The Handbook of Speech Perception
Blackwell Handbooks in Linguistics

This outstanding multi-volume series covers all the major subdisciplines within linguistics today and, when complete, will offer a comprehensive survey of linguistics as a whole.

Already published:

*The Handbook of Child Language*
Edited by Paul Fletcher and Brian MacWhinney

*The Handbook of Phonological Theory*
Edited by John A. Goldsmith

*The Handbook of Contemporary Semantic Theory*
Edited by Shalom Lappin

*The Handbook of Sociolinguistics*
Edited by Florian Coulmas

*The Handbook of Phonetic Sciences*
Edited by William J. Hardcastle and John Laver

*The Handbook of Morphology*
Edited by Andrew Spencer and Arnold Zwicky

*The Handbook of Japanese Linguistics*
Edited by Natsuko Tsujimura

*The Handbook of Linguistics*
Edited by Mark Aronoff and Janie Rees-Miller

*The Handbook of Contemporary Syntactic Theory*
Edited by Mark Baltin and Chris Collins

*The Handbook of Discourse Analysis*
Edited by Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton

*The Handbook of Language Variation and Change*
Edited by J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes

*The Handbook of Historical Linguistics*
Edited by Brian D. Joseph and Richard D. Janda

*The Handbook of Language and Gender*
Edited by Janet Holmes and Miriam Meyerhoff

*The Handbook of Second Language Acquisition*
Edited by Catherine Doughty and Michael H. Long

*The Handbook of Bilingualism*
Edited by Tej K. Bhatia and William C. Ritchie

*The Handbook of Pragmatics*
Edited by Laurence R. Horn and Gregory Ward

*The Handbook of Applied Linguistics*
Edited by Alan Davies and Catherine Elder

*The Handbook of Speech Perception*
Edited by David B. Pisoni and Robert E. Remez
The Handbook of Speech Perception

Edited by

David B. Pisoni and Robert E. Remez
## Contents

<table>
<thead>
<tr>
<th>List of Contributors</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>ix</td>
</tr>
<tr>
<td><strong>Michael Studdert-Kennedy</strong></td>
<td></td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td><strong>David B. Pisoni and Robert E. Remez</strong></td>
<td></td>
</tr>
</tbody>
</table>

### Part I  Sensing Speech

1. Acoustic Analysis and Synthesis of Speech  
   **James R. Sawusch**  
   7
2. Perceptual Organization of Speech  
   **Robert E. Remez**  
   28
3. Primacy of Multimodal Speech Perception  
   **Lawrence D. Rosenblum**  
   51
4. Phonetic Processing by the Speech Perceiving Brain  
   **Lynne E. Bernstein**  
   79
5. Event-Related Evoked Potentials (ERPs) in Speech Perception  
   **Dennis L. Molfese, Alexandra P. Fonaryova Key, Mandy J. Maguire, Guy O. Dove, and Victoria J. Molfese**  
   99

### Part II  Perception of Linguistic Properties

6. Features in Speech Perception and Lexical Access  
   **Kenneth N. Stevens**  
   125
7. Speech Perception and Phonological Contrast  
   **Edward Flemming**  
   156
8. Acoustic Cues to the Perception of Segmental Phonemes  
   **Lawrence J. Raphael**  
   182
9. Clear Speech  
   **Rosalie M. Uchanski**  
   207
10. Perception of Intonation  
    **Jacqueline Vaissière**  
    236
11. Lexical Stress  
    **Anne Cutler**  
    264
12 Slips of the Ear  
Z. S. Bond  

Part III Perception of Indexical Properties  
13 Perception of Dialect Variation  
Cynthia G. Clopper and David B. Pisoni  
14 Perception of Voice Quality  
Jody Kreiman, Diana Vanlancker-Sidtis, and Bruce R. Gerratt  
15 Speaker Normalization in Speech Perception  
Keith Johnson  
16 Perceptual Integration of Linguistic and Nonlinguistic Properties of Speech  
Lynne C. Nygaard  

Part IV Speech Perception by Special Listeners  
17 Speech Perception in Infants  
Derek M. Houston  
18 Speech Perception in Childhood  
Amanda C. Walley  
19 Age-Related Changes in Spoken Word Recognition  
Mitchell S. Sommers  
20 Speech Perception in Deaf Children with Cochlear Implants  
David B. Pisoni  
21 Speech Perception following Focal Brain Injury  
William Badecker  
22 Cross-Language Speech Perception  
Núria Sebastián-Gallés  
23 Speech Perception in Specific Language Impairment  
Susan Ellis Weismer  

Part V Recognition of Spoken Words  
24 Spoken Word Recognition: The Challenge of Variation  
Paul A. Luce and Conor T. McLennan  
25 Probabilistic Phonotactics in Spoken Word Recognition  
Edward T. Auer, Jr. and Paul A. Luce  

Part VI Theoretical Perspectives  
26 The Relation of Speech Perception and Speech Production  
Carol A. Fowler and Bruno Galantucci  
27 A Neuroethological Perspective on the Perception of Vocal Communication Signals  
Timothy Q. Gentner and Gregory F. Ball  

