horn}} = \frac{c}{4L_{\text{horn}}} \quad (3.3)$$

where L_{horn} is the length of the brass tube when it is unrolled. However, in brass instruments, the length of the tube is associated with the *pulse rate* of the note rather than the frequency of the lowest body resonance. So the $F0$ is associated

with the pitch of the note that the instrument is playing rather than its brassy timbre. The relationship between the F0 of the instrument and the pulse rate at any particular moment is complicated by the fact that the pulse rate is also affected by the tension of the lips, and the fact that it is not actually possible to excite the instrument with a pulse rate as low as its F0. The length of the French horn is about 3.65 m, so its F0 is about 23.3 Hz. This is actually below the lower limit of melodic pitch (Krumbholz et al. 2000; Pressnitzer et al. 2001). If for the sake of this illustration, however, we take this F0 to be c1, then the instrument can be made to produce pulse rates that are harmonics of C1, beginning with C2, that is, C2, G2, C3, E3, G3, etc., by increasing the tension of the lips. The point of the example, however, is that the equation for pulse rate in brass instruments contains a size variable, e.g., L_{horn} , and as the size of the instrument increases, the pulse rate decreases because the length of the tube is in the denominator on the right-hand side of the equation.

The broad mid-frequency resonance that defines the timbre of all brass instruments is strongly affected by the form of the mouthpiece. The mouthpiece can be modeled as an internally excited Helmholtz resonator (Fletcher and Rossing 1998). The vibration of a Helmholtz resonator is much more complicated than that of a tube, but it is nevertheless instructive with respect to the effects of source size on the acoustic variables of the French horn sound. If we designate the resonance frequency $F1_{\text{horn}}$ by analogy with $F1_{\text{voice}}$, then the resonance rate of the formant is

$$F1_{\text{horn}} = \frac{c}{2\pi} \sqrt{\frac{A_{\text{stem}}}{L_{\text{stem}} V_{\text{bowl}}}} \quad (3.4)$$

Here A_{stem} and L_{stem} are the area and length of the stem that connects the bowl of the mouthpiece to the tube, and V_{bowl} is the volume of the bowl of the mouthpiece. This is a much more complex equation involving three size variables, and the balance of these variables is crucial to the sound of a brass instrument. For present purposes, however, it is sufficient to note that the most important size variable is the volume of the bowl, and it is in the denominator; so once again, the rate of the body resonance decreases as the size of the bowl increases.

Similar relationships are observed in the other families of sustained-tone orchestral instruments, such as the woodwinds and strings; as the size of the components in the vibrating source and the resonant parts of the body increases, the pulse rate and the resonance rate decrease. This is the form of the size information in the sounds that animals use to communicate at a distance, and it is the form of the size information in the notes of sustained-tone instruments. Musical notes have a more uniform amplitude envelope and a more uniform pulse rate than animal syllables. Nevertheless, the size information has a similar form because of the basic properties of vibrating sources; as the components get larger in terms of mass, length, or volume, they oscillate more slowly.

The same physical principles also apply to the percussive sources described in Chapter 2, which produce single-cycle pulse-resonance sounds. For example, in the equation that specifies the natural frequencies of a struck bar (Chapter 2, Eq. [2.3a]), the length term is in the denominator, so the natural frequencies decrease as bar length increases. Similarly, in the equation for the natural frequencies of a struck plate (Lutfi, Chapter 2, Eq. [2.5]), the length and width terms are both in the denominator. Thus, size information is ubiquitous in mechanical sound sources. We turn now to the perception of source size in speech sounds and musical sounds.

4. The Form of Size Information in the Human Auditory System

The representation of size information in the auditory system has been illustrated by Irino and Patterson (2002) using a pair of /a/ vowels like those in the right-hand column of Figure 3.1. The two vowels were simulated using the cross-area function of a Japanese male saying the vowel /a/ (Yang and Kasuya 1995). In one case, the vocal tract length was that appropriate for an average male (15 cm); in the other, the length was reduced by one-third (10 cm), which would be appropriate for a small woman. The glottal pulse rate (GPR) was the same in the two vowels as it is in Figures 3.1c and 3.1d. The auditory image model (AIM, Patterson et al. 1992, 1995) was used to simulate the internal representation of the two vowels; the resulting “stabilized auditory images” are shown in Figure 3.3 which is a modified version of Figure 3 in Irino and Patterson (2002). Briefly, a

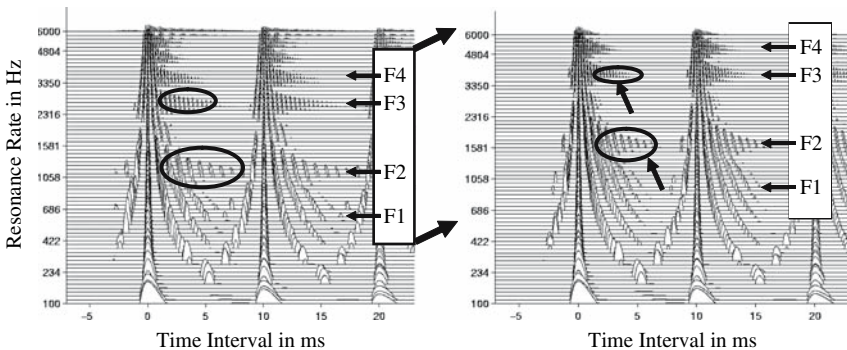


FIGURE 3.3. Auditory images of the vowel /a/ produced by a person with a long vocal tract (left column) and a short vocal tract (right column), showing the form that a change in source size takes in the internal auditory representation of these pulse-resonance sounds. The formants (F1–F4) move up as a unit on the tonotopic axis (the ordinate); that is, the resonance rates of the formants increase proportionately, and they have proportionately shorter duration.

gammatone auditory filterbank is used to simulate the basilar membrane motion produced by the vowel, and the resulting neural activity is simulated by applying half-wave rectification and adaptive compression separately to each channel of the filterbank output (Patterson and Holdsworth 1996). The repeating waveform of the vowel sound produces a repeating pattern of neural activity in the auditory nerve. In AIM, the pattern is stabilized by (1) calculating time intervals from the neural pulses produced by glottal pulses to the neural pulses produced by the remaining amplitude peaks within the glottal cycle, and (2) cumulating the time intervals in a dynamic, interval histogram (one histogram for each channel of the filterbank). The result of this “strobed temporal integration” (Patterson et al. 1992; Patterson 1994) is an array of dynamic, interval histograms that represents the auditory image; it is intended to simulate the first internal representation of the sound of which you are aware. The stabilization mechanism is assumed to be in the brainstem or thalamus.

The GPR of the synthetic /a/ vowels was 100 Hz, so the time between glottal pulses was 10 ms in both cases. The glottal pulses excite most of the channels in the filterbank, and so there are peaks at 0 ms and multiples of 10 ms in each channel, and these peaks form vertical ridges in the auditory image (Figure 3.3). This is the form of voice pitch in the auditory image: a vertical ridge that moves left as pitch increases and right as pitch decreases.

The rightward-pointing triangles on the vertical ridges are the formants of the vowels in this representation (marked by F1–F4 and arrows). They show that the vocal tract resonates longer at these frequencies. The overall shape of the patterns is quite similar, since it is the same vowel, /a/. The formants in the right-hand auditory image are shifted up, as a unit, along the quasi-log-frequency dimension, and a comparison of the fine structure of the formants in the corresponding ellipses shows that the formants ring faster in the auditory image of the vowel from the shorter vocal tract. This is the form of a change in vocal tract length in the auditory image: the resonances move up as a group (that is, the resonance rates increase), and the resonances decay away faster, so that the pattern shrinks in width. The same form of change occurs when the body resonators of musical instruments are reduced in size and when the struck bars and plates described in Chapter 2 are reduced in size. In the latter case, the resonance structure is attached to the 0-ms vertical, and the resonance structure extends across the full width of the image, because the density and stiffness of bars and plates means that the resonances ring much longer than those of the vocal tract or sustained-tone instruments.

The dimensions of the auditory image are both forms of frequency; the ordinate is acoustic frequency, which for narrow resonators is the resonance rate; the abscissa is the reciprocal of pulse rate. So the auditory image segregates the two components of size information and presents them, as frequencies, in a simple orthogonal form. The time interval between the vertical ridges in the auditory image is directly related to the size of the vocal cords, and the period of the individual resonances is directly related to the size of the resonators in the body of the source. Thus, in this image, changes in the size of the

excitation source are reflected in proportional changes in the time intervals between the vertical ridges, and changes in the size of the vocal resonators are reflected in proportional changes in the time intervals in the triangular structures that represent the formants. Irino and Patterson (2002) and Turner et al. (2006) have demonstrated how the auditory image can be converted into a Mellin image in which the pattern of the “message,” /a/, is truly scale-invariant. This aspect of size processing is beyond the scope of the current chapter.

