Acoustic Methods
for Speech Science

Raymond D. Kent, PhD
Ferenc Bunta, PhD

Plural Publishing Inc.

Plural Publishing
Inc.

9177 Aero Drive, Suite B
San Diego, CA 92123

Email: information@pluralpublishing.com
Website: https://www.pluralpublishing.com

Typeset in 11/13 Adobe Garamond by Flanagan's Publishing Services, Inc.
Printed in the United States of America by Thomson Reuters

NOTICE TO THE READER
Care has been taken to confirm the accuracy of the indications, procedures, drug dosages, and diagnosis and remediation protocols presented in this book and to ensure that they conform to the practices of the general medical and health services communities. However, the authors, editors, and publisher are not responsible for errors or omissions or for any consequences from application of the information in this book and make no warranty, expressed or implied, with respect to the currency, completeness, or accuracy of the contents of the publication. The diagnostic and remediation protocols and the medications described do not necessarily have specific approval by the Food and Drug administration for use in the disorders and/or diseases and dosages for which they are recommended. Application of this information in a particular situation remains the professional responsibility of the practitioner. Because standards of practice and usage change, it is the responsibility of the practitioner to keep abreast of revised recommendations, dosages and procedures.

Library of Congress Cataloging-in-Publication Data

Names: Kent, Raymond D. author | Bunta, Ferenc author
Title: Acoustic methods for speech science / Raymond D. Kent, Ferenc Bunta.
Description: San Diego, CA : Plural Publishing, [2027] | Includes bibliographical references and index.
Identifiers: LCCN 2025047099 (print) | LCCN 2025047100 (ebook) | ISBN 9781635507904 paperback | ISBN 9781635505177 ebook
Subjects: LCSH: Phonetics, Acoustic | Speech processing systems | Communicative disorders—Patients—Rehabilitation
Classification: LCC P221.5 $.K47 2027 (print) | LCC P221.5 $.K47 2027 (ebook)
LC record available at https://lccn.loc.gov/2025047099
LC ebook record available at https://lccn.loc.gov/2025047100

Contents

Preface

"Speech Science is defined as the interdisciplinary study that involves phonetics, speech science, and engineering to understand the biological and physical aspects of speech production, with the aid of computer technology for data collection and analysis on a large scale." This definition, generated by artificial intelligence (AI) and based on an article by John Laver (2006), is a suitable introduction to this text—except that it is somewhat circular in its use of the term *speech science* as both the item to be defined and as part of the definition. Alternatively, we can define *speech science* as the interdisciplinary study of the process of speech communication, including the production, perception, and transmission of speech. As noted in the opening quotation, it is important to recognize the interdisciplinary nature of the scientific study of speech along with sweeping advances made possible by computers.

This book focuses on the acoustic aspects of speech science, including theories, methods, and applications. The acoustic signal of speech is a bridge between speech production and speech perception, so that knowledge of acoustics is key to a general understanding of speech communication. This text offers a systematic and practical account of how acoustic analysis can be applied in audiology, language teaching, linguistics, speech-language pathology, psychology, and other specialties. The text is designed for the general reader and does not assume an intensive background in mathematics or the physical sciences. Activities based on the freeware Praat enable the reader to perform acoustic analyses of speech with a personal computer. Step-by-step instructions and analysis guidelines pave the way to a hands-on experience with modern methods for acoustic measurements.

Learning acoustics is not a casual undertaking, but a systematic discussion of speech acoustics in this text makes for a manageable, and perhaps even satisfying, course of study. Tiwari (2012) wrote,

> In order to understand and evaluate the speech, it is important to have at least a basic understanding of science of speech acoustics: how the acoustics of speech are produced, how they are described, and how differences, both between speakers and within speakers, arise in an acoustic output. (p. 24)

This book covers these topics along with discussion of an emerging era in which AI is a partner in the study and application of speech science. In so doing, the book discusses a trajectory of discoveries from about 1940 to the present, while keeping an eye on the practicality of acoustic analysis in a computer-based world.

Watershed Events: A Brief History of Speech Acoustics

To understand the evolution of speech acoustics, it is helpful to review watershed events that have shaped contemporary understanding and applications.