Index
Edward T. Auer, Jr., House Ear Institute. Email: eauer@hei.org

William Badecker, Johns Hopkins University. Email: badecker@jhu.edu

Gregory F. Ball, Johns Hopkins University. Email: gball@jhu.edu

Lynne E. Bernstein, House Ear Institute. Email: lbernstein@hei.org

Z. S. Bond, Ohio University. Email: zbond1@ohiou.edu

Cynthia G. Clopper, Indiana University. Email: cclopper@indiana.edu

Anne C. Cutler, Max Planck Institute for Psycholinguistics. Email: anne.cutler@mpi.nl

Guy O. Dove, University of Louisville. Email: Guy.dove@louisville.edu

Edward Flemming, Stanford University. Email: flemming@csli.stanford.edu

Alexandra P. Fonaryova Key, University of Louisville. Email: Alexandra.key@louisville.edu

Carol A. Fowler, Haskins Laboratories. Email: cowler@haskins.yale.edu

Bruno Galantucci, Haskins Laboratories. Email: Bruno.galantucci@haskins.yale.edu

Timothy Q. Gentner, University of Chicago. Email: t-gentner@uchicago.edu

Bruce R. Gerratt, University of California – Los Angeles. Email: bgerratt@ucla.edu

Derek M. Houston, Indiana University School of Medicine. Email: dmhousto@indiana.edu
Contributors

Keith A. Johnson, Ohio State University. Email: kjohnson@ling.ohio-state.edu

Jody Kreiman, University of California – Los Angeles. Email: jkreiman@ucla.edu

Paul A. Luce, The State University of New York – Buffalo. Email: luce@buffalo.edu

Mandy J. Maguire, University of Louisville. Email: Mandy.maguire@louisville.edu

Conor T. McLennan, The State University of New York – Buffalo. Email: mclennan@buffalo.edu

Dennis L. Molfese, University of Louisville. Email: dlmolfese@mac.com

Victoria J. Molfese, University of Louisville. Email: tori@louisville.edu

Lynne C. Nygaard, Emory University. Email: lnygaard@emory.edu

David B. Pisoni, Indiana University. Email: Pisoni@indiana.edu

Lawrence J. Raphael, Adelphi University. Email: raphael@adelphi.edu

Robert E. Remez, Barnard College. Email: remez@columbia.edu

Lawrence D. Rosenblum, University of California – Riverside. Email: larry.rosenblum@verizon.net

James R. Sawusch, University at Buffalo. Email: jsawusch@acsu.buffalo.edu

Núria Sebastián-Gallés, Parc Cientific de Barcelona- Hospital de Sant Joan de Déu. Email: nsebastian@ub.edu

Mitchell S. Sommers, Washington University. Email: Msommers@artsci.wustl.edu

Kenneth N. Stevens, Massachusetts Institute of Technology. Email: stevens@speech.mit.edu

Rosalie M. Uchanski, CID at Washington University School of Medicine. Email: UchanskiR@ent.wustl.edu

Jacqueline Vaissière, Laboratoire de Phonetique et de Phonologique. Email: jacqueline.vaissiere@univ-paris3.fr

Diana Vanlancker-Sidtis, New York University. Email: Diana.vanlancker@nyu.edu

Amanda C. Walley, University of Alabama – Birmingham. Email: awalley@uab.edu

Susan Ellis Weismer, University of Wisconsin – Madison. Email: ellisweismer@facstaff.wisc.edu
Historically, the study of audition has lagged behind the study of vision, partly, no doubt, because seeing is our first sense, hearing our second. But beyond this, and perhaps more importantly, instruments for acoustic control and analysis demand a more advanced technology than their optic counterparts: having a sustained natural source of light, but not of sound, we had lenses and prisms long before we had sound generators and oscilloscopes. For speech, moreover, early work revealed that its key perceptual dimensions are not those of the waveform as it impinges on the ear (amplitude, time), but those of its time-varying Fourier transform, as it might appear at the output of the cochlea (frequency, amplitude, time). So it was only with the invention of instruments for analysis and synthesis of running speech that the systematic study of speech perception could begin: the sound spectrograph of R. K. Potter and his colleagues at Bell Telephone Laboratories in New Jersey during World War II, the Pattern Playback of Franklin Cooper at Haskins Laboratories in New York, a few years later. With these devices and their successors, speech research could finally address the first task of all perceptual study: definition of the stimulus, that is, of the physical conditions under which perception occurs.

Yet a reader unfamiliar with the byways of modern cognitive psychology who chances on this volume may be surprised that speech perception, as a distinct field of study, even exists. Is the topic not subsumed under general auditory perception? Is speech not one of many complex acoustic signals to which we are exposed, and do we not, after all, simply hear it? It is, of course, and we do. But due partly to the peculiar structure of the speech signal and the way it is produced, partly to the peculiar equivalence relation between speaker and hearer, we also do very much more.

