CHAPTER 1

Introduction

From human prehistory to the new media of the future, speech communication has been and will be the dominant mode of human social bonding and information exchange. The spoken word is now extended, through technological mediation such as telephony, movies, radio, television, and the Internet. This trend reflects the primacy of spoken communication in human psychology.

In addition to human-human interaction, this human preference for spoken language communication finds a reflection in human-machine interaction as well. Most computers currently utilize a *graphical user interface* (GUI), based on graphically represented interface objects and functions such as windows, icons, menus, and pointers. Most computer operating systems and applications also depend on a user's keyboard strokes and mouse clicks, with a display monitor for feedback. Today's computers lack the fundamental human abilities to speak, listen, understand, and learn. Speech, supported by other natural modalities, will be one of the primary means of interfacing with computers. And, even before speech-based interaction reaches full maturity, applications in home, mobile, and office segments are incorporating spoken language technology to change the way we live and work.

A spoken language system needs to have both speech recognition and speech synthesis capabilities. However, those two components by themselves are not sufficient to build a useful spoken language system. An understanding and dialog component is required to manage interactions with the user; and domain knowledge must be provided to guide the system's interpretation of speech and allow it to determine the appropriate action. For all these components, significant challenges exist, including robustness, flexibility, ease of integration, and engineering efficiency. The goal of building commercially viable spoken language systems has long attracted the attention of scientists and engineers all over the world. The purpose of this book is to share our working experience in developing advanced spoken language processing systems with both our colleagues and newcomers. We devote many chapters to systematically introducing fundamental theories and to highlighting what works well based on numerous lessons we learned in developing Microsoft's spoken language systems.

1.1. MOTIVATIONS

What motivates the integration of spoken language as the primary interface modality? We present a number of scenarios, roughly in order of expected degree of technical challenges and expected time to full deployment.

1.1.1. Spoken Language Interface

There are generally two categories of users who can benefit from adoption of speech as a control modality in parallel with others, such as the mouse, keyboard, touch-screen, and joystick. For novice users, functions that are conceptually simple should be directly accessible. For example, raising the voice output volume under software control on the desktop speakers, a conceptually simple operation, in some GUI systems of today requires opening one or more windows or menus, and manipulating sliders, check-boxes, or other graphical elements. This requires some knowledge of the system's interface conventions and structures. For the novice user, to be able to say *raise the volume* would be more direct and natural. For expert users, the GUI paradigm is sometimes perceived as an obstacle or nuisance and shortcuts are sought. Frequently these shortcuts allow the power user's hands to remain on the keyboard or mouse while mixing content creation with system commands. For example, an operator of a graphic design system for CAD/CAM might wish to specify a text formatting command while keeping the pointer device in position over a selected screen element.

Speech has the potential to accomplish these functions more powerfully than keyboard and mouse clicks. Speech becomes more powerful when supplemented by information streams encoding other dynamic aspects of user and system status, which can be resolved by the semantic component of a complete multimodal interface. We expect such multimodal interactions to proceed based on more complete user modeling, including speech, visual orientation, natural and device-based gestures, and facial expression, and these will be coordinated with detailed system profiles of typical user tasks and activity patterns.

In some situations you must rely on speech as an input or output medium. For example, with wearable computers, it may be impossible to incorporate a large keyboard. When driving, safety is compromised by any visual distraction, and hands are required for controlling the vehicle. The ultimate speech-only device, the telephone, is far more widespread than the PC. Certain manual tasks may also require full visual attention to the focus of the work. Finally, spoken language interfaces offer obvious benefits for individuals challenged with a variety of physical disabilities, such as loss of sight or limitations in physical motion and motor skills. Chapter 18 contains a detailed discussion on spoken language applications.

1.1.2. Speech-to-Speech Translation

Speech-to-speech translation has been depicted for decades in science fiction stories. Imagine questioning a Chinese-speaking conversational partner by speaking English into an unobtrusive device, and hearing real-time replies you can understand. This scenario, like the spoken language interface, requires both speech recognition and speech synthesis technology. In addition, sophisticated multilingual spoken language understanding is needed. This highlights the need for tightly coupled advances in speech recognition, synthesis, and understanding systems, a point emphasized throughout this book.