1940s and 1950s. In 1942, Tsutomu Chiba and Masato Kajiyama published *The Vowel: Its Nature and Structure*, a monograph that described and modeled the production and perception of vowels by combining information from physiology, physics, and psychology. Unfortunately, this seminal publication was not widely distributed because of the Pacific war. However, it laid the foundation for the modern understanding of vowels and should be recognized for its technological prowess and theoretical insights. In 1947, Ralph K. Potter, George A. Kopp, and Harriet C. Green published *Visible Speech*, demonstrating how the sound spectrograph produces a visible representation of speech signals for the objective measurement of phonetic events. Before this advance, visual display of speech relied primarily on oscillograms (waveforms), which were often opaque to the underlying speech processes and phonetic elements. With the sound spectrograph, speech became visible in a way that transformed the field of acoustic phonetics. In 1952, Peterson and Barney published a classic paper that presented acoustic data on vowels produced by multiple speakers, including men, women, and children, demonstrating the effects of speaker age and sex on acoustic measurements. Peterson and Barney's data derived from a hardware spectrograph remain a standard source for acoustic measurements of vowels.

1960s. Gunnar Fant's *Acoustic Theory of Speech Production* was published in 1960. This classic book heralded a rigorous approach to understanding the acoustics of speech. It provided a mathematically based introduction of the source-filter theory, a dominant theory in speech acoustics. The article "Perception of the Speech Code" by Liberman et al. was published in 1967. This article, one of the most influential papers in the study of speech, galvanized the joint study of speech as an acoustic signal and a perceptual phenomenon. Liberman et al. proposed the motor theory of speech perception to account for the way in which the motor processes of speech production are key to relating speech acoustics to speech perception. Whatever opinion one may have regarding this theory, the paper stimulated considerable research interest and is frequently cited decades after its appearance.

1970s. Digital signal processing of the acoustic signal of speech was well underway by this time. With this innovation, speech signals could be recorded, stored, and analyzed with digital computers. Increased memory capacity and processing speed ushered in a new era in which dedicated hardware instruments such as the sound spectrograph became museum pieces. Acoustic analysis became ever more convenient, rapid, flexible, and accurate. Important advances included new analysis methods (e.g., linear predictive coding, cepstral analysis), natural-sounding speech synthesis, text-to-speech systems, reading machines for the blind, and talking toys such as Speak & Spell.

1980s. In 1980, Dennis H. Klatt published his article, "Software for a Cascade/Parallel Formant Synthesizer," which outlined the basic elements of parametric speech synthesis by computer. Famed theoretical physicist, Stephen Hawking, who lost his speech to amyotrophic lateral sclerosis, used Perfect Paul, the default male voice of DECTalk, which was the commercialized version of Klatt's synthesized speech, and chose to continue using it despite the availability of newer programs. Other achievements in this period were the use of hidden Markov models in automatic speech recognition (e.g., Dragon Dictate) and development of a cochlear implant (the earliest successful neural prosthesis) that uses speech feature processing (such as the Nucleus 22-channel cochlear implant system).

1990s. In 1998, Kenneth N. Stevens's book *Acoustic Phonetics* summarized the modern understanding of speech acoustics. The first version of Praat (Dutch for "talk") appeared in 1991. Praat (Boersma & Weenink, 2025) is the basic software used for activities in this book and is available for download at no cost. It is compatible with several operating systems, including Unix, Linux, Mac, and Microsoft Windows. Other technological advances in this period included the design of connectionist networks for automatic speech recognition, construction of large speech databases, use of speech recognition processing in hearing aids, development and marketing of low-cost speech analysis systems based on digital signal processing, and introduction of Furby, a toy that not only talks (in Furbish) but gradually seems to "learn" English words.

2000s. In 2006, Google announced Google Translate, a web-based translation service available for free. This early version operated as a statistical machine translation in which translations were generated from statistical models whose parameters were based on the analysis of bilingual text corpora.

2010s. In 2014, Amazon announced the virtual assistant technology Alexa with the smart speaker Echo as the launch platform. In 2016, Google Assistant was launched. Neural text-to-speech (TTS) is now used for speech synthesis based on deep learning. Unlike rule-based speech synthesis, this approach readily incorporates prosodic features, such as stress, intonation, and rhythm, which can be readily converted from one voice to another. Examples of this technology are WaveNet, Tacotron, and Deep Voice. These systems produce natural-sounding speech by generating audio waveforms directly from text. Deep learning is developed for automatic speech recognition and natural language processing. The Language Environment Analysis (LENA) system was introduced. It is a wearable audio recorder with software for automated vocal analysis of children's communicative interactions.