To get a sense of how odd speech is, consider writing and reading. Speech is unique among systems of animal communication in being amenable to transduction into an alternative perceptuomotor modality. The more or less continuously varying acoustic signal of an utterance in any spoken language can be transcribed as a visual string of discrete alphabetic symbols, and can then be reproduced from that string by a reader. How we effect the transforms from analog signal to discrete message, and back again, and the nature of the percept that mediates these transforms are central problems of speech research.
Notice that without the alphabet as a means of notation, linguistics itself, as a field of study, would not exist. But the alphabet is not merely a convenient means of representing language; it is also the primary objective evidence for our intuition that we speak (and language achieves its productivity) by combining a few dozen discrete phonetic elements to form an infinite variety of words and sentences. Thus, the alphabet, recent though it is in human history, is not a secondary, purely cultural aspect of language. The inventors of the alphabet brought into consciousness previously unexploited segmental properties of speech and language, much as, say, the inventors of the bicycle discovered previously unexploited cyclic properties of human locomotion. The biological nature and evolutionary origins of the discrete phonetic categories represented by the alphabet are among many questions on which the study of speech perception may throw light.

To perceive speech is not merely to recognize the holistic auditory patterns of isolated words or phrases, as a bonobo or some other clever animal might do; it is to parse words from a spoken stream, and segments from a spoken word, at a rate of several scores of words per minute. Notice that this is not a matter of picking up information about an objective environment, about banging doors, passing cars, or even crying infants; it is a matter of hearers recognizing sound patterns coded by a conspecific speaker into an acoustic signal according to the rules of a natural language. Speech perception, unlike general auditory perception, is intrinsically and ineradicably intersubjective, mediated by the shared code of speaker and hearer.

Curiously, however, the discrete linguistic events that we hear (segments, syllables, words) cannot be reliably traced in either an oscillogram or a spectrogram. In a general way, their absence has been understood for many years as due to their manner of production: extensive temporal and spectral overlap, even across word boundaries, among the gestures that form neighboring phonetic segments. Yet how a hearer separates the more or less continuous flow into discrete elements is still far from understood. The lack of an adequate perceptual model of the process may be one reason why automatic speech recognition, despite half a century of research, is still well below human levels of performance.

The ear’s natural ease with the dynamic spectrotemporal patterns of speech contrasts with the eye’s difficulties: oscillograms are impossible, spectrograms formidably hard, to read – unless one already knows what they say. On the other hand, the eye’s ease with the static linear string of alphabetic symbols contrasts with the ear’s difficulties: the ear has limited powers of temporal resolution, and no one has ever devised an acoustic alphabet more efficient than Morse code, for which professional rates of perception are less than a tenth of either normal speech or normal reading. Thus, properties of speech that lend themselves to hearing (exactly what they are, we still do not know) are obstacles to the eye, while properties of writing that lend themselves to sight are obstacles to the ear.

Beyond the immediate sensory qualities of speech, a transcript omits much else that is essential to the full message. Most obvious is prosody, the systematic variations in pitch, loudness, duration, tempo, and rhythm across words, phrases, and sentences that convey a speaker’s intentions, attitudes, and feelings. What a transcript leaves out, readers put back in, as best they can. Some readers are so good at this that they become professional actors.
Certain prosodic qualities may be peculiar to a speaker’s dialect or idiolect, of which the peculiar segmental properties are also omitted from a standard transcript. What role, if any, these and other indexical properties (specifying a speaker’s sex, age, social status, person, and so on) may play in the perception of linguistic structure remains to be seen. I note only that, despite their unbounded diversity within a given language, all dialects and idiolects converge on a single phonology and writing system. Moreover, and remarkably, all normal speakers of a language can, in principle if not in fact, understand language through the artificial medium of print as quickly and efficiently as through the natural medium of speech.

Alphabetic writing and reading have no independent biological base; they are, at least in origin, parasitic on spoken language. I have dwelt on them here because the human capacity for literacy throws the biological oddity of speech into relief. Speech production and perception, writing and reading, form an intricate biocultural nexus at the heart of modern western culture. Thanks to over 50 years of research, superbly reviewed in all its diversity in this substantial handbook, speech perception offers the student and researcher a ready path into this nexus.

Michael Studdert-Kennedy
Haskins Laboratories
New Haven, Connecticut
To

Alvin M. Liberman

and

Peter W. Jusczyk
The major goal of *The Handbook of Speech Perception* is to present the research and theory that has guided our understanding of human speech perception. Over the last three decades, enormous theoretical and technical changes have occurred in perceptual research on speech. From its origins in psychophysical assessments of basic phonetic attributes in telecommunication systems, the research agenda has broadened in scope considerably over the years to encompass multisensory speech perception, speech perception with sensory prostheses, speech perception across the life span, speech perception in neuropathological disorders, as well as the study of the perception of linguistic, paralinguistic, and indexical attributes of speech. Growth in these diverse areas has spurred theoretical developments reflecting a variety of perspectives for explaining and modeling speech perception in its various manifestations. *The Handbook of Speech Perception* was conceived to provide a timely forum for the research community by presenting a collection of technical and theoretical accomplishments and challenges across the field of research.