Finally, note that whereas strobed temporal integration preserves the details of the resonances as they arise in basilar membrane motion, pitch mechanisms based on autocorrelation and the autocorrelogram do not (Licklider 1951; Slaney and Lyon 1990; Meddis and Hewitt 1991; Yost et al. 1996). Autocorrelation averages periodicity information over the glottal cycle. Whenever the resonance period is not an integer divisor of the glottal period, the periodicity information provided by the autocorrelation differs from the resonance rate of the formant. Thus, although autocorrelation can be used to predict the pitch and pitch strength of a vowel with great accuracy, the calculation smears the fine structure of the formant information (cf., for example, Figures 3.2c and 3.3c of Patterson et al. 1995), and consequently, it reduces the fidelity of any subsequent size-invariant representation of the message.

The pulse rate and resonance rate of a sound do not describe the size of a source in absolute terms. They are acoustic variables that describe properties of the sound wave as it travels from the sender to the listener. The acoustic variables change in a predictable way as the resonators in the sender’s body grow. However, the brain does not have the equations required to convert a pulse rate into a mass or a length, and even if it had the equations, there would still be difficulties. The information about all of the physical variables involved in the production of the sound has to be transmitted to the listener via only two acoustic variables: pulse rate and resonance rate. These acoustic variables often vary with the product of several physical variables like mass and length, so a given pulse rate could be produced by many different combinations of mass and length. So what the listener receives is one pulse-rate value that summarizes the aggregate effect of all of the physical variables on the vibration source, and one resonance-rate value that summarizes the aggregate effect of another set of physical variables on resonance rate.

Moreover, the brain is not actually interested in the mass, length, or volume of the physical components of the sounder, such as the size of the vocal cords or the length of the vocal tract. What matters to the listener is the size of the sender’s body: some perceptual and/or cognitive combination of the sender’s height, mass, and volume, and within a species, whether one sender is much bigger or smaller than another. In order to estimate the sender’s body size, a more central mechanism must combine the pulse-rate and resonance-rate information with some form of stored knowledge about the structure of the sender and/or a body of experience with a range of individuals from the specific population.

This is a complex problem, to which we return in Section 6. The next section is concerned with the much simpler problem of comparing the relative size of two individuals from the same species, or two musical instruments from the same family.

5. The Perception of Relative Size in Communication Sounds

Broadly speaking, the resonators in animals maintain their shape and composition as the animal grows, because the resonators are part of the sender's body. So within a population of senders, the function that relates the physical variables describing resonator components to the acoustic variables remains the same, and the constants maintain their fixed values. Thus, the changes are typically limited to the specific values of a small number of size-related variables, whose growth patterns are correlated and whose effects all go in the same, predictable, direction. As a result, differences in pulse rate and resonance rate provide useful information about the relative size of individuals within a population of senders. In this section, we describe perceptual experiments designed to demonstrate that listeners perceive the size information provided by the resonance rate and the pulse rate, and that they can discriminate relatively small changes in resonance rate as well as pulse rate. The results support the hypothesis that resonance rate is a dimension of auditory perception like pitch, and that together, resonance rate and pulse rate largely determine our perception of the relative size of animals and musical instruments.

5.1 Discriminating Speaker Size from Changes in Vocal Tract Length

Recently, two high-quality voice processing systems have been developed that make it possible to dissect segments of natural speech and manipulate the vocal tract length (VTL) and glottal pulse rate (GPR) information without changing the other qualities that specify the message and the speaker's identity. These voice coders, or vocoders, are referred to by the acronyms STRAIGHT (Kawahara et al. 1999; Kawahara and Irino 2004) and PRAAT (Boersma 2001), and they have made it possible to perform experiments on the perception of size information in natural speech with precise stimulus control. PRAAT has the advantage that it can extract formant contours as well as the voice pitch from utterances. The advantage of STRAIGHT is that the spectral envelope of the speech that carries the vocal-tract information is smoothed, as it is extracted, to remove the interference that occurs between the harmonic structure associated with the glottal-pulse rate of the stimulus, and the transfer function of the analysis window in the short-term Fourier transform. This helps to avoid the problem that LPC analysis has with the first formant when the GPR is relatively high and there are

only one or two harmonics of the voice pitch to define the first formant. When operating on speech, both PRAAT and STRAIGHT can produce resynthesized utterances of extremely high quality, even when the speech is resynthesized with pulse rates and vocal tract lengths beyond the normal range of human speech.

Smith et al. (2005) used STRAIGHT to control VTL in an experiment designed to measure a listener's ability to discriminate speaker size from differences in resonance rate. If acoustic scale functions as a dimension of sound as suggested by Cohen (1993), then we might expect to find that listeners can readily make fine discriminations in VTL, and thus speaker size, just as they can for the intensity of sound (loudness) or light (brightness). Moreover, if this is a general mechanism of auditory perception, we should expect to find that listeners can make size judgments even when the speech sounds are scaled to simulate humans much larger and smaller than those that the listeners have ever encountered.

Smith et al. (2005) prepared a set of "canonical" vowels /a/, /e/, /i/, /o/, /u/ from recordings made of author RP saying the vowels in natural /hVd/ sequences, i.e., "haad, hayed, heed, hoed, who'd." The vowels were edited to a common length of 600 ms by extracting the central sustained portion of the vowel and gating them on and off with a smooth cosine-squared envelope. The vowels were normalized to the same intensity level and the GPR was scaled to 113 Hz, which is near to the average for men. The VTL of these vowels was then scaled using STRAIGHT, which is actually a sophisticated speech-processing package that dissects and analyzes an utterance at the level of individual glottal cycles. It performs a "pitch synchronous" spectral analysis with a high-resolution fast Fourier transform (FFT), and then the envelope is smoothed to remove the zeros introduced by the position of the Fourier analysis window relative to the time of the glottal pulse. The resultant sequence of spectral envelopes describes the resonance behavior of the vocal tract in a form that is largely independent of pitch.

Once STRAIGHT has segregated a voiced sound into a GPR contour and a sequence of spectral envelope frames, the frequency dimension of the spectral envelope can be expanded or contracted independently of the GPR, and vice versa. Then the vowel can be resynthesized with its new GPR and VTL. The operations are largely independent, with the restriction that the GPR must never be higher than about half the frequency of the lowest formant for satisfactory resynthesis. When GPR is changed while keeping VTL constant, we hear one person repeating an utterance using different pitches, like singing a word on different notes; when VTL is changed keeping GPR constant, we hear something quite different, as though a set of people of different sizes were lined up on a stage, each saying the same word one after another, and all on the same pitch. Utterances recorded from a man can be transformed to sound like women and children. A demonstration of the manipulations possible with STRAIGHT is provided on the website² of the Centre for Neural Basis of Hearing. Liu and

²<http://www.pdn.cam.ac.uk/cnbh/>

Kewley-Port (2004) have reviewed STRAIGHT and commented favorably on the quality of its production of resynthesized speech. Assmann and Katz (2005) have also shown that a listener's ability to identify vowels is not adversely affected when they are manipulated by STRAIGHT over a reasonable range of GPR and VTL.

5.1.1 Speaker Size Discrimination with Vowel Sounds

In Smith et al. 2005, the scaling of VTL was accomplished simply by compressing or expanding the spectral envelope of the speech linearly along a linear frequency axis. On a logarithmic frequency axis, the spectral envelope shifts along the axis as a unit, and this is the form of size change in the frequency domain for information associated with resonance rate. The JND for speaker size was initially measured with single vowels using a two-alternative, forced-choice (2IFC) procedure. One vowel was presented in each interval, and the listener had to choose the interval corresponding to the speaker who sounded smaller. Psychometric functions showing percentage correct as a function of the difference in VTL between the speakers were measured for a variety of test voices with GPR values ranging from 40 to 640 Hz and VTL values from 7 to 24 cm (the average for adult males is about 16 cm). The results showed that detecting a change in speaker size based on a change in VTL is a relatively easy task. The JND was on average about 8%, which compares favorably with the JND for the intensity of a noise (loudness), which is about 10% (Miller 1947). The only exception was for vowels with long VTLs and the highest GPR (640 Hz); in this bottom-right corner of the GPR-VTL plane, the resonance of the vowel becomes long relative to the period of the sound, and the vowel becomes difficult to recognize; the lowest harmonic moves up in frequency beyond the position of the first formant.