Speech Acoustics Today

The foregoing historical journey, admittedly highly selective from a vast array of achievements, brings us to the middle 2020s, where we stand at a remarkable point of progress and very likely at the threshold of an era in which AI will profoundly affect speech communication, facilitating human-machine interaction through speech, giving access to large speech databases representing speakers of multiple languages and dialects, and implementing powerful algorithms for various types of analysis to capture the nuances of speech, including its linguistic content, emotion, and speaker attributes. The use of machines to understand and produce speech is now ubiquitous, with ongoing improvements in accuracy, convenience, and naturalness. Spoken language, which some regard as the most important feature of human evolution, is now shared with machines that understand and produce speech.

Kent and Read (2002) wrote that, "Because the acoustic signal intermediates between a speaker's production of speech and a listener's perception of the speech, acoustic analysis helps in the understanding of both speech production and speech perception" (p. 2). This book, a revised and updated edition of Kent and Read, echoes and expands on the principle that the acoustic signal is a bridge between these two facets of speech. Understanding this bridge is central to modern speech science and its applications.

References

Boersma, P., & Weenink, D. (2025). *Praat* (Version 6.4.27) [Computer software]. http://www.fon.hum.uva.nl/praat

Chiba, T., & Kajiyama, M. (1942). *The vowel: Its nature and structure*. Tokyo-Kaiseikan.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton.

Kent, R. D., & Read, C. (2002). *Acoustic analysis of speech* (2nd ed.). Singular Publishing.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*(3), 971–995. https://doi.org/10.1121/1.383940

Laver, J. (2006). Speech. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd ed., pp. 636–648). Elsevier.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Potter, R. K., Kopp, G. A., & Green, H. C. (1947). *Visible speech*. D. Van Nostrand.

Stevens, K. N. (1998). *Acoustic phonetics* (Vol. 30). MIT Press.

Tiwari, M. (2012). Speech acoustics: How much science? *Journal of Natural Science, Biology, and Medicine*, *3*(1), 24–31. https://doi.org/10.4103/0976-9668.95942

About the Authors

Raymond D. Kent, PhD, is Professor Emeritus of Communicative Sciences and Disorders at the University of Wisconsin–Madison. His publications include more than 250 journal articles, book chapters, and reviews on various topics in speech science and speech pathology. He has authored or edited 18 books, including: *Clinical Phonetics*, *Intelligibility in Speech Disorders*, *Reference Manual for Communicative Sciences and Disorders: Speech-Language Pathology*, *The Speech Sciences*, *Handbook of Voice Quality Measurement*, *The MIT Encyclopedia of Communication Disorders*, and *Handbook on Children's Speech: Development, Disorders, and Variations*. He served as editor of the *Journal of Speech and Hearing Research*, associate founding editor of *Clinical Linguistics and Phonetics*, and associate editor of *Folia Phoniatrica et Logopaedica*. His awards include Honors of the American Speech-Language-Hearing Association; Docteur Honoris Causa from the Université de Montréal; Honorary Professor, The University of Queensland, Australia; Visiting Erskine Fellow, University of Canterbury, New Zealand; and an Honorary Doctorate from the University of Oulu, Finland.

Ferenc Bunta, PhD, is Professor of Communication Sciences and Disorders at the University of Houston and Adjunct Professor at Baylor College of Medicine. His publications include more than 50 journal articles, book chapters, and other scholarly works primarily on multilingual phonology and speech in typically developing children and their peers with hearing loss who use cochlear implants. His research has received funding from the National Institutes of Health/National Institute on Deafness and Other Communication Disorders, Department of Education, Cochlear Americas, and the Spencer Foundation. He has served as guest editor and editorial board member of scholarly journals, including the *Journal of Speech and Hearing Research*, and reviewed grant proposals for various agencies (including the European Research Council and National Science Foundation). He is also the founding director of the PhD program in Communication Sciences and Disorders at the University of Houston.

Chapter 1

Theoretical Foundations

Speech is a phenomenon that is studied from a variety of perspectives and often by individuals (such as speech-language pathologists, linguists, and foreign language instructors) whose first reaction may not be to look at it as a physical, acoustic phenomenon. Yet, the very existence of the speech signal is testament to the necessity of understanding acoustic phenomena by professionals who work with speech, including speech-language pathologists, audiologists, linguists, language teachers, and other communication specialists. This text is for everyone who works with speech and wants to learn through both reading and hands-on activities that illustrate the utility of acoustic analyses of speech.

Acoustics is a branch of physics concerned with the study of mechanical waves in gases, liquids, and solids. This book is concerned primarily with the acoustics of an air-filled tube, the vocal tract that extends from the larynx to the mouth or nares. A basic understanding of acoustic theory is requisite to performing acoustic analyses of speech that are valid and accurate. The relevant theory often is discussed in mathematical terms involving calculus and differential equations, but here we offer a less mathematical presentation that we hope captures the essential concepts with minimal reliance on mathematics. A major goal in this chapter is to reveal the theoretical foundations of modern methods for the acoustic analysis of speech.