The scope of the topics encompassed here matches the interdisciplinary nature of the research community that studies speech perception. This includes several neighboring fields: audiology, speech and hearing sciences, behavioral neuroscience, cognitive science, computer science and electrical engineering, linguistics, physiology and biophysics, and experimental psychology. We estimate that the chapters are accessible to non-specialists while also engaging to specialists. While *The Handbook of Speech Perception* takes a place among the many excellent companion volumes in the Blackwell series on language and linguistics, the topics collected here are motivated by the specific concerns of the perception of spoken language, and therefore it is unique in the series.

The 27 chapters are organized into six sections. Each chapter provides an informed and critical introduction to the topic under consideration by including: (1) a synthesis of current research and debate; (2) a narrative comprising clear examples and findings from the research literature and the author’s own research program; and (3) a look toward the future in terms of anticipated developments in the field.

In Part I, “Sensing Speech,” five chapters cover a wide range of foundational issues in the field. James Sawusch provides a technical summary of current
techniques for the analysis and synthesis of speech; Robert Remez discusses several theoretical problems about the perceptual organization of speech and how it differs from other auditory signals; Lawrence Rosenblum presents empirical and theoretical arguments for the primacy of multimodal speech perception; Lynne Bernstein discusses the neural substrates of speech perception; Dennis Molife et al. describe recent electrophysiological findings on speech perception and language development.

In Part II, “Perception of Linguistic Properties,” seven chapters survey the major areas of the field of human speech perception. Kenneth Stevens describes the role of linguistic features in speech perception and lexical access; Edward Flemming discusses the relations between speech perception and phonological contrast within an optimality theoretic framework; Lawrence Raphael provides a detailed summary of the major acoustic cues to segmental phonetic perception; Rosalie Uchanski describes the rapidly growing literature on the perception of clear speech; Jacqueline Vaissière provides an extensive review and interpretation of the contribution of intonation to speech perception; Anne Cutler describes the role of lexical stress in speech perception; and Zinny Bond discusses perceptual functions from the perspective of mishearing, or slips of the ear.

The four chapters in Part III focus on the “Perception of Indexical Properties” of speech. Cynthia Clopper and David Pisoni describe recent findings on the perception of dialect variation; Jody Kreiman, Diana Van Lancker-Sidtis and Bruce Gerratt present a summary and theoretical framework on the perception of voice quality; Keith Johnson discusses talker normalization in speech perception; and Lynne Nygaard reviews research on the integration of linguistic and non-linguistic properties of speech.

Part IV is concerned with “Speech Perception by Special Listeners.” Derek Houston offers a perspective on the development of speech perception in infancy; Amanda Walley provides an extensive review of speech perception in childhood; Mitchell Sommers describes recent findings on age-related changes in speech perception and spoken word recognition; David Pisoni reviews findings on the speech perception of deaf children with cochlear implants; William Badecker discusses speech perception following brain injury; Núria Sebastián-Gallés considers speech perception across languages; and Susan Ellis-Weismer examines the recent literature on speech perception in children with specific language impairment.

Part V presents two chapters on “Recognition of Spoken Words.” Paul Luce and Conor McLennan discuss the challenges of phonetic variation in word recognition; Edward Auer and Paul Luce examine the conceptualization of probabilistic phonotactics in word recognition.

The final section, Part VI, contains two chapters that present quite different “Theoretical Perspectives” on speech perception. Carol Fowler and Bruno Galantucci discuss the relation between speech perception and speech production while Timothy Gentner and Gregory Ball present a neuroethological perspective on the perception of vocal communication signals.

There are many decisions that face an editor in composing an ideal handbook, one that can be useful for the student and researcher alike. Early in our discussions, we understood that we would not be creating a comprehensive review of method and theory in research on speech perception. For one reason, technical methods and technical problems evolve rapidly as researchers explore one or
another opportunity. For another, the Annual Reviews already exist and can satisfactorily offer a snapshot of a field at a particular instant. Aiming higher, we asked each of the contributors to produce a lively essay expressing a point of view to introduce the reader to the major issues and findings in the field. The result is a broad-ranging and authoritative collection that articulates a perspective on exactly those critical questions that are likely to move a rapidly changing field of research.

The advent of a handbook can be viewed as a sign of growth and maturity of a discipline. The Handbook of Speech Perception brings the diverse field of speech perception together for the researcher who, while focusing on a specific aspect of speech perception, might desire a clearer understanding of the aims, methods, and prospects for advances across the field. In addition to the critical survey of developments across a wide range of research on human speech perception, we also imagine the Handbook facilitating the development of multi-disciplinary research on speech perception.