By its nature, a change in vocal-tract length produces a predictable shift of the vowel spectrum, as a unit, along a logarithmic frequency axis, and the tonotopic axis along the basilar membrane is quasi-logarithmic. So, it might be possible for a listener to focus on one formant peak and perform the task by noting whether the peak shifted up or down in the second interval. Accordingly, Smith et al. (2005) ran a second version of the discrimination experiment with a more speechlike paradigm, which effectively precluded the possibility of using a simple spectral cue. The paradigm is presented in quasi-musical notation in Figure 3.4. Each interval of the trial contained a sequence of four vowels chosen randomly without replacement from the five used in the experiment, and the vowels were presented with one of four pitch contours, again chosen randomly. The duration of the vowels was shortened to about 400 ms to make the sequences sound more natural. The starting point for each pitch contour was varied randomly over a 9% range, and the level of the vowels in a given interval was roved in intensity over a 6-dB range. The only fixed parameter within an interval was VTL, and the only consistent change between intervals for all of the vowels was VTL. As before, the listener's task was simply to choose the interval with the smaller speaker. In this paradigm, the listener cannot do the task by

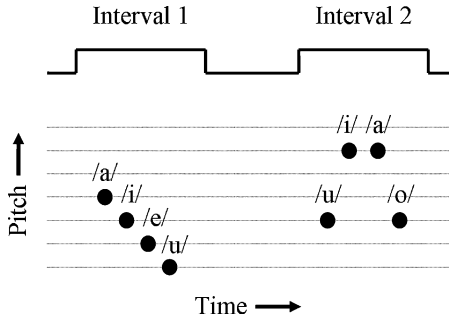


FIGURE 3.4. Schematic of the vowel-phrase paradigm for the VTL discrimination experiment of Smith et al. (2005). The only consistent difference between the vowels in the two intervals is vocal tract length.

listening to a single spectral component and noting whether it shifts up or down in the second interval. It is also the case that this paradigm naturally prompts the listener to think of the sounds in the two intervals as coming from two different speakers; the natural prosody of the sequences discourages listening for spectral peaks.

The experiment with speechlike vowel sequences produced JNDs that were similar to those obtained with single vowels. Together, the experiments with single vowels and vowel sequences show that listeners can make fine judgments about the relative size of two speakers, even when other properties of speech are varying, and that they can make size judgments for vowels scaled well beyond the normal range in both VTL and GPR. The JND for VTL information in vowels was less than 10% over a wide area of the GPR–VTL plane. When the GPR is 160 Hz, there are approximately 10 JNDs in speaker size between VTL values of 7 and 24 cm, so a JND corresponds to a VTL difference of about 2 cm. This supports the hypothesis that acoustic scale functions as a dimension in auditory perception (Irino and Patterson 2002).

5.1.2 Speaker Size Discrimination with Syllable Phrases

The experiments of Smith et al. (2005) on speaker size discrimination have since been extended to syllable phrases by Ives et al. (2005) in a study that greatly increased the variability of the stimulus set and made the task much more like that experienced in everyday speech. Ives et al. created a large, balanced database containing 90 consonant–vowel (CV) syllables and 90 vowel–consonant (VC) syllables. Each of the CV and VC categories contained three groups of 30 syllables distinguished by consonant category: sonorants, stops, and fricatives. In each group, six consonants of a specific type were paired with each of five vowels (/a/, /e/, /i/, /o/, and /u/). The full set of 180 syllables is presented in Ives et al. (2005), Table (1). The perceptual center of each syllable was determined (Marcus 1981; Scott 1993) and then used to ensure that the rhythm in the syllable phrases was fairly even, as it is in speech.

The vowels were scaled in GPR and VTL using STRAIGHT to simulate five categories of speaker with GPRs ranging from 80 to 320 Hz and VTLs ranging from 9 to 17 cm. The speaker types included three common categories with typical combinations of GPR and VTL, a large male, a large female, and a small child; and two unusual categories, one with a short vocal tract and a low pitch and one with a long vocal tract and a high pitch. The listeners were presented with two phrases of four syllables in a 2IFC discrimination paradigm very similar to that of Smith et al. (2005). There was a consistent difference in VTL between the two phrases, and the difference was varied over trials to determine the JND for VTL. The syllables in each phrase were selected randomly, with replacement, from one of the six groups within the database (e.g., CV-sonorants, CV-stops, CV-fricatives, VC-sonorants, VC-stops, or VC-fricatives). The level of the syllables in each phrase was roved between phrases over a 6-dB range, and the GPR of each of the syllables within the phrase was varied along one of four pitch contours.

The JNDs for the adult male and female speakers are just over 4% for all six syllable types. The JNDs for the other three categories are slightly larger (5%–6%), due mainly to worse performance on syllables containing stop consonants (/b/, /d/, /g/, /p/, /t/, /k/). Thus, the average for all speaker categories and syllable types is about 5%, which is considerably less than the value observed with vowels in a similar paradigm, and the reduction in the JND occurs despite the increased complexity of the stimulus set. Ives et al. (2005) attribute the improvement in performance to the greater naturalness of the speech in the syllable experiment. Although Smith et al. (2005) recorded natural vowels, they extracted the sustained portion in the center of the vowel and applied a cosine onset envelope to all of the vowels, which made them more similar and less natural. In the syllable experiment, the natural onset of each individual syllable was preserved, and the stimuli sounded considerably more natural as a result.

Finally, a note of caution is due with respect to predicting a speaker's height from his or her voice. Although there is a strong correlation between vocal tract length and speaker height over the full range of heights (Fitch and Giedd 1999), and although this makes it easy to distinguish children from adults, it is nevertheless the case that within small groups of adult men or adult women, you cannot expect to predict height differences from the voice differences with great accuracy. There are two reasons for this: First, the standard deviation for height in adult populations is relatively small, only about 4% of mean height, both for adult men and adult women. So the average height difference is relatively small in percentage terms. Second, the correlation between VTL and height is not perfect; on average, in the data of Fitch and Giedd (1999), the standard deviation for VTL, *given height*, is still about 6%. Thus, it is not surprising to find that the correlation between formant frequency and height is weak in small groups of adult men or adult women (González 2004; Owren and Anderson 2005; Rendall et al. 2005). It is also the case that in syllable phrases, the JND for the perception of a change in VTL is about 5% (Ives et al. 2005). So, with your eyes closed, you are not likely to be able to reliably discriminate the height of two men,

or two women, drawn randomly from the population, because the difference in VTL will probably be only one or two JNDs.

5.1.3 Resonance Rate Discrimination and Profile Analysis

In retrospect, it seems odd that the perception of speaker size has received so little attention in hearing and speech research. In spectral terms, the effect of a change in speaker size is theoretically very simple: If the GPR is fixed and the frequency axis is logarithmic, the profile for a given vowel has a fixed shape, and VTL changes simply shift the profile as a unit, toward the origin as size increases and away from it as size decreases. The analysis of spectral profiles is a well-known topic in psychoacoustics since it was introduced by Spiegel et al. (1981); see Green (1988) for a review. However, in the main, people have elected to follow Green and colleagues and concentrate on profiles constructed from sets of equal-amplitude sinusoids whose frequencies are equally spaced on a *logarithmic* axis. These stimuli are not like the voiced parts of speech; they do not have a regular harmonic structure, the excitation is not pulsive, and they sound nothing like vowels. Moreover, the task in traditional profile analysis (PA) is to detect an increment in one of the sinusoidal components, which is very different from detection of a shift in the spectral location of the profile as a whole.

Drennan (1998) provides an excellent overview of PA research that includes a few PA experiments in which the stimuli are composed of sets of harmonically related components that are intended to simulate vowel sounds to a greater or lesser degree (see, e.g., Leek et al. 1987; Alcántara and Moore 1995). However, there is no attempt to simulate the filtering action of the vocal tract and produce realistic vowel profiles; nor is there any attempt to simulate changes in VTL or measure sensitivity to coherent spectral shifts.