Source-Filter Theory

This theory is of such fundamental importance that it is given first position in this book. Primary sources for a formal description are Fant (1960), Stevens and House (1961), and Stevens (2000). The **source-filter theory** (sometimes called the **acoustic theory of speech production**) models the acoustic signal of speech as resulting from a source and a filter. The source is either a voice source provided by the vibrating vocal folds in the larynx or a noise generated somewhere in the vocal tract. The filter is a set of resonances associated with a particular configuration of the articulatory system. Figure 1–1 identifies these primary components. Figure 1–1A shows a simplified version consisting only of a source and a vocal tract filter. Figure 1–1B adds another component, the radiation characteristic, which is the effect of transmitting sound from the mouth into the atmosphere. This transmission increases at frequencies above 300 to 500 Hz at a rate of 6 dB per octave (an octave corresponds to a doubling in frequency). The sound radiation is usually modeled as a simple spherical source of energy equal to the mouth volume velocity that transmits the energy uniformly in all directions (also called an *infinite baffle*). Another term for the filter is the **transfer function**, which is a mathematical function that models the output of a system

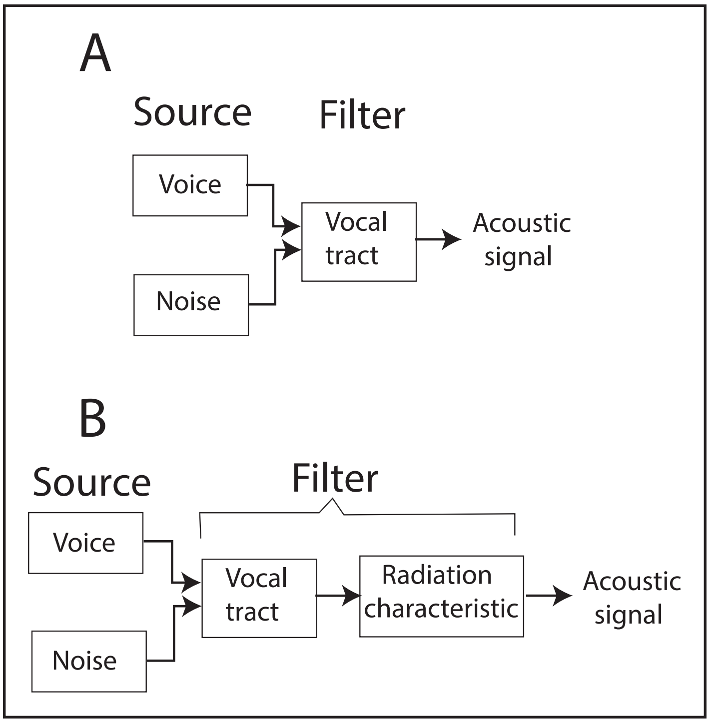

Figure 1–1. Basic components of the source-filter theory. **A.** A simple version. **B.** An elaborated version that includes the influences of both the vocal tract and the radiation characteristic.

for each admissible input (hence, an input–output function). For speech production, the transfer function includes both the vocal tract filtering and the radiation characteristic, as illustrated in Figure 1–1B.

For most purposes, the source and filter are regarded as independent and noninteractive, so that the behavior of the source is not affected by the behavior of the filter, or vice versa. For example, speakers can change the pitch of their voices without affecting the shape of the vocal tract for a given vowel. We can produce the vowel /i/ in the word *see* with a low pitch or higher pitch, a normal voice or breathy voice. However, in some circumstances, interaction can occur between source and filter. We neglect these effects for now, but note that they can be important (see Zhang, 2023, for extended discussion).

Figure 1–2 shows how the source-filter theory applies to vowels and voiceless fricatives. For vowels, vibrations of the vocal folds in the larynx are the energy source, and the entire vocal tract serves as the filter. For voiceless fricatives such as the /s/ in *see*, a constriction in the vocal tract generates noise as the energy source, and the portion of the vocal tract downstream from the source is the primary filter (neglecting for the moment the possible contribution of the cavity behind the noise source). For certain sounds, such as the voiced fricative in the word *zoo*, voicing and noise energy are combined.