1.1.3. Knowledge Partners

The ability of computers to process spoken language as proficient as humans will be a landmark to signal the arrival of truly intelligent machines. Alan Turing [29] introduced his famous *Turing test*. He suggested a game, in which a computer's use of language would form the criterion for intelligence. If the machine could win the game, it would be judged intelligent. In Turing's game, you play the role of an interrogator. By asking a series of questions via a teletype, you must determine the identity of the other two participants: a machine and a person. The task of the machine is to fool you into believing it is a person by responding as a person to your questions. The task of the other person is to convince you the other participant is the machine. The critical issue for Turing was that using language as humans do is sufficient as an operational test for intelligence.

The ultimate use of spoken language is to pass the Turing test in allowing future extremely intelligent systems to interact with human beings as knowledge partners in all aspects of life. This has been a staple of science fiction, but its day will come. Such systems require reasoning capabilities and extensive world knowledge embedded in sophisticated search, communication, and inference tools that are beyond the scope of this book. We expect that spoken language technologies described in this book will form the essential enabling mechanism to pass the Turing test.

1.2. SPOKEN LANGUAGE SYSTEM ARCHITECTURE

Spoken language processing refers to technologies related to speech recognition, text-to-speech, and spoken language understanding. A spoken language system has at least one of the following three subsystems: a speech recognition system that converts speech into words, a text-to-speech system that conveys spoken information, and a spoken language understanding system that maps words into actions and that plans system-initiated actions.

There is considerable overlap in the fundamental technologies for these three subareas. Manually created rules have been developed for spoken language systems with limited success. But, in recent decades, data-driven statistical approaches have achieved encouraging results, which are usually based on modeling the speech signal using well-defined statistical algorithms that can automatically extract knowledge from the data. The data-driven approach can be viewed fundamentally as a pattern recognition problem. In fact, speech recognition, text-to-speech conversion, and spoken language understanding can all be regarded as pattern recognition problems. The patterns are either recognized during the runtime operation of the system or identified during system construction to form the basis of runtime generative models such as prosodic templates needed for text-to-speech synthesis. While we use and advocate the statistical approach, we by no means exclude the knowledge engineering approach from consideration. If we have a good set of rules in a given problem area, there is no need to use the statistical approach at all. The problem is that, at time of this writing, we do not have enough knowledge to produce a complete set of high-quality rules. As scientific and theoretical generalizations are made from data collected to construct data-driven systems, better rules may be constructed. Therefore, the rule-based and statistical approaches are best viewed as complementary.

1.2.1. Automatic Speech Recognition

A source-channel mathematical model described in Chapter 3 is often used to formulate speech recognition problems. As illustrated in Figure 1.1, the speaker's mind decides the source word sequence $\mathbf{W}$ that is delivered through his/her text generator. The source is passed through a noisy communication channel that consists of the speaker's vocal apparatus to produce the speech waveform and the speech signal processing component of the speech recognizer. Finally, the speech decoder aims to decode the acoustic signal $\mathbf{X}$ into a word sequence $\hat{\mathbf{W}}$, which is hopefully close to the original word sequence $\mathbf{W}$.

A typical practical speech recognition system consists of basic components shown in the dotted box of Figure 1.2. Applications interface with the decoder to get recognition results that may be used to adapt other components in the system. *Acoustic models* include the representation of knowledge about acoustics, phonetics, microphone and environment variability, gender and dialect differences among speakers, etc. *Language models* refer to a system's knowledge of what constitutes a possible word, what words are likely to co-occur, and in what sequence. The semantics and functions related to an operation a user may wish to perform may also be necessary for the language model. Many uncertainties exist in these areas, associated with speaker characteristics, speech style and rate, recognition of basic

speech segments, possible words, likely words, unknown words, grammatical variation, noise interference, nonnative accents, and confidence scoring of results. A successful speech recognition system must contend with all of these uncertainties. But that is only the beginning. The acoustic uncertainties of the different accents and speaking styles of individual speakers are compounded by the lexical and grammatical complexity and variations of spoken language, which are all represented in the language model.

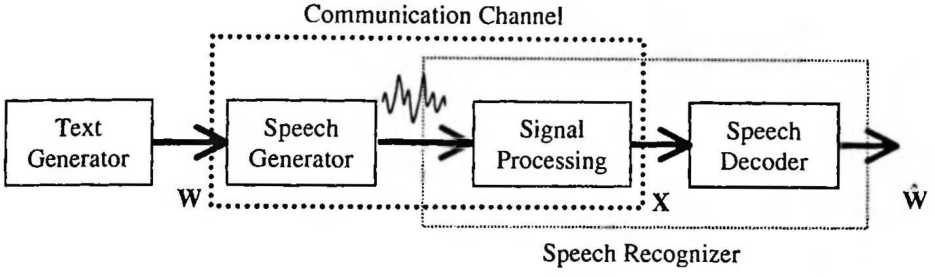

Figure 1.1 A source-channel model for a speech recognition system [15].