5.2 *Discriminating Instrument Size from Changes in Resonance Rate*

Further support for the hypotheses that acoustic scale is perceived as a dimension of auditory perception is provided by a recent study on the perception of size in musical instrument sounds. The instruments of the orchestra come in families, and within a family, the different instruments have the same shape and construction. The members of a family differ mainly in size. Musical sounds are pulse-resonance sounds, and although the mechanisms they use to produce their notes are sometimes very different from the way humans produce syllables, the notes of music carry size information in the form of a pulse rate and a resonance rate.

van Dinther and Patterson (2006) performed an experiment with scaled musical notes from four instrument families to determine the JND for a change in instrument size over a large range of pulse rates and resonance rates. The experiment focused on the baritone range in four instrument families: strings, woodwind, brass, and voice. Thus for the string family, it was the cello; for the woodwind family, the tenor saxophone; for the brass family, the French horn,

and for the human voice, the baritone. The notes were taken from a high-fidelity database of musical sounds (Goto et al. 2003). Each note was extracted with its natural onset to preserve the attack timbre of the instrument, and a cosine-squared envelope was applied to the end of the waveform to produce a smooth 50-ms offset. The total duration of the waveform was 350 ms. These specific instrument families were chosen because they produce sustained notes with similar temporal envelopes, and the notes have the pulse-resonance structure that STRAIGHT is most successful in scaling (Kawahara and Irino 2004).

STRAIGHT was used to modify the pulse rate (PR) and resonance rate (RR) of the notes and produce the small changes required for the discrimination experiment. The JND was measured for five combinations of pulse rate and resonance rate in a pattern similar to that used in Ives et al. (2005). The pulse rates were G_1 , G_2 , and G_3 (49, 98, and 198 Hz), and the resonance rate was scaled up or down by two-thirds of an octave; the design is illustrated in Figure 3.5. The procedure was similar to that employed in the speechlike version of the discrimination experiment in Smith et al. (2005). Each interval of a trial contained a short melody, instead of a single note, to preclude listeners from performing the task on the basis of a shift in a single spectral peak. The melodies also promote a musical mode of listening (synthetic rather than analytic). Scaled versions of the notes were then used to generate two-sided psychometric functions showing how much the resonance rate of the instrument has to be decreased or increased from that of the standard for reliable discrimination. The psychometric functions

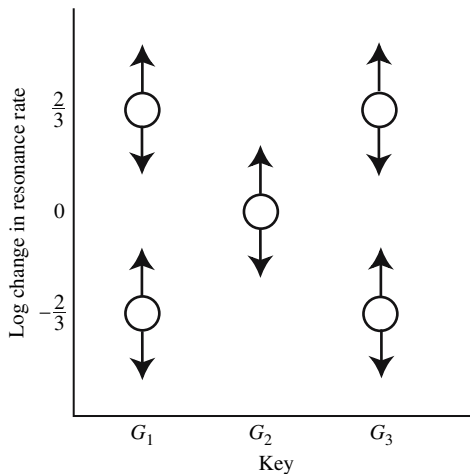


FIGURE 3.5. The combinations of pulse rate and resonance rate used as “standards” in the experiment of van Dinther and Patterson (2006) on discrimination of instrument size. The abscissa shows the pulse rate in musical notation; the ordinate shows the factor by which the resonance rate was modified (in log units). The arrows show the direction in which the JNDs were measured.

were measured for each PR–RR combination of each of the four instruments (cello, saxophone, French horn, and baritone voice).

The results show that listeners are able to discriminate size in instrument sounds, and they can specify which is the smaller of two instruments from short melodies. Within a family, the JND is relatively consistent, typically varying by no more than a factor of two across conditions. The JNDs for the baritone voice are comparable to those observed in the discrimination experiments of Ives et al. (2005), averaging around 5%. The JNDs for the French horn are around 8%, while those for the saxophone are around 12%. So the JNDs for these two instruments are similar to those for other sensory dimensions (Miller 1947), and about double the JND for speech in syllable phrases (Ives et al. 2005). The JNDs in the music study are largest for the cello. The JND is about 10% when the pitch is low and the instrument is small, or the pitch is high and the instrument is large, but they increase to around 20% when the pitch is low and the instrument is large or the pitch is high and the instrument is small. And overall, listeners have slightly more difficulty when the instrument is large and plays a low-pitched melody.

In summary, the psychometric functions associated with discriminating a change in source size as a function of resonance rate are steep and consistent, supporting the argument that resonance rate functions as a dimension of auditory perception. The slope of the psychometric function shows that listeners need a change of 5%–10 % in resonance rate to discriminate a change in the size of the resonators in the vocal tract or the bodies of musical instruments.

5.3 Discriminating Pulse Rate in Vowel Sounds and Click Trains

The basic relationship between the acoustic variable “glottal pulse rate” and the perceptual variable “voice pitch” is straightforward: Pitch increases with pulse rate. Indeed, the relationship is so simple that the pitch of the voice is normally expressed in terms of glottal pulse rate in Hertz. Voice pitch is also expressed as the fundamental, F_0 , of the harmonic series observed in the magnitude spectrum of the short-term Fourier transform of the note. But for most purposes, it is simpler just to think of voice pitch as glottal pulse rate.

As a child grows up into an adult, its vocal cords become longer and heavier (Titze 1989), and the GPR decreases from about 260 Hz for small children of both sexes to about 220 Hz for women and about 120 Hz for men. The change proceeds smoothly with height as girls grow up into women. For boys, however, when they reach puberty there is an increase in testosterone, which accelerates growth in the laryngeal cartilages (Beckford et al. 1985). As a result, there is a sudden drop in GPR by almost an octave at around 13 years of age. In the adult population, once the effect of sex is removed, there is no direct correlation between body size and GPR; that is, the range of heights in the relatively small populations of men, or women, included in studies of GPR (e.g., Lass and Brown 1978; Künzel 1989; Hollien et al. 1994) is not large enough to

reveal a correlation with height, given the variability in GPR. Thus, there is size information in GPR, in the sense that one can reliably distinguish small children from large adults, but within a group of men or women, a difference in GPR is not a reliable indicator of a difference in height. As noted earlier, speakers vary GPR by varying the tension of the vocal cords to indicate prosodic distinctions in speech. So, for a given speaker, the long-term average value of his or her voice pitch, over a sequence of utterances, is size information, but the short-term changes in pitch, over the course of an utterance, are speech information rather than size information.

Although the relationship between voice pitch and the perception of size is somewhat complicated, the discrimination of a change in GPR per se is not. Smith et al. (2005) includes an experiment in which they measured the JND for voice pitch using synthetic vowel sounds for a wide range of combinations of GPR and VTL. On each trial, listeners were presented two vowels with the same VTL and that differed a little in GPR, and over the course of the trials, the GPR difference was varied to produce a psychometric function from which the JND was determined. When the GPR was in the normal range for the human voice, the JND was less than 2%. This performance also extended beyond the range of the human voice up to 640 Hz, and it was largely unaffected by the value of the VTL. That is, the discrimination of changes in voice pitch would appear to be largely independent of the properties of the resonance in speech sounds. When the GPR was reduced to 40 Hz, close to the lower limit of human pitch perception, the JND rose to about 9%.

The majority of data on the discrimination of pulse rate, however, come from research that is ostensibly on pitch perception as opposed to size perception. (In the next section, we describe an experiment that shows that changes in pulse rate interact with changes in resonance rate in the perception of changes in source size.) Temporally regular trains of very-short-duration pulses without resonances produce a sound with a strong pitch and a buzzy, mechanistic timbre. These “click trains” and sets of regularly spaced harmonics (which are their spectral equivalent) have been used to study what has been referred to as “residue pitch” (Schouten 1938), “periodicity pitch” (Licklider 1951), “repetition pitch” (Thurlow and Small 1955), “virtual pitch” (Terhardt 1974), and more recently, “melodic pitch” (Krumbholz et al. 2000; Pressnitzer et al. 2001). The discrimination of click rate is similar to the discrimination of the glottal pulse rate in speech sounds, but without any confounding influence from the vocal resonances. In these discrimination studies, matched pairs of click trains with slightly different click rates are compared (typically in a 2IFC paradigm) to determine the JND for a range of click rates. Krumbholz et al. (2000) provide a review of the studies dating back to Ritsma and Hoekstra (1974). They show that “rate discrimination threshold” (RDT), as it is called, is less than 2% for a wide range of click rates, provided that the stimulus contains energy in the region below 1000 Hz.