The size and shape of the speech production system vary considerably from one indi-

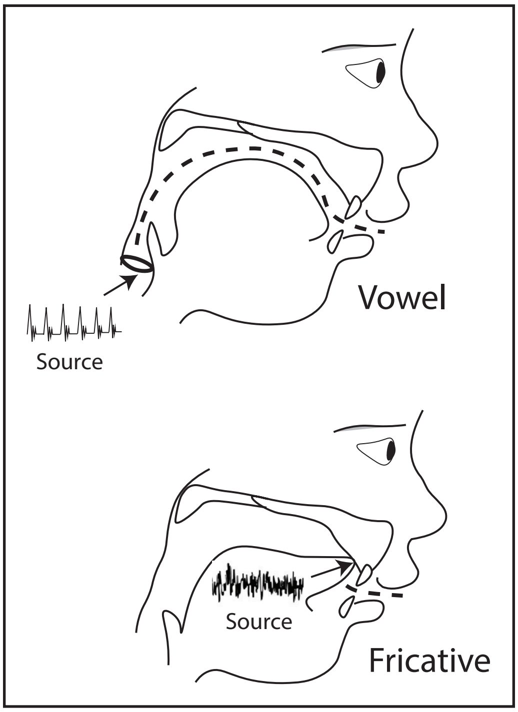

Figure 1–2. Source-filter theory for vowels and fricatives. For the vowel, the source is the vibration of the vocal folds. For the fricative, the source is a noise generated at a point in the vocal tract. Both the vowel and fricative activate resonances of the vocal tract.

vidual to another with age, sex, and other factors. Among the major factors contributing to the acoustic properties of speech are the sizes of the vocal folds and the vocal tract. As shown in Figure 1–3, the length of the vocal folds differs between children and adults and between men and women, on average. This figure also shows that the length of the vocal tract differs between children and adults and between men and women. Age and sex differences in speech acoustics are a recurrent theme in this book, and it is emphasized that many acoustic values depend on speaker age and gender as well as some other factors (such as the size of a person's vocal tract and vocal folds). Speech and voice are highly sexually dimorphic, and age figures prominently in acoustic values of many speech parameters. Therefore, typical reference values are often driven by age and sex (we are referring to biological sex, not gender identities). Historically, discussions of speech acoustics often have taken the adult male as the primary example, but this book has a wider view and presents examples and data from individuals of various groups of speakers. In addition, age and gender are important in selecting analysis methods and adjusting analysis parameters. Much more is said about this issue in forthcoming chapters. The essential point is that speech acoustics are not a "one-size-fits-all" proposition.

Frequency

Before expanding on the source-filter theory, the concept of **frequency** warrants introduction and explanation. As its name suggests, frequency is the number of times something occurs over time. For example, the frequency with which midnight occurs is once per day. The duration of the daily cycle is approximately 24 hr (or 86,400 s). This can be expressed in the following simple formula:

Equation 1–1

$$f = 1/t$$

where f is frequency in hertz, and t is time in seconds.

As evident from our example and equation, frequency is in inverse relationship with time. The longer the period of a cycle, the lower the frequency will be. The shorter the period, the higher is the frequency. Using our example of how frequently midnight occurs, its frequency is approximately 0.000011574 Hz (f = 1/86,400 s = 0.000011574 Hz). The frequency of the daily cycle is very low