We cannot conclude without acknowledging the many individuals on whose creativity, knowledge, and cooperation this endeavor depended, namely, the authors whose essays compose The Handbook of Speech Perception. A venture of this scope cannot succeed without the conscientious care of a publisher to protect the project, and we have received the benefit of this attention from Blackwell’s Tami Kaplan and Sarah Coleman; thanks also to our copy-editor, Anna Oxbury. The skill and resourcefulness of Luis Hernandez was critical to the production of the work, and we are grateful for his timely good deeds on our behalf. And, for her extraordinary versatility and assiduousness in steering the authors and the editors to the finish line, we offer our sincere thanks to Darla Sallee. We also wish to acknowledge the valuable contributions of Cynthia Clopper and Susannah Levi who helped with the final proof of the entire book.

David B. Pisoni and Robert E. Remez
Bloomington and New York
Part I  Sensing Speech
1 Acoustic Analysis and Synthesis of Speech

JAMES R. SAWUSCH

1.1 Overview

The speech signal is the end point for speaking and the starting point for listening. While descriptions of language and language processes use terms like word, phrase, syllable, intonation, and phoneme, it is important to remember that these are explanatory constructs and not observable events. The observable events are the movements of the articulators and the resulting sound. Consequently, understanding the nature of speech sounds is critical to understanding both the mental processes of production and perception. It is also important to be able to create sounds that have particular acoustic qualities for studies of perception. The focus in this chapter is speech analysis and synthesis as an aid to understanding the processes of speech perception. In analysis, we seek to characterize the energy at each frequency at each point in time and whether the signal is periodic or aperiodic. These qualities are related to the processes and structures of articulation and may be exploited by the listener in perception. In synthesis, we seek to reproduce speech from a small set of values (parameters) that describe the desired articulatory or acoustic qualities of the signal. Our starting point will be the nature of the articulatory system and a characterization of how sound is produced and modified in speech.

1.2 The Speech Signal

In overview, the production of speech sounds involves an air source that passes through the vocal folds. The folds are either held open or vibrate. The “sound” (air flow) is then modified as it passes through the vocal tract. A representation of the human vocal tract, in cross-section, is shown in Figure 1.1(a). The net effect of this chain of events is the speech signal. This characterization of speech production is known as the source-filter model (see Fant, 1960) and is shown schematically in panels b, c, d, and e in Figure 1.1.
1.2.1 Source-filter

As the air stream from the lungs passes through the larynx, it can set the vocal folds vibrating. The rate at which the vocal folds vibrate is determined by their size and the muscle tension placed on them. Adult males generally have longer and more massive vocal folds than children, and adult females are intermediate. Like the strings on a piano, longer, more massive vocal folds produce a lower rate of vibration. Listeners hear this as a lower pitch voice. When the vocal folds are vibrating, we refer to the resulting speech signal as voiced. However, it is also possible for the talker to pass air through the larynx without causing the vocal folds to vibrate. In this case, the resulting speech signal is voiceless. A schematic representation of the spectrum (energy at different frequencies) of voicing is shown in Figure 1.1(b). There is energy at the fundamental frequency (the rate of vocal fold vibration) and at integer multiples of the fundamental (the harmonics).

The air stream then enters the vocal tract. One very simplified way to describe the vocal tract is as a series of tubes. In fluent speech, the talker moves the tongue, lips and jaw from one configuration to another to produce the sounds of speech. At any one point in time, the position of the tongue, lips, and jaw can be approximated as a series of one or more tubes of different lengths and uniform cross-sectional areas. This is shown schematically in Figure 1.1(a) where the tongue, lips, and jaw position are appropriate for the vowel /æ/ as in “bat.” The effect of the vocal tract on the air stream passing through it is to pass some frequencies and attenuate others. Every tube has a natural or resonant frequency, set by the length of the tube. Again, like a piano, the long tube (string) has a lower resonant frequency and the short tube (string) has a higher resonant frequency. The resonance characteristics of the vocal tract for the vowel /æ/ are shown schematically in Figure 1.1(c). The peaks in panel (c) represent the resonant frequencies. The
vocal tract thus shapes the air stream and the resulting sound represents the combined effects of the larynx (source) and the vocal tract (filter).

Finally, the sound radiates out of the vocal tract (the effect of which is shown in panel d) and results in the sound spectrum shown in panel (e). The sounds of speech are the result of a source (voiced or voiceless) that is passed through a filter (the vocal tract). The complex nature of the speech signal is due to the dynamic nature of the speech production process. The movement of the tongue, lips, and jaw means that the frequency composition of speech changes as the shape of the vocal tract changes. The bottom part of Figure 1.2 shows an example: the speech signal for the sentence “The bottle deposit is five cents.” The vertical scale is in units of pressure while the horizontal scale displays time.