Krumbholz et al. (2000) extended the research and used RDT to measure the lower limit of “melodic” pitch (LLMP) as a function of the spectral location of

the energy in the stimulus. Their results are similar to those of Pressnitzer et al. (2001), who used bandpass-filtered click trains to construct four-note melodies and measure the LLMP in a musical context. The LLMP is about 32 Hz when the stimulus contains low-frequency energy down to 200 Hz. The LLMP increases as the energy moves up in frequency; the rate of increase is initially slow (the LLMP is still below 50 Hz when the lowest component is 800 Hz), but as the lowest frequency in the stimulus moves above about 1000 Hz, the rate of increase accelerates, and when the energy is all above 3200 Hz, the LLMP is about 300 Hz.

6. The Interaction of Resonance Rate and Pulse Rate in the Perception of Source Size

Estimating the *absolute* size of a source from a single auditory event is, theoretically, a much more difficult task than making a judgment about the relative size of two similar sources. The listener has to use experience and/or context to interpret the size information in the sound. Nevertheless, when the radio or the telephone presents us with a new, unknown speaker, we can tell whether the speaker is a child or an adult, which suggests that we have the relevant experience. We also know that there is size information in speech sounds. The length of the vocal tract is highly correlated with speaker height (Fitch and Giedd 1999), and the longer the vocal tract, the lower the formant frequencies (Chiba and Kajiyama 1942; Fant 1970). Specifically, as a child grows between the ages of 4 and 12, the formant frequencies of males decrease by about 32% from their values at age 4, while the formant frequencies of females decrease by about 20% over the same age range (Hollien et al. 1994; Huber et al. 1999).

The contrast between the theoretical problem of estimating the absolute size of a sound source and our apparent ability to do it with relative ease for humans prompted Smith and Patterson (2005) to measure listeners' ability to estimate speaker height for isolated vowels with a wide range of GPRs and VTLs. The data are of particular interest because they reveal an interaction between GPR and VTL in the estimation of speaker height.

6.1 The Interaction of GPR and VTL in the Estimation of Speaker Size

Listeners were presented isolated vowels scaled over a large range of GPR and VTL values, and requested to make two judgments about each vowel: the height of the speaker (on a seven-point scale from very short to very tall) and the speaker's natural category (man, woman, boy, or girl). The experiment was performed for two ranges of GPR and VTL values. The narrower range was similar to that encountered in the normal population: GPR varied from 80 to 400 Hz in six logarithmic steps, and VTL ranged from 22.2 cm to 7.8 cm in

six logarithmic steps. The wider range was chosen to extend the judgments well beyond the values encountered in everyday speech; GPR varied from 61 to 523 Hz in six logarithmic steps, and VTL ranged from 26.8 cm to 6.5 cm in six logarithmic steps. These VTLs simulate speakers ranging from a small child of height 0.6 m (a VTL of 6.5 cm) to a giant of height 3.7 m (a VTL of 26.8 cm). The data showed that the effect of range was small; that is, judgments of size made during the experiment with the extended range, for combinations of VTL and GPR that are commonly encountered, were essentially the same as the judgments made when vowels with similar combinations of VTL and GPR were presented in the experiment with the smaller range.

The results from the two experiments combined are shown in Figure 3.6 as a size surface over the GPR–VTL plane. The figure shows that listeners reliably reported that vowels spoken with a low GPR and a long VTL came from a very tall person (the upper back part of the surface) and that vowels spoken with a high GPR and a short VTL came from a small person (the lower front part of

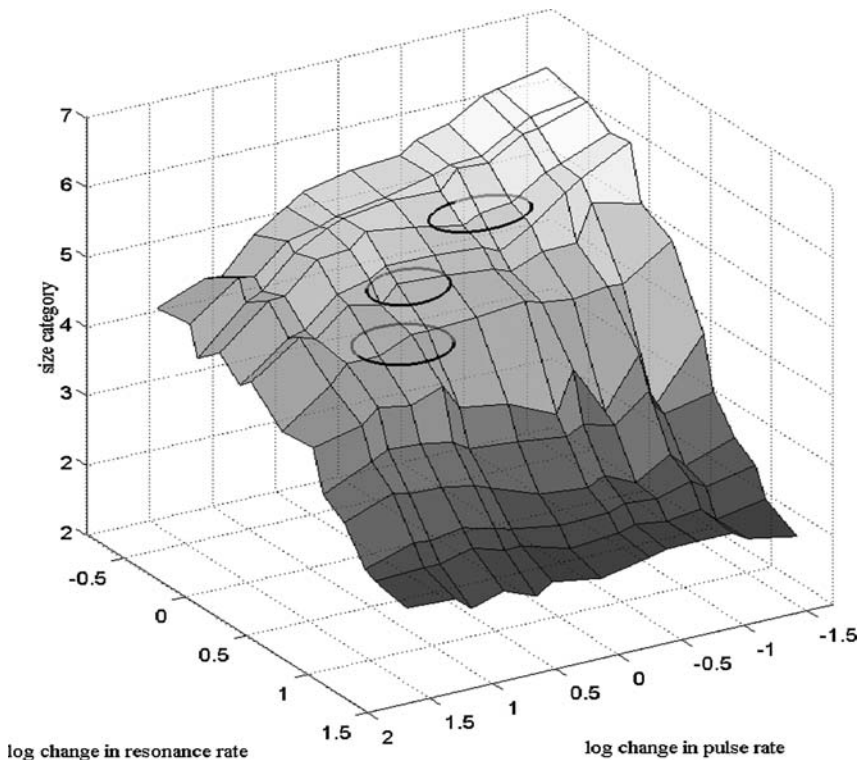


FIGURE 3.6. A surface of size judgments showing the average estimated size for voices with a wide range of combinations of glottal pulse rate and vocal tract length. The surface was constructed from the data of Smith and Patterson (2005). The three ellipses show the normal combinations of GPR and VTL for men, women, and children, estimated from the data of Peterson and Barney (1952).

the surface). However, the surface is not planar, indicating that in at least part of the space, GPR and VTL interact in the determination of the perceived size of the speaker. Broadly speaking, on these log-log coordinates, GPR has a nearly linear effect on perceived size for those VTLs in the normal range for adults and for VTLs longer than those typically encountered in everyday life. The ellipses show the normal range of GPR and VTL values in speech for men, women, and children, derived from the vowels of 76 men, women, boys, and girls speaking ten vowels (Peterson and Barney 1952). The estimates of VTL were calibrated with measurements of the VTL taken from magnetic resonance images (Fitch and Giedd 1999). Each ellipse represents the mean ± 2 standard deviations in each dimension for each category of speaker.

In contrast, as VTL decreases through the range associated with children, the effect of GPR decreases *rapidly* for *low* GPRs and *slowly* for *high* GPRs. As a result, for VTLs in the range of most children, and for shorter VTLs, changes in GPR have very little effect on the perceived height of the speaker; all of the vowels are perceived as emanating from very small people. It is also worth noting that the data revealed very little evidence of learning: Listeners could perform at near-asymptotic levels after a few minutes' experience with the task.

6.2 *The Interaction of GPR and VTL in Instrument Identification*

The study by van Dinther and Patterson (2006) of size perception in musical instruments included an experiment to determine the extent to which listeners could recognize instrument sounds when their resonance rates and pulse rates had been increased or decreased with STRAIGHT. They used four families of instruments: strings, woodwinds, brass and voice, and chose four members with different sizes from each family. The specific instruments are listed in their Table 1 with the pitch range, or register. The sixteen starting notes that identify the instruments were scaled up and down by 5, 7, or 12 semitones and up by 7 or 12 semitones *in pulse rate*, and they were scaled up and down by one-third and two-thirds of an octave *in resonance rate*, making a total of 5×5 , or 25, versions of each note. A 16-alternative forced-choice procedure was used to measure recognition performance using a graphical interface with 16 buttons labeled with the 16 instrument names in the layout shown in their Table 1. On each trial, one of the 25 notes for one of the 16 instruments was selected and played to the listener three times. The listener's task was to identify the instrument from one of the 16 options.

The results showed that listeners could identify the scaled instrument notes reasonably accurately, even for notes scaled well beyond the normal range for that instrument. Performance was above 55% correct for all combinations of pulse rate and resonance rate, and it rose to about 85% correct for the unscaled notes. Chance performance for instrument identification in this task is 6.25% correct. An analysis of the errors showed that listeners were essentially perfect on the identification of instrument family. Moreover, when both the pulse rate

and the resonance rate were decreased, if the listener made an error, it was very likely that they would choose a larger instrument from within the same family. Similarly, when both the pulse rate and the resonance rate were increased, if the listener made an error, it was very likely that they would choose a smaller instrument from within the same family.