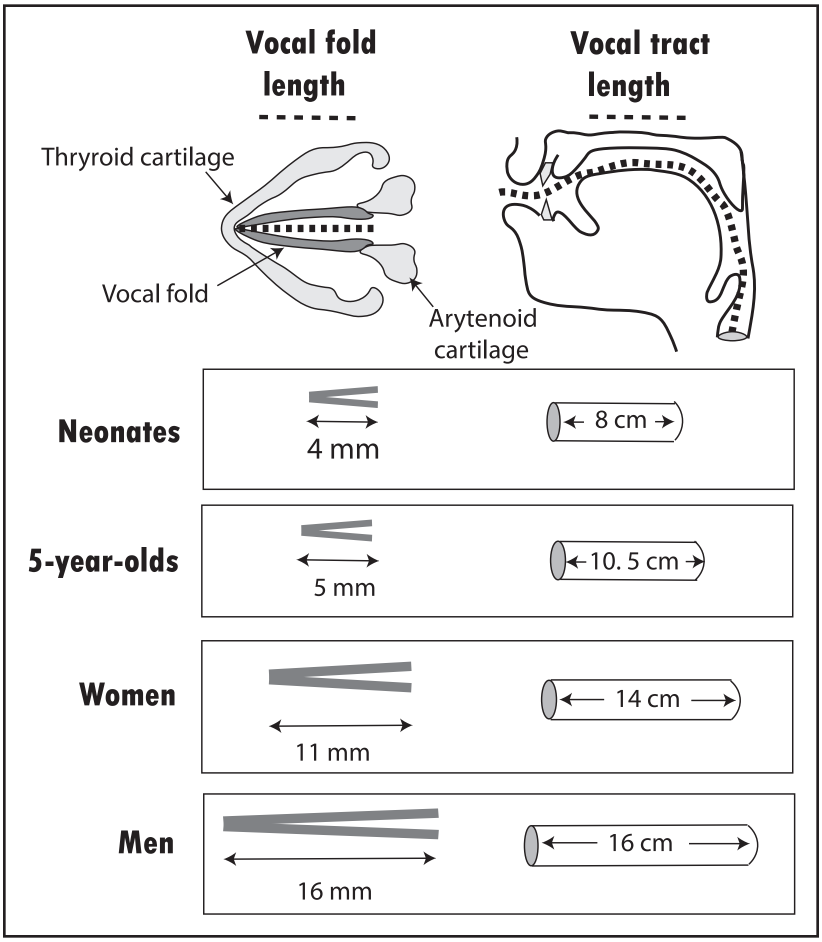

Figure 1–3. Age and gender differences in the lengths of the vocal folds and the vocal tract. Typical values are shown for a neonate, 5-year-old child, woman, and man.

compared to the frequency of speech sounds. For example, a young child may have a vocal fundamental frequency of 300 Hz when producing a voiced sound (such as a vowel), and the period of that frequency is approximately 0.00333 s. Compare the duration of a typical day that takes about 86,400 s to the duration of 0.00333 s for a 300-Hz sound. Figure 1–4 illustrates the measurement of a 300-Hz sine wave whose period is 0.00333 s. Later in this book, the activities, in particular Activity 8 (Harmonics and Resonance), provide hands-on exercises related to these issues.

Assumptions of the Source-Filter Theory

Applications of the source-filter theory generally assume that the system is linear, time invariant, and stationary. A system is **linear** if it conforms to the additivity principle, which in essence holds that a system response to two or more stimuli is the sum of the responses to each individual stimulus. A system is **time invariant** if its response to inputs is constant over time, so that any delay or advance in the input is matched by the same delay or advance

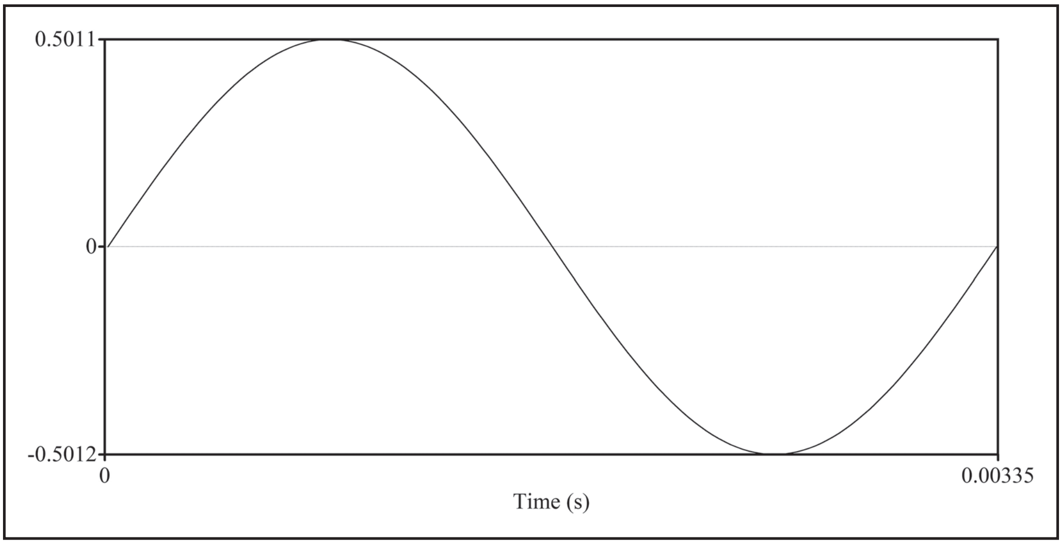

Figure 1–4. Duration of a 300-Hz sine wave is 0.00333 s.

in the output. A system is **stationary** if it has statistical properties that do not change over time. Stationary signals generally are produced by **linear time-invariant (LTI) systems**. These assumptions are important because LTI systems can be analyzed using powerful and readily available methods such as Fourier analysis, as discussed in Chapter 2. If a system does not conform to the LTI assumptions, then other mathematical methods should be used.