1.2.2 The digital domain

Before proceeding, a few words are in order about the process of recording and converting the sound signal into a digital form. A microphone converts pressure in air to a voltage (the electrical equivalent). This is still a continuous, analog signal. Since most modern speech analysis and synthesis is done by computer, we need to convert the signal into a digital form that has sufficient fidelity to preserve the details of the signal that listeners use in speech recognition. This is accomplished by a process of sampling. At regular time intervals, the voltage of the signal is converted to a numerical (digital) form using an analog-to-digital converter. The two key aspects or parameters of this process are the sampling rate and the precision of the conversion (resolution).
The upper limit for human hearing is approximately 20 kHz. So, for high fidelity, the sampling rate needs to represent frequencies up through 20 kHz. To represent a frequency, we need a minimum of two values: one for the increase in pressure and one for the decrease in pressure. Without at least two values, we cannot capture and represent the change over time that corresponds to that frequency. Thus, our sampling rate has to be at least twice the highest frequency that we wish to represent. For human hearing, this means that our sampling rate must be at least twice 20 kHz. With a sampling rate of 44.1 kHz (used for compact disk recordings), we can capture and represent frequencies in the range of human hearing. While musical instruments produce harmonics at high frequencies, there are few acoustic qualities that are correlated with linguistic structure in human speech above 10 kHz. Most of the acoustic information that signals linguistic properties in voiced speech occurs below 5 kHz. Consequently, a sampling rate near 20 kHz is quite adequate for speech and most speech research related to perception uses sampling rates in the 10 kHz to 20 kHz range.

The second key parameter is the precision of the conversion between analog and digital formats. Again, using our CD example, each point in time is represented by a 16-bit number. This provides a range of numbers from \(-2^{15}\) to \(2^{15}\) or \(-32,768\) to \(32,767\) for representing the sound pressure at each point in time. This is sufficient to represent most of the dynamic range of human hearing and can faithfully represent and reproduce loud and soft sounds in the same recording. While the range of intensity in speech is not as large as in music, a 16-bit representation is a good choice to preserve the details of the signal such as the change from a soft /f/ or /θ/ to a vowel such as /aɪ/ (e.g. “five” in Figure 1.2).

One final note. When converting an analog signal such as speech into digital form, frequencies higher than one-half the sampling rate need to be removed (filtered) before the conversion to digital form. This is because these frequencies cannot be accurately represented. Unless the signal is filtered before sampling, these frequencies will appear in the resulting digital waveform as distortion at low frequencies (referred to as “aliasing”). Typically, a low-pass filter is used to remove the frequencies above one-half the sampling rate.

1.3 Acoustic Analysis

1.3.1 Fourier analysis

The starting point for speech analysis on a computer is the Fourier transform. Jean Baptiste Joseph Fourier (1768–1830) proposed that any periodic signal can be represented as the sum of a set of sinusoids (pure tones) with particular amplitudes and phases. The Discrete Fourier Transform (DFT) represents a mathematical means of determining the amplitudes and phases of a set of sinusoids that represent the frequency composition of a sound. Put another way, the DFT converts a function (the sound) in the time domain (change in pressure over time) to a function in the frequency domain (the spectrum) which represents the intensity at each frequency. Figure 1.3 shows a brief part of the acoustic signal from Figure 1.2 on the bottom and the corresponding power spectrum for this part of the signal on the top. If we take repeated short samples of the signal and
convert each into a spectrum, we can display the information in the form shown in the top half of Figure 1.2. This is called a sound spectrogram. Time is on the horizontal axis and frequency is on the vertical axis. The darkness in the graph represents intensity with white representing low intensity and black high intensity at that frequency and time. The dark concentrations of energy over time in Figure 1.2 are called formants and these represent the natural resonances of the vocal tract. The lowest frequency formant is termed the first formant (F1), the next is the second formant (F2), and so forth.

This brief description hides a wealth of important details about speech (see Childers, 2000; Flanagan, 1972; Oppenheim & Schafer, 1975). Some of these details are important, so we will expand on them. The first is that this particular analysis was done after the signal had been pre-emphasized. Pre-emphasis has the effect of tilting the spectrum so that the high frequencies are more intense. This process makes the higher frequency formants show up more clearly in the spectrogram. The second important detail is the duration of the acoustic signal that is transformed into a spectral representation. We will focus on durations that are appropriate for examining the acoustic consequences of articulation. There is a trade-off between the length of the signal and the frequency resolution in the

Figure 1.3  Spectral cross section at the release of the consonant /b/ in “bottle” in Figure 1.2.
spectrum. Time and frequency are reciprocals since frequency is change over time. If a long signal is examined, then we can get a very detailed picture in the frequency domain. That is, we achieve good frequency resolution. If the sample comes from a part of the speech sound where the articulators were moving rapidly, then the frequency composition of the sound would also be changing. However, because we are analyzing this part of the sound as a single entity, the rapid changes will be smeared over time and our detailed resolution spectrum will show the average intensity at each frequency during this time interval. The detailed frequency resolution comes at the cost of low temporal resolution. Figure 1.3 shows a relatively long stretch of our sentence (46.3 ms), centered at 200 ms, which is near the onset of the stop consonant /b/. The individual harmonics in the spectrum are resolved quite well. However, any rapid changes in the spectrum such as formant frequency transitions or the details of a release burst have been lost. A spectrum computed with a long temporal window such as this one is referred to as a narrow-band spectrum because it captures a detailed frequency resolution of the signal.