This prompted van Dinther and Patterson (2006) to summarize the within-family error data in terms of a surface that shows the trading relationship between pulse rate and resonance rate, in order to examine the interaction of pulse rate (PR) and resonance rate (RR) in the perception of instrument size. Specifically, for within-family errors, the percentage of cases in which each listener chose a larger member of a family was calculated as a function of the *difference* in pulse rate and the *difference* in resonance rate between the scaled and unscaled versions of the note. The results were presented as a contour plot (Figure 3.7) in which the dependent variable was the percentage of cases in which the listener chose a larger member of a family given a specific combination of pulse rate and resonance rate.

Consider the 50% correct contour line. It shows that there is a strong trading relation between a change in pulse rate and a change in resonance rate. When the pulse rate is increased on its own, it increases the percentage of cases in which the listener will choose a smaller member of the family. However, this

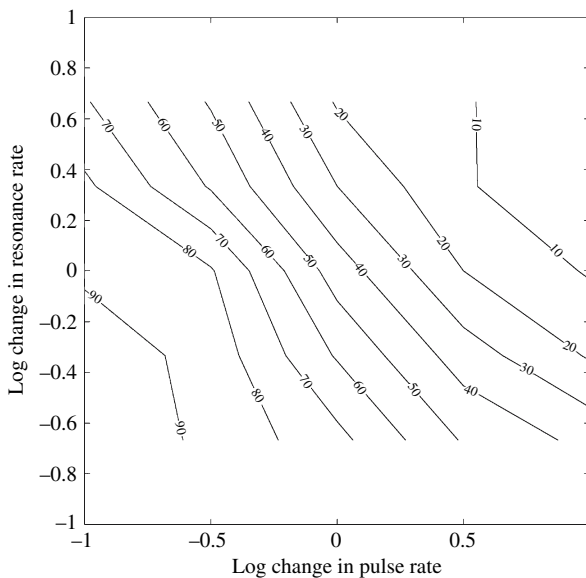


FIGURE 3.7. Contours showing the percentage of within-family errors where the listener chose a larger member of the family, plotted as a function of the difference in pulse rate (abscissa) and resonance rate (ordinate) between the scaled and unscaled versions of the note. The differences are plotted on an octave scale (base-2 logarithmic scale) for both the abscissa and the ordinate. The contours were constructed from the data of van Dinther and Patterson (2006).

can be entirely counteracted by a decrease in resonance rate, which makes the instrument seem larger. Moreover, the contour is essentially a straight line, and on these log-log coordinates, the slope of the line is on the order of -1 ; that is, in log units, the two variables have roughly the same effect on instrument identification. Very similar, essentially linear, trading relationships are observed for all of the contours between about 20% and 80% correct, and the spacing between the lines is approximately equal. Together, these observations mean that the errors are highly predictable on the basis of just two numbers: the logarithm of the change in pulse rate and the logarithm of the change in resonance rate. The fact that the surface is roughly planar means that the trading relationship can be characterized, except at the extremes, by a plane. For any point in the central range of the plane, a change in pulse rate of PR log units can be counteracted by a change in resonance rate of -1.3 PR log units. This means that when measured in log units, the effect of a change in pulse rate on the perception of size is a little greater than the effect of a change in resonance rate. However, the JND for resonance rate is larger than that for pulse rate, so if we express the relationship in terms of JNDs instead of log units, the relative importance of resonance rate increases. The JND for resonance rate was observed to be about 10%. The JND for pulse rate is more like 2% (Krumbholz et al. 2000; Figure 3.5). Therefore, one JND in resonance rate has about the same effect on the perception of size as four JNDs in pulse rate. The primary point, however, is that there is a very strong interaction between PR and RR in the perception of instrument size. Indeed, van Dinther and Patterson (2006) suggest that much of the difference in timbre between instruments within a family is size information associated with the pulse rate and the resonance rate.

6.3 The Interaction of GPR and VTL in Size Discrimination

The size surface of Smith and Patterson (2005) shows that (1) for long VTLs, the slope of the surface is shallow and uniform, (2) for shorter VTLs, in the range of normally sized children, the slope is steep for low GPRs and shallow for high GPRs, and (3) for the shortest VTLs, beyond the normal range, VTL still affects perceived size but GPR does not. The complexity of this surface prompted Gomersall et al. (2004) to develop a method for measuring the slope of the size surface directly using a 2AFC size-discrimination experiment.

It was assumed that the surface in a local region could be approximated by a plane, specifically, that a local region on the surface is reasonably well described by the first-order terms in a two-dimensional Taylor expansion. Listeners were required to discriminate between (1) a four-vowel phrase spoken by a “standard” speaker with a fixed combination of GPR and VTL (that is, a given point on the size surface) and (2) four-vowel phrases from test speakers with combinations of GPR and VTL that differed sufficiently to make their voices discriminable from the test speaker, but not so different as to violate the Taylor expansion criterion

(that coefficients above first order in the expansion be small relative to the first-order terms). The vowels for the test and standard speakers were generated using STRAIGHT from recordings of the vowels of one *female* speaker pronounced in /hVd/ format.

The JND for VTL is roughly three times the JND for voice pitch, so on log GPR versus log VTL coordinates, we might expect that the locus of speakers that are equally discriminable from the standard would have combinations of GPR and VTL values that form an ellipse about the standard speaker. The paradigm is illustrated in Figure 3.8, where the open circles show the GPR and VTL combinations for five test speakers, and the filled circles about the open circles show the GPR and VTL combinations for the respective test speakers. On a given trial, a random four-vowel phrase from the standard speaker (with one fixed combination of GPR and VTL values) is presented in one stimulus interval; another random four-vowel phrase from one of the test speakers (with a different, but fixed, combination of GPR and VTL values) is presented in the other interval, and the listener had to choose the interval with the smaller speaker. There were eight test speakers for each standard speaker spaced evenly about the ellipse, as indicated in the figure. The axis of the ellipse was tilted relative to the

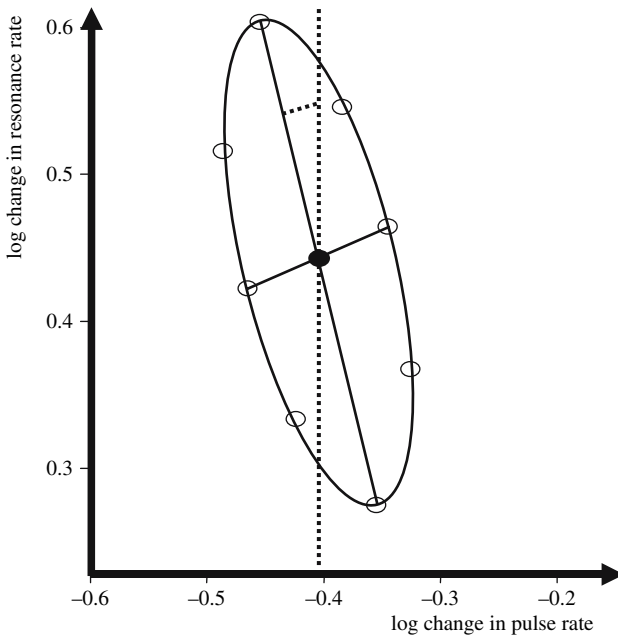


FIGURE 3.8. The ellipse of stimulus conditions used to measure the slope of the size surface for the standard voice representing an adult male. Each of the test voices (open circles) is compared repeatedly with the standard (filled circle) to determine which speaker sounds larger. The data are used to estimate the slope of the plane within the ellipse on the size surface. (Adapted from Gomersall et al. 2004.)

GPR–VTL coordinate system to ensure that both of the experimental variables changed from the first to the second interval on every trial. This helps to prevent listeners from focusing solely on the GPR or the VTL cue.

Test voices with higher GPR values and shorter VTL values tend to be heard as smaller than the standard speaker, and speakers with lower GPR values and longer VTL values tend to be heard as larger than the standard speaker. The eight probabilities, estimated by repeated pairings of a standard voice with each of their respective test voices, can be used to fit a plane to each ellipse of data (Gomersall et al. 2004). The line of steepest descent on the plane provides an estimate of the slope of the size surface at the point of the standard voice, and when the line of steepest descent is projected onto the GPR–VTL plane, the angle of the projected line reveals the tradeoff between VTL and GPR in the determination of perceived size.