The speech signal is processed in the signal processing module that extracts salient feature vectors for the decoder. The decoder uses both acoustic and language models to generate the word sequence that has the maximum posterior probability for the input feature vectors. It can also provide information needed for the adaptation component to modify either the acoustic or language models so that improved performance can be obtained.

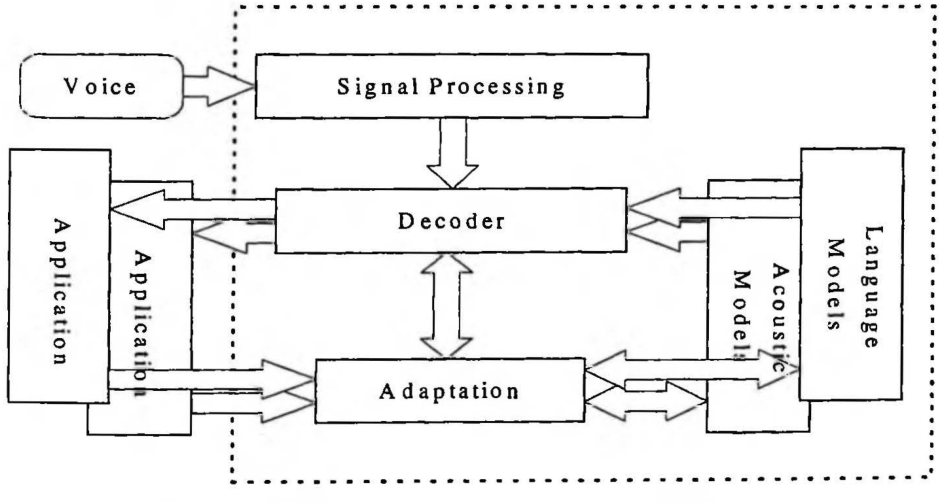

Figure 1.2 Basic system architecture of a speech recognition system [12].

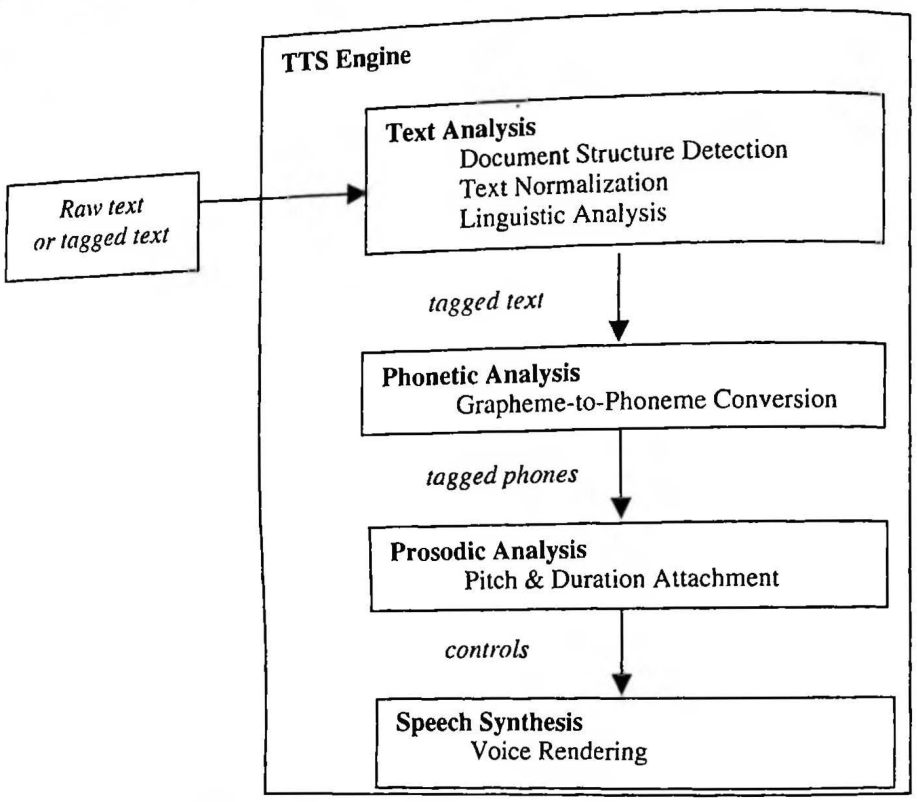

Figure 1.3 Basic system architecture of a TTS system.