Figure 1.4 shows the same part of the sound as Figure 1.3. Here, a much shorter segment of 5.8 ms centered at the same point in time as in the previous

![Figure 1.4](image)

**Figure 1.4** Spectral cross section at the release of the consonant /b/ in “bottle” in Figure 1.2.
example has been selected. Because the duration of the sound segment is relatively short, the frequency resolution in the spectrum will be relatively coarse. This can be seen in the top panel of Figure 1.4. The individual harmonics of the spectrum are no longer resolved here. Only broad peaks representing multiple harmonics are present. However, if this 5.8 ms part of the signal were compared to the 5.8 ms before and the 5.8 ms after, we could determine the nature of any changes in the spectrum that would correspond to rapid movements of the speech articulators. This computation with a short temporal window is referred to as a wide-band spectrum. In Figure 1.2 the location of the spectral sections shown in Figures 1.3 and 1.4 is marked by the arrow at 200 ms. Both spectral sections represent the onset of voicing in the consonant /b/. The speech sound at this point contains a release burst followed by rapid formant transitions. These temporal changes are preserved in the wide-band spectrogram and spectral section (Figures 1.2 and 1.4). A narrow-band spectrogram of the sound would partly or completely obscure these rapid spectrum changes.

Our next issue is illustrated by the bell-shaped curve superimposed on the sound in the bottom panels in both Figures 1.3 and 1.4. This curve is referred to as a window function and represents the portion of the sound that will be transformed into the spectrum. The purpose of a window function is to determine which part of the sound will be transformed. In addition to the length of the window, which we have just described, there are also different shaped windows. The bell shaped curve in Figure 1.4 is called a Hamming window. The amplitude of the sound is adjusted at each point to reflect the height of the window function relative to the horizontal axis. The result is that the part of the sound at the center of the window is treated at its full original amplitude and sounds near the edge of the window are attenuated to near zero. A window that tapers at the beginning and end reduces distortion in the spectrum. A fuller treatment of this can be found in Saito and Nakata (1985; also Oppenheim & Schafer, 1975).

The last step is to repeat this process for the entire utterance: select a short portion of the sound, impose the window function, and process with the FFT. This procedure results in a spectrum like those shown in Figures 1.3 and 1.4 for each point in time. To ensure that our analysis does not miss any of the information in the signal, this is usually done so that successive segments of the sound overlap. In order to display the frequency information over time, the height on the graph in Figure 1.4 is represented on a black (intense) to white (quiet) scale and each of our samples represents the information for one point in time in Figure 1.2. The sound spectrogram shown in Figure 1.2 is based on a series of short (5.8 ms) time windows such as those in Figure 1.4 and is a wide-band spectrogram. In this display, the resonant frequencies of the vocal tract show up as the dark concentrations of energy over time. These are referred to as formants.

The fundamental frequency (F0) or voice pitch does not show up as a concentration of energy in this display. Rather, the fundamental appears as the alternating light and dark vertical striations that are seen most clearly in the vowels such as the /a/ of the word “bottle” in Figure 1.2. In order to understand why F0 appears this way, look at the bottom panel of Figure 1.4. Here, the 5.8 ms window is shown superimposed upon the first vocal pulse of the consonant /b/. The 5.8 ms segment is shorter in duration than one vocal pulse (which is approximately 10 ms). The overall amplitude of the signal in the window will change
with the precise positioning of the window on the waveform. When the window is positioned over a more intense part of the waveform, the resulting spectrum will have more energy and this appears as a darker vertical band. When the window is positioned over a portion of the waveform with less energy, the resulting spectrum will have less energy overall and a lighter vertical band will appear in the spectrogram. Since the intensity of the sound rises and falls with each vocal pulse, the alternating light and dark areas of the spectrogram recur at regular intervals corresponding to the fundamental frequency (rate of vocal fold vibration). Since the resolution in frequency is inversely proportional to the resolution in time (as described above), the short temporal window used in the wide-band spectrogram of Figure 1.2 does not resolve the individual harmonics of the male talker’s voice. Since the individual harmonics are not resolved, F0 does not appear as a separate horizontal dark band. The narrow-band spectrum of Figure 1.3 does resolve the individual harmonics.

The description thus far illustrates the choices that must be made in any speech analysis: sampling rate, window type, window duration. These factors determine the frequency and temporal resolution of the spectrum. Even with careful choices for these alternatives, the spectrogram has some limits. First, since our analysis is of the combined effects of the source and the filtering of the vocal tract, the concentrations of energy that we have called formants are not necessarily the same as the resonant frequencies of the vocal tract. From the standpoint of investigating the acoustic structure of speech and its implications for speech perception and production, this can be a problem. The question is one of what information we need from the sound. If it is the information available to the listener, then the spectrogram is a good first order approximation. A more detailed approximation would be based on an understanding of the workings of the ear and the neural representation of sound. However, if we need to measure the resonant frequencies of the vocal tract, then the spectrogram is limited because the spectrum represents the source as modified by the vocal tract rather than the vocal tract itself. A further complication is that an estimate of the bandwidth of the formants is sometimes also desired and is not easily measured from the spectrogram or a cross-sectional spectrum. The bandwidths of formants are useful in formant synthesizers (discussed later).