The lines of steepest descent for 16 standard voices are presented in Figure 3.9. On this two-dimensional plot, the length of the vector shows the steepness of

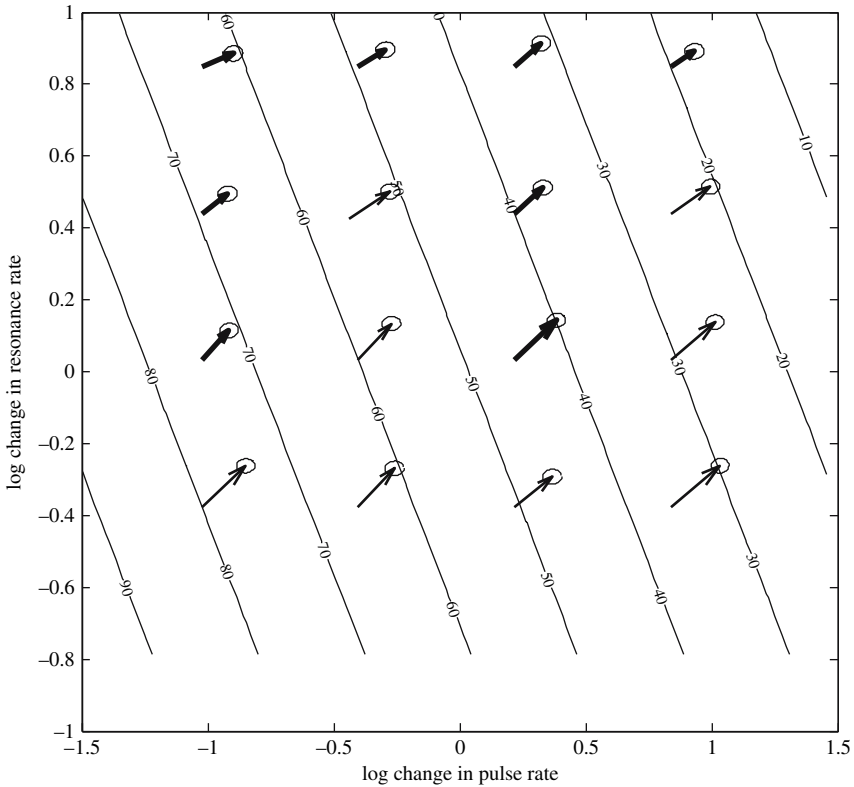


FIGURE 3.9. Size-surface slope vectors showing the angle of steepest descent for the ellipses associated with the 16 standard voices presented in the discrimination experiment of Gomersall et al. (2004).

the gradient (in relative terms), and the angle shows the direction in which the surface goes down fastest. The ellipse at the end of the vector shows the confidence limits for the estimate; that is, there is a 95% probability that the vector ends in the ellipse.

The figure shows that the gradient vectors do not vary substantially, either in terms of their length or their angle, across the GPR–VTL plane as much as might have been expected from the size surface generated with absolute judgments by Smith and Patterson (2005). The results are more like the uniform trading relationship derived from the within-family errors in the musical instrument study (van Dinther and Patterson 2006). The vectors are a little longer and the angles a little larger for the longer VTLs, but the differences are much less than the differences in slope associated with short and long VTLs in the absolute judgments.

7. Speaker Normalization

In Chapter 10 of this volume, Lotto and Sullivan argue that the “source” in the case of speech is the message of the communication rather than the pitch of the voice or the shape and length of the speaker’s vocal tract. Similarly, at the end of the introduction to this chapter, attention was drawn to the fact that pulse-rate normalization and resonance-rate normalization are not just mechanisms for estimating source size; they also adapt auditory processing to the size of a source and help the auditory system produce a largely size-invariant representation of the message. Indeed, it has been argued, in this chapter and elsewhere (e.g., Irino and Patterson 2002), that pulse-rate adaptation and resonance-rate normalization are general auditory mechanisms that evolved with animal communication to make auditory perception generally robust to variation in source size. It also seems likely that if one could develop an auditory preprocessor that combined pulse-rate adaptation and resonance-rate normalization with source-laterality processing and grouping by common onset, the preprocessor would very likely enhance the robustness of speech-recognition machines considerably.

In speech research, the processes that confer robustness on recognition are collectively referred to as “vowel normalization” (Miller 1989), “vocal tract normalization” (Welling and Ney 2002), or “talker normalization” (Lotto and Sullivan, Chapter 10), depending on the aspect of communication under consideration. Lotto and Sullivan provide a comprehensive review of the adverse effects of the many different forms of speaker variability on the robustness of automatic speech recognition (see their Section 4). They distinguish *intrinsic* normalization techniques “...that rely on information contained solely within the vowel...” from *extrinsic* normalization techniques, which use longer-term aspects of speaker variability, typically at the sentence level rather than the syllable level, for additional normalization (e.g., Ladefoged and Broadbent 1957). Whereas

the two classes of normalization techniques receive about equal attention in Chapter 10, the current chapter focuses entirely on just two of the intrinsic techniques: pulse-rate adaptation and resonance-rate normalization. There are several reasons for this: (1) These mechanisms function like mappings that can be applied to the time–frequency representation produced in the cochlea without reference to the context of the communication. (2) They automatically take care of a large portion of the acoustic variability associated with variation in speaker size, and so simplify the process of producing a size-invariant representation of the message at the syllable level. (3) The production of a largely size-invariant version of the message at an early stage in the processing facilitates subsequent extrinsic normalizations involving context; indeed, it may be essential for efficient functioning of extrinsic normalization. (4) There are algorithms for implementing these intrinsic normalizations and integrating them into the mainstream of computational auditory scene analysis; specific algorithms for extrinsic normalization are still in the early stages of development.

In Chapter 10, the primary example of intrinsic normalization is VTL normalization, which is based on formant ratio theory (FRT) (see Miller 1989 for a review). The theory originated with the historic conjecture by Lloyd (1890) that vowels are more readily identified by the ratios of their formant frequencies than by the absolute frequencies of the formants. A physical explanation of how the formant ratios of a vowel might remain largely unchanged as a child grows up and the absolute values of their formant frequencies decrease was provided by the early vocal tract models of Chiba and Kajiyama (1942) and Fant (1970). The basics of FRT from the auditory perspective were presented in Section 3.1 of the current chapter.

In Chapter 10, the importance of intrinsic pulse-rate normalization to speaker normalization and message invariance is largely overlooked, as is often the case in speech and language research. Miller (1989) developed an “auditory-perceptual” approach to speaker normalization, in which FRT was to be augmented by the inclusion of a “sensory reference” (SR). The SR was based on the individual speaker’s average pitch, measured relative to the average pitch of the population (Miller 1989, Appendix A), and it was used to adjust formant ratios to improve recognition rates. It is a form of intrinsic GPR normalization, inasmuch as it is used to scale the initial “auditory-perceptual” representation of a sound and to place it within an “auditory perceptual space” that is conceptually similar to the auditory image space described in Section 3. It is technically quite different, however, inasmuch as GPR adaptation precedes the calculation of formant ratios in AIM, whereas it is applied after the calculation of formant ratios in Miller’s model. It is also the case that it was developed to accommodate a subset of vowels associated with anomalous pitches, rather than all vowels, so it involves context in a way that is more typical of extrinsic normalization processes. In the end, however, we expect that optimal speech recognition will probably require both of these forms of GPR normalization in either intrinsic or extrinsic forms.

8. Summary

The pulse-resonance “syllables” that animals use to communicate, and the pulse-resonance notes that we use to make music, contain information about the size of the source in the pulse rate and the resonance rate. Both decrease as the size of the animal or the instrument increases. Humans perceive changes in resonance rate as changes in source size— either speaker size or instrument size—and they are very sensitive to changes in source size. Resonance rate appears to be a dimension of auditory perception just like musical pitch, and there is a tradeoff between pulse rate and resonance rate in the perception of source size.

The perceptual data support the hypothesis of Irino and Patterson (2002), Smith et al. (2005), and van Dinther and Patterson (2006) that the auditory system adapts to the pulse rate and normalizes for resonance rate as it constructs a largely size-invariant representation of the message of the syllable or the musical note.

Acknowledgments. Support for the writing of this paper and much of the research described in it was provided by the UK Medical Research Council (G990369, G0500221) and the German Volkswagen Foundation (VWF 1/79 783).