1.2.2. Text-to-Speech Conversion

The term *text-to-speech*, often abbreviated as TTS, is easily understood. The task of a text-to-speech system can be viewed as speech recognition in reverse – a process of building a machinery system that can generate human-like speech from any text input to mimic human speakers. TTS is sometimes called *speech synthesis*, particularly in the engineering community.

The conversion of words in written form into speech is nontrivial. Even if we can store a huge dictionary for most common words in English; the TTS system still needs to deal with millions of names and acronyms. Moreover, in order to sound natural, the intonation of the sentences must be appropriately generated.

The development of TTS synthesis can be traced back to the 1930s when Dudley's *Voder*, developed by Bell Laboratories, was demonstrated at the World's Fair [18]. Taking advantage of increasing computation power and storage technology, TTS researchers have been able to generate high-quality commercial multilingual text-to-speech systems, although the quality is inferior to human speech for general-purpose applications.

The basic TTS components are shown in Figure 1.3. The text analysis component normalizes the text to the appropriate form so that it becomes speakable. The input can be

either raw text or tagged. These tags can be used to assist text, phonetic, and prosodic analysis. The phonetic analysis component converts the processed text into the corresponding phonetic sequence, which is followed by prosodic analysis to attach appropriate pitch and duration information to the phonetic sequence. Finally, the speech synthesis component takes the parameters from the fully tagged phonetic sequence to generate the corresponding speech waveform.

Various applications have different degrees of knowledge about the structure and content of the text that they wish to speak so some of the basic components shown in Figure 1.3 can be skipped. For example, some applications may have certain broad requirements such as rate and pitch. These requirements can be indicated with simple command tags appropriately located in the text. Many TTS systems provide a set of markups (tags), so the text producer can better express their semantic intention. An application may know a lot about the structure and content of the text to be spoken to greatly improve speech output quality. For engines providing such support, the *text analysis* phase can be skipped, in whole or in part. If the system developer knows the phonetic form, the phonetic analysis module can be skipped as well. The prosodic analysis module assigns a numeric duration to every phonetic symbol and calculates an appropriate pitch contour for the utterance or paragraph. In some cases, an application may have prosodic contours precalculated by some other process. This situation might arise when TTS is being used primarily for compression, or the prosody is *transplanted* from a real speaker's utterance. In these cases, the quantitative prosodic controls can be treated as special tagged field and sent directly along with the phonetic stream to speech synthesis for voice rendition.

1.2.3. Spoken Language Understanding

Whether a speaker is inquiring about flights to Seattle, reserving a table at a Pittsburgh restaurant, dictating an article in Chinese, or making a stock trade, a spoken language understanding system is needed to interpret utterances in context and carry out appropriate actions. Lexical, syntactic, and semantic knowledge must be applied in a manner that permits cooperative interaction among the various levels of acoustic, phonetic, linguistic, and application knowledge in minimizing uncertainty. Knowledge of the characteristic vocabulary, typical syntactic patterns, and possible actions in any given application context for both interpretation of user utterances and planning system activity are the heart and soul of any spoken language understanding system.

A schematic of a typical spoken language understanding system is shown in Figure 1.4. Such a system typically has a speech recognizer and a speech synthesizer for basic speech input and output, and a *sentence interpretation* component to parse the speech recognition results into semantic forms, which often need *discourse analysis* to track context and resolve ambiguities. The *Dialog Manager* is the central component that communicates with applications and the spoken language understanding modules such as discourse analysis, sentence interpretation, and response generation.

While most components of the system may be partly or wholly generic, the dialog manager controls the flow of conversation tied to the action. The dialog manager is respon-

sible for providing status needed for formulating responses, and maintaining the system's idea of the state of the discourse. The discourse state records the current transaction, dialog goals that motivated the current transaction, current objects in focus (temporary center of attention), the object history list for resolving dependent references, and other status information. The discourse information is crucial for sentence interpretation to interpret utterances in context. Various systems may alter the flow of information implied in Figure 1.4. For example, the dialog manager may be able to supply contextual discourse information or pragmatic inferences, as feedback to guide the recognizer's evaluation of hypotheses at the earliest level of search.