Source-Filter Theory for Vowels

We now consider in more detail the source-filter theory as it relates to vowels. In keeping with the goal of a diverse population of speakers, we consider in this example vowel production by a young child. Figure 1–5 shows the major components of the theory. The source of energy takes the form of a harmonic spectrum that is typical of voiced sounds such as vowels. That is, the energy is located at integer multiples of the fundamental frequency of the voice. A voiced source produced by a human is a complex sound with a fundamental frequency (f0) that is also the first harmonic, and at integer multiples of that f0 are the higher harmonics (second harmonic = 2 × f0, third harmonic = 3 × f0, etc.; Figure 1–6). For example, if f0 were 200 Hz, the second harmonic would be twice that or 400 Hz, the third harmonic would be 600 Hz, the fourth harmonic would be 800 Hz, and so on. If f0 were 300 Hz, the second harmonic would be 600 Hz, the third harmonic would be 900 Hz, and so on. For voiced sounds produced by humans, the energy of the source falls off at a rate of approximately 12 dB per octave, so most of the energy is in the lower frequencies, and therefore, lower harmonics have relatively more amplitude than higher harmonics (illustrated by Figure 1–6).

In the example provided in Figure 1–6, the value of f0, which is also the first harmonic, is 220 Hz, so that the successive harmonics have frequencies of 440, 660, 880 Hz, and so on. Because the energy of the source falls off at a rate of 12 dB per octave as previously noted, most of the energy is in the lower frequencies. This energy source activates resonances

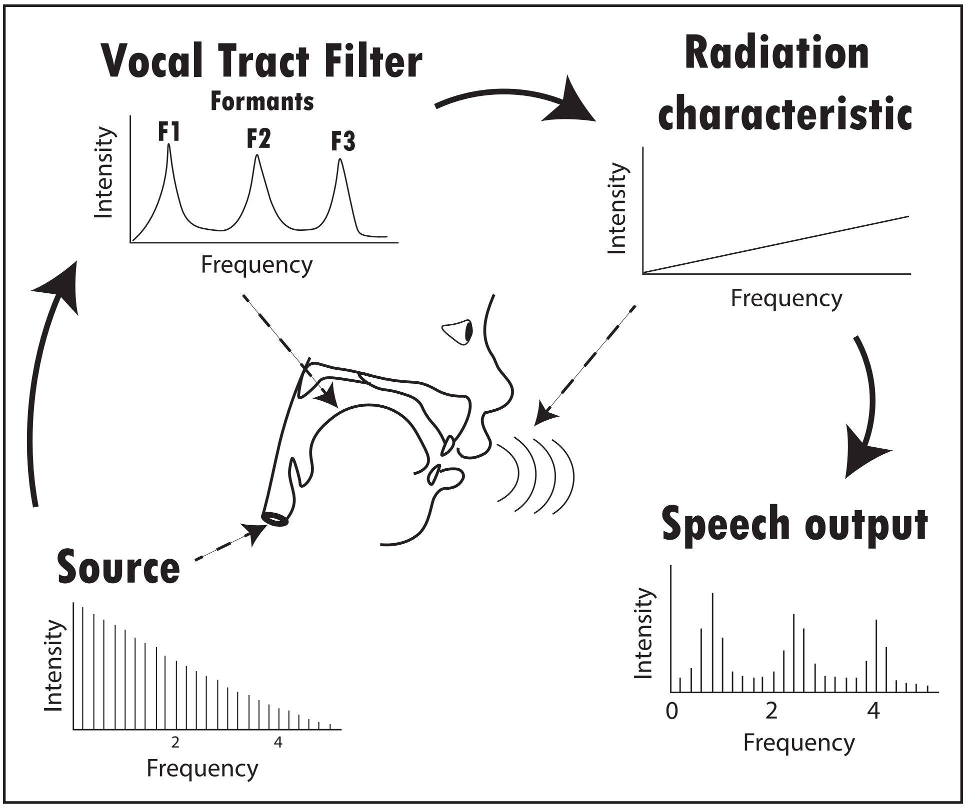

Figure 1–5. Source-filter theory of vowel acoustics, with a child as speaker.