References

- Alcántara JI, Moore BCJ (1995) The identification of vowel-like harmonic complexes: Effect of component phase, level and fundamental frequency. *J Acoust Soc Am* 97:3813–3824.
- Assmann PF, Katz WF (2005) Synthesis fidelity and time-varying spectral change in vowels. *J Acoust Soc Am* 117:886–895.
- Beckford NS, Rood SR, Schaid D (1985) Androgen stimulation and laryngeal development. *Ann Otol Rhinol Laryngol* 94: 634–640.
- Boersma P (2001) Praat, a system for doing phonetics by computer. *Glott Int* 5(9/10): 341–345.
- Chiba T, Kajiyama M (1942) *The Vowel: Its Nature and Structure*. Tokyo: Tokyo-Kaiseikan.
- Cohen L (1993) The scale transform. *IEEE Trans Acoust Speech Signal Proc* 41:3275–3292.
- Cooke M (2006) A glimpsing model of speech perception in noise. *J Acoust Soc Am* 119:1562–1573.
- Drennan W (1998) Sources of variation in profile analysis: Individual differences, extended training, roving level, component spacing and dynamic contour. PhD thesis, Indiana University.
- Fant G (1970) *Acoustic Theory of Speech Production*, 2nd ed. Paris: Mouton.
- 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 B* 268:1669–1675.
- Fletcher NH, Rossing TD (1998) *The Physics of Musical Instruments*. New York: Springer-Verlag.

- Gomersall P, Walters T, Turner R, Patterson RD (2004) The relative contribution of glottal pulse rate and vocal tract length in size discrimination judgements. Poster presented at the British Society of Audiology meeting, Sept. London (available on the CNBH Website: <http://www.pdn.cam.ac.uk/cnbh/>).
- González, J (2004) Formant frequencies and body size of speaker: A weak relationship in adult humans. *J Phonet* 32:277–287.
- Goto M, Hashiguchi H, Nishimura T, Oka R (2003) RWC music database: Music genre database and musical instrument sound database. In ISMIR is International Symposium on Music Information Retrieval. pp. 229–230.
- Green DM (1988) *Profile Analysis*. London: Oxford University Press.
- Hollien H, Green R, Massey K (1994) Longitudinal research on adolescent voice change in males. *J Acoust Soc Am* 96:3099–3111.
- Huber JE, Stathopoulos ET, Curione GM, Ash T, Johnson K (1999) Formants of children, women and men: The effects of vocal intensity variation. *J Acoust Soc Am* 106:1532–1542.
- Irino T, Patterson RD (2002) Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilized wavelet-Mellin transform. *Speech Commun* 36:181–203.
- Ives DT, Smith, DRR, Patterson RD (2005) Discrimination of speaker size from syllable phrases. *J Acoust Soc Am* 118:3816–3822.
- Kawahara H, Irino T (2004) Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In: Divenyi P (ed) *Speech Segregation by Humans and Machines*. Dordrecht: Kluwer Academic, pp. 167–180.
- Kawahara H, Masuda-Kasuse I, de Cheveigne A (1999) Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency-based F_0 extraction: Possible role of repetitive structure in sounds. *Speech Commun* 27(3–4):187–207.
- Krumholz K, Patterson RD, Pressnitzer D (2000) The lower limit of pitch as determined by rate discrimination. *J Acoust Soc Am* 108:1170–1180.
- Künzel HJ (1989) How well does average fundamental frequency correlate with speaker height and weight? *Phonetica* 46:117–125.
- Ladefoged P, Broadbent DE (1957) Information conveyed by vowels. *J Acoust Soc Am* 29:98–104.
- Lass NJ, Brown WS (1978) Correlational study of speakers, heights, weights, body surface areas and speaking fundamental frequencies. *J Acoust Soc Am* 63:1218–1220.
- Leek MR, Dorman MF, Summerfield Q (1987) Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 81:148–154.
- Licklider JCR (1951) A duplex theory of pitch perception. *Experientia* 7:128–133.
- Liu C, Kewley-Port D (2004) STRAIGHT: A new speech synthesizer for vowel formant discrimination. *ARLO* 5:31–36.
- Lloyd RJ (1890) Speech sounds: Their nature and causation (I). *Phonetica Studien* 3:251–278.
- Marcus SM (1981) Acoustic determinants of perceptual centre (P-centre) location. *Percept Psychophys* 30:247–256.
- Meddis R, Hewitt MJ (1991) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J Acoust Soc Am* 89:2866–2882.
- Miller GA (1947) Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *J Acoust Soc Am* 19:609–619.

- Miller JD (1989) Auditory-perceptual interpretation of the vowel. *J Acoust Soc Am* 85:2114–2133.
- Owren MJ, Anderson JD (2005) Voices of athletes reveal only modest acoustic correlates of stature. *J Acoust Soc Am* 117:2375.
- Patterson RD (1994) The sound of a sinusoid: Time-interval models. *J Acoust Soc Am* 96:1419–1428.
- Patterson RD, Holdsworth J (1996) A functional model of neural activity patterns and auditory images. In: Ainsworth WA (ed) *Advances in Speech, Hearing and Language Processing*, Vol. 3, Part B. London: JAI Press.
- Patterson RD, Robinson K, Holdsworth J, McKeown D, Zhang C, Allerhand MH (1992) Complex sounds and auditory images. In: Cazals Y, Demany L, Horner K (eds) *Auditory Physiology and Perception, Proceedings of the 9th International Symposium on Hearing*. Oxford: Pergamon Press, pp. 429–446.
- Patterson RD, Allerhand M, 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, Anderson TR, Francis K (2006) Binaural auditory images for noise-resistant speech recognition. In: Ainsworth W, Greenberg S (eds) *Listening to Speech: An Auditory perspective*. The Publisher, LEA, is Lawrence Erlbaum Associates City is Mahwah, NJ pp. 257–269.
- Peterson GE, Barney HL (1952) Control methods used in the study of vowels. *J Acoust Soc Am* 24:175–184.
- Pressnitzer D, Patterson RD, Krumbholz K (2001) The lower limit of melodic pitch. *J Acoust Soc Am* 109:2074–2084.
- Rendall D, Vokey JR, Nemeth C, Ney C (2005) Reliable but weak voice-formant cues to body size in men but not women. *J Acoust Soc Am* 117:2372.
- Ritsma RJ, Hoekstra A (1974) Frequency selectivity and the tonal residue. In: Zwicker E, Terhardt E (eds) *Facts and Models in Hearing*. Berlin: Springer, pp. 156–163.
- Schouten JF (1938) The perception of subjective tones. *Proc Kon Ned Akad Wetensch* 41:1086–1093.
- Scott SK (1993) P-centres in speech an acoustic analysis. PhD thesis, University College London.
- Slaney M, Lyon RF (1990) A perceptual pitch detector. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, Albuquerque, New Mexico.
- 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 R, Kawahara H, Irino T (2005) The processing and perception of size information in speech sounds. *J Acoust Soc Am* 117:305–318.
- Spiegel MF, Picardi MC, Green DM (1981) Signal and masker uncertainty in intensity discrimination. *J Acoust Soc Am* 70:1015–1019.
- Sprague MW (2000) The single sonic twitch model for the sound production mechanism in the weakfish, *Cynoscion regalis*. *J Acoust Soc Am* 108:2430–2437.
- Terhardt E (1974) Pitch, consonance, and harmony. *J Acoust Soc Am* 55:1061–1069.
- Thurlow WR, Small AM Jr (1955) Pitch perception for certain periodic auditory stimuli. *J Acoust Soc Am* 27:132–137.
- Titze IR (1989) Physiologic and acoustic differences between male and female voices. *J Acoust Soc Am* 85:1699–1707.

- Turner RE, Al-Hames MA, Smith DRR, Kawahara H, Irino T, Patterson RD (2006) Vowel normalisation: Time-domain processing of the internal dynamics of speech. In: Divenyi P, Greenberg S, Meyer G. (eds) *Dynamics of Speech Production and Perception*. Amsterdam: IOS Press, pp. 153–170.
- van Dinther R, Patterson RD (2006) Perception of acoustic scale and size in musical instrument sounds. *J Acoust Soc Am* 120:2158–2176.
- Welling L, Ney H (2002) Speaker adaptive modelling by vocal tract normalization. *IEEE Trans Speech Audio Process* 10:415–426.
- Yang C-S, Kasuya H (1995) Dimension differences in the vocal tract shapes measured from MR images across boy, female and male subjects. *J Acoust Soc Jpn E* 16:41–44.
- 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.