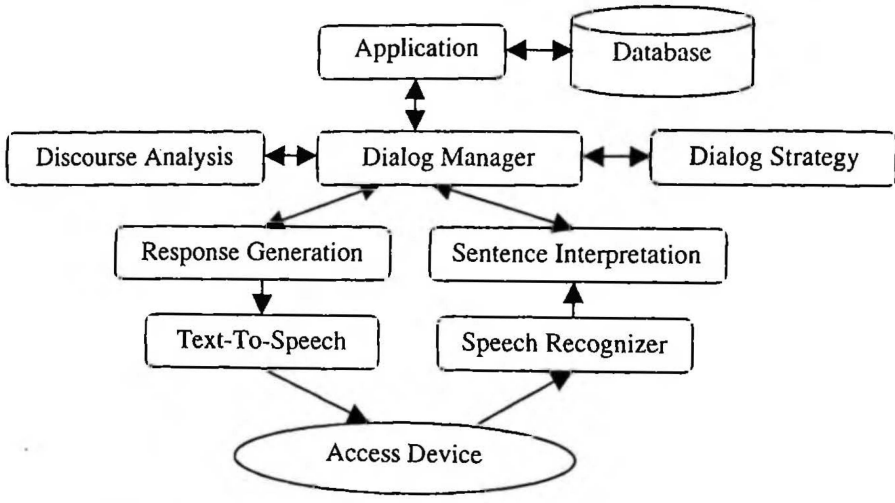

Figure 1.4 Basic system architecture of a spoken language understanding system.

1.3. BOOK ORGANIZATION

We attempt to present a comprehensive introduction to spoken language processing, which includes not only fundamentals but also a practical guide to build a working system that requires knowledge in speech signal processing, recognition, text-to-speech, spoken language understanding, and application integration. Since there is considerable overlap in the fundamental spoken language processing technologies, we have devoted Part I to the foundations needed. Part I contains background on speech production and perception, probability and information theory, and pattern recognition. Parts II, III, IV, and V include chapters on speech processing, speech recognition, speech synthesis, and spoken language systems, respectively. A reader with sufficient background can skip Part I, referring back to it later as needed. For example, the discussion of speech recognition in Part III relies on the pattern recognition algorithms presented in Part I. Algorithms that are used in several chapters

within Part III are also included in Parts I and II. Since the field is still evolving, at the end of each chapter we provide a historical perspective and list further readings to facilitate future research.

1.3.1. Part I: Fundamental Theory

Chapters 2 to 4 provide you with a basic theoretic foundation to better understand techniques that are widely used in modern spoken language systems. These theories include the essence of linguistics, phonetics, probability theory, information theory, and pattern recognition. These chapters prepare you fully to understand the rest of the book.

Chapter 2 discusses the basic structure of spoken language including speech science, phonetics, and linguistics. Chapter 3 covers probability theory and information theory, which form the foundation of modern pattern recognition. Many important algorithms and principles in pattern recognition and speech coding are derived based on these theories. Chapter 4 introduces basic pattern recognition, including decision theory, estimation theory, and a number of algorithms widely used in speech recognition. Pattern recognition forms the core of most of the algorithms used in spoken language processing.

1.3.2. Part II: Speech Processing

Part II provides you with necessary speech signal processing knowledge that is critical to spoken language processing. Most of what discuss here is traditionally the subject of electrical engineering.

Chapters 5 and 6 focus on how to extract useful information from the speech signal. The basic principles of digital signal processing are reviewed and a number of useful representations for the speech signal are discussed. Chapter 7 covers how to compress these representations for efficient transmission and storage.

1.3.3. Part III: Speech Recognition

Chapters 8 to 13 provide you with an in-depth look at modern speech recognition systems. We highlight techniques that have been proven to work well in real systems and explain in detail how and why these techniques work from both theoretic and practical perspectives.

Chapter 8 introduces hidden Markov models, the most prominent technique used in modern speech recognition systems. Chapters 9 and 11 deal with acoustic modeling and language modeling respectively. Because environment robustness is critical to the success of practical systems, we devote Chapter 10 to discussing how to make systems less affected by environment noises. Chapters 12 and 13 deal in detail with how to efficiently implement the decoder for speech recognition. Chapter 12 discusses a number of basic search algorithms, and Chapter 13 covers large vocabulary speech recognition. Throughout our discussion, Microsoft's Whisper speech recognizer is used as a case study to illustrate the methods introduced in these chapters.

1.3.4. Part IV: Text-to-Speech Systems

In Chapters 14 through 16, we discuss proven techniques in building text-to-speech systems. The synthesis system consists of major components found in speech recognition systems, except that they are in the reverse order.