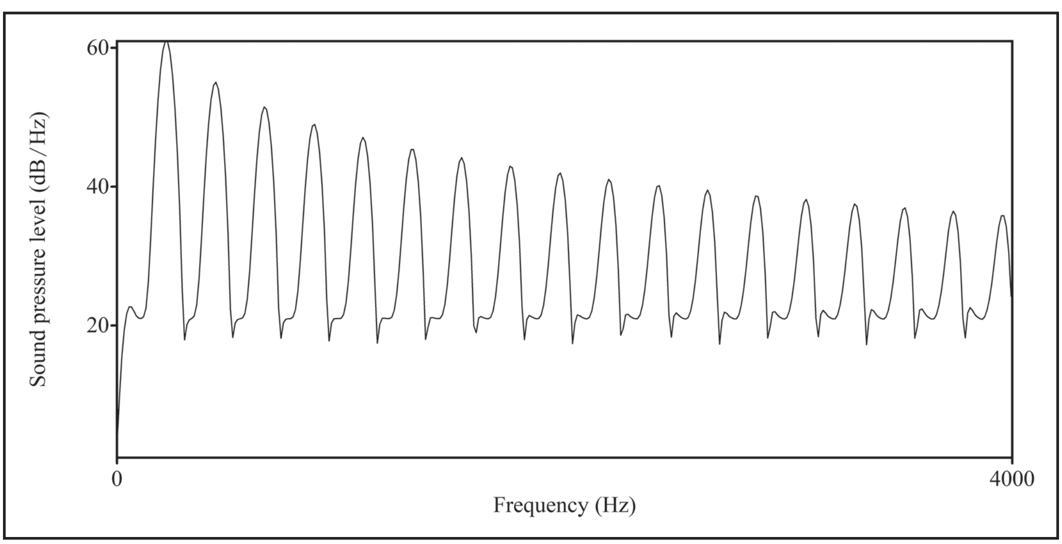

Figure 1–6. Simulation of a voiced sound source with a fundamental frequency of 220 Hz and its harmonics.

(called formants) associated with a specific vocal tract configuration. The word **formant** has somewhat different meanings in studies of speech production and perception. In this book, a formant is a resonance of the vocal tract, the presence of which can be inferred or estimated from acoustic analyses of the speech signal. Formants can serve as important cues to differentiating resonant sounds such as the vowels in the words *sea* versus *sue*. In engineering applications, a formant may also be called a **pole**. Formants define the filter (or transfer function) of the vocal tract. In the example shown in Figure 1–5, the formant frequencies are appropriate for a 5-year-old child. Formants have no energy in themselves but shape the speech sound by virtue of the filtering action imposed on an energy source. In other words, formants "respond to" or "reinforce" the source energy at certain frequencies, creating bands of energy that are critical for speech. The significance of formants in speech production and perception is expanded on in the chapters covering vowels and resonant consonants (such as liquids and glides; Chapters 3, 4, and 5). The effects of the filter are evident in the output spectrum in which the individual harmonics have different levels of intensity. As mentioned earlier in this chapter, account also must be taken of the radiation characteristic, modeled here as a 6-dB/octave increase over most of the frequency range.

The Simple Tube Model for One Vowel

We now consider a vocal tract model with a simple geometry and a simple pattern of resonance. The tube illustrated in Figure 1–7A is closed at one end, open at the other, and is 17.5 cm long. Like any tube of this nature, this tube will exhibit resonance under certain physical conditions. These conditions depend on the relation between the frequency of a sound source and the length of the tube. The relation is known as the **odd-quarter wavelength relation**. The **wavelength** of a sound is the distance traveled by a sine wave during one period of vibration. The wavelength therefore depends on two physical quantities: the frequency of vibration (inversely related to the period of vibration) and the speed of sound. Wavelength can be calculated using Equation 1–2.

Equation 1–2

$$\lambda = s/f$$

where λ is the wavelength in meters, s is the speed of sound in meters/second, and f is the frequency in hertz.

Equation 1–2 tells us that the wavelength of a sinusoid can be measured in physical space. For example, the wavelength of a 1-kHz tone is about 0.343 m ($\lambda = 343/1{,}000 = 0.343$ m). If this tone were played into a hallway and we could visualize its wave motion, then the same point on the wave would be encountered every 0.343 m. In this sense, the wavelength is a spatial phenomenon.

The odd-quarter wavelength relation describes the relation between a sound source and a resonating tube. To show how resonance is created, we use the situation shown in Figure 1–8. A sound source, a pure tone or sinusoid, is located just outside the open end of the tube. The tube has a length l, and the tone has a wavelength of $4l$ (i.e., the wavelength of the tone is exactly four times the length of the tube). The sinusoid can be conceptualized as a wave of condensation and a wave of rarefaction separated by one-half wavelength. Therefore, the opening of the tube will receive a train of condensations and rarefactions. In the diagram labeled 1, condensation C is just entering the tube with a particle motion from left to right. This condensation is followed