What is needed is a method of separating the influences of the source (excitation) from the filter (vocal tract). There are two widely used approaches that attempt to do this: Linear Predictive Coding and cepstral analysis. We will examine Linear Predictive Coding because of its widespread use in extracting formant frequencies and bandwidths for formant synthesis. For a treatment of cepstral analysis, the reader should see Deller, Hansen, and Proakis (1993, ch. 6) or Wakita (1996).

### 1.3.2 Linear Predictive Coding

In Linear Predictive Coding (LPC), the speech waveform is modeled (predicted) based on a source function and a transfer function. This approach treats human speech production as a source exciting or driving a set of resonators. However, the analogy is not quite exact. The glottal source in human voiced speech is more complex than the impulse excitation that is built into LPC so the transfer function
in LPC does contain influences of the glottis (see Atal & Hanauer, 1971). In spite of this caveat, LPC does produce a smoothed spectrum that can be used for estimating formant frequencies and bandwidths. In turn, this information can be used to remove the influence of the transfer function (vocal tract) from the speech signal in a process known as inverse filtering. This allows researchers to examine the nature of the source (see Klatt & Klatt, 1990; Price, 1989 for examples). The variant of LPC that is described here is known as the all-pole model (Markel & Gray, 1976). The transfer function can be thought of as a set of resonators or tubes. This neglects any influence of zeros (anti-resonators) where energy in the source is effectively damped or canceled by the transfer function. In human speech there are zeros in the transfer function for certain classes of speech sounds, such as nasals. In all-pole LPC, the effects of zeros are modeled by using additional poles (resonators). We will return to this limitation of LPC later.

Mathematically, LPC models the sound. Each sample in the waveform is predicted based on a linear combination of a set of immediately preceding samples. The first step in LPC is to determine the coefficients of the equation that yields the best prediction of the next sample of the waveform based on the set of previous sample points. This represents the transfer function in the time domain. Using an FFT, the time domain function is converted to the frequency domain. Figure 1.5 shows the resulting LPC spectrum for an 11.4 ms part of the waveform that is centered at the same point in the sound as the wide-band spectrum of Figure 1.4. The LPC shows a smoothed spectrum that makes it easier to measure formant frequencies and bandwidths. In Figure 1.5, the formant frequencies for the first five formants are shown by the dashed vertical lines through the peaks in the LPC spectrum. The formant bandwidth for F5 is shown as the distance (in frequency) between points that are 3 dB below the peak (center frequency of the formant) on either side of the peak. The mathematics for estimating the peak and the bandwidth are fairly straightforward (see Markel & Gray, 1976) and widely used in software for speech analysis.

This brief overview hides a wealth of important details about LPC analysis. For example, the analysis can be done asynchronously to the speech signal or pitch synchronously with one analysis frame per vocal pulse. The order of the LPC can be adjusted depending upon the source characteristics of the portion of the waveform with more coefficients used for voiced parts of the signal and fewer for voiceless parts. A fuller description of these choices and others is necessary to understanding the limits and uses of LPC, but it is beyond the bounds of this chapter. A good starting point for the adventurous (and mathematically inclined) reader is Markel and Gray (1976) or Deller et al. (1993). Childers (2000) provides a good technical overview with software to illustrate the process.

The LPC can be used to measure formant frequencies and bandwidths because it makes assumptions about the nature of the speech signal: Speech can be modeled accurately using an all-pole model excited by an impulse. Then, we separate the transfer function from the impulse source. This results in a temporal signal that has less influence of the source (glottis) and more clearly reflects the influence of the vocal tract than the original sound. Thus, the peaks in the spectrum of the predicted function are a reasonable approximation to the resonant frequencies of the vocal tract. In performing an LPC analysis, as in the basic spectral analysis, a sample of the waveform of a particular size is chosen and a windowing function
such as the Hamming window is imposed upon the speech segment. Also, like the basic spectral analysis, the speech signal is typically pre-emphasized to tilt the spectrum upward with increasing frequency. This increases the intensity of the higher frequency energy in the signal and leads to a more accurate estimation of the higher frequency formants.

One additional parameter must be chosen: the order of the LPC. The researcher must choose the number of poles to use in the model. Using too few poles will result in a formant being missed or having two closely spaced formants merged in the analysis. Using too many may result in spurious peaks in the estimated spectrum. The choice of the number of poles is reducible to the sampling rate of the sound (which determines the upper frequency limit of the spectrum) and the length of the talker’s vocal tract. For a male talker, Markel and Gray suggest that a good rule of thumb is the sampling rate in kHz plus 4 or 5. With a sampling rate of 10 kHz, 14 or 15 would be appropriate as the LPC order. The LPC spectrum in Figure 1.5 uses 14 poles with a sampling rate of 11,050 Hz. With an adult female, the vocal tract is shorter and the order of the LPC can be slightly less because fewer formants are present in the spectrum (see Atal & Hanauer, 1971 or Markel & Gray, 1976).