Chapter 14 covers the analysis of written documents and the text needed to support spoken rendition, including the interpretation of audio markup commands, interpretation of numbers and other symbols, and conversion from orthographic to phonetic symbols. Chapter 15 focuses on the generation of pitch and duration controls for linguistic and emotional effect. Chapter 16 discusses the implementation of the synthetic voice, and presents algorithms to manipulate a limited voice data set to support a wide variety of pitch and duration controls required by the text analysis. We highlight the importance of trainable synthesis, with Microsoft's Whistler TTS system as an example.

1.3.5. Part V: Spoken Language Systems

As discussed in Section 1.1, spoken language applications motivate spoken language R&D. The central component is the spoken language understanding system. Since it is closely related to applications, we group it together with application and interface design.

Chapter 17 covers spoken language understanding. The output of the recognizer requires interpretation and action in a particular application context. This chapter details useful strategies for dialog management, and the coordination of all the speech and system resources to accomplish a task for a user. Chapter 18 concludes the book with a discussion of important principles for building spoken language interfaces and applications, including general human interface design goals, and interaction with other modalities in specific application contexts. Microsoft's MiPad is used as a case study to illustrate a number of issues in developing spoken language and multimodal applications.

1.4. TARGET AUDIENCES

This book can serve a variety of audiences:

Integration engineers: Software engineers who want to build spoken language systems, but who do not want to learn detailed speech technology internals, will find plentiful relevant material, including application design and software interfaces. Anyone with a professional interest in aspects of speech applications, integration, and interfaces can also achieve enough understanding of how the core technologies work, to allow them to take full advantage of state-of-the-art capabilities.

Speech technology engineers: Engineers and researchers working on various subspe cialties within the speech field will find this book a useful guide to understanding related technologies in sufficient depth to help them gain insight on where their own approaches overlap with, or diverge from, their neighbors' common practice.

Graduate students: This book can serve as a primary textbook in a graduate or advanced undergraduate speech analysis or language engineering course. It can serve as a sup-

plementary textbook in some applied linguistics, digital signal processing, computer science, artificial intelligence, and possibly psycholinguistics course.

Linguists: As the practice of linguistics increasingly shifts to empirical analysis of real-world data, students and professional practitioners alike should find a comprehensive introduction to the technical foundations of computer processing of spoken language helpful. The book can be read at different levels and through different paths, for readers with differing technical skills and background knowledge.

Speech scientists: Researchers engaged in professional work on issues related to normal or pathological speech may find this complete exposition of the state-of-the-art in computer modeling of generation and perception of speech interesting.

Business planners: Increasingly, business and management functions require some level of insight into the vocabulary and common practices of technology development. While not the primary audience, managers, marketers, and others with planning responsibilities and sufficient technical background will find portions of this book useful in evaluating competing proposals, and in making business decisions related to the speech technology components.

1.5. HISTORICAL PERSPECTIVE AND FURTHER READING

Spoken language processing is a diverse field that relies on knowledge of language at the levels of signal processing, acoustics, phonology, phonetics, syntax, semantics, pragmatics, and discourse. The foundations of spoken language processing lie in computer science, electrical engineering, linguistics, and psychology. In the 1970s an ambitious speech understanding project was funded by DARPA, which led to many seminal systems and technologies [17]. A number of human language technology projects funded by DARPA in the 1980s and 1990s further accelerated the progress, as evidenced by many papers published in *The Proceedings of the DARPA Speech and Natural Language/Human Language Workshop*. The field is still rapidly progressing and there are a number of excellent review articles and introductory books. We provide a brief list here. More detailed references can be found within each chapter of this book. Gold and Morgan's *Speech and Audio Signal Processing* [10] also has a strong historical perspective on spoken language processing.

Hyde [14] and Reddy [24] provided an excellent review of early speech recognition work in the 1970s. Some of the principles are still applicable to today's speech recognition research. Waibel and Lee assembled many seminal papers in *Readings in Speech Recognition Speech Recognition* [31]. There are a number of excellent books on modern speech recognition [1, 13, 15, 22, 23].

Where does the state of the art speech recognition system stand today? A number of different recognition tasks can be used to compare the recognition error rate of people vs. machines. Table 1.1 shows five typical recognition tasks with vocabularies ranging from 10 to 5000 words speaker-independent continuous speech recognition. The Wall Street Journal Dictation (WSJ) Task has a 5000-word vocabulary as a continuous dictation application for the WSJ articles. In Table 1.1, the error rate for machines is based on state of the art speech

recognizers such as systems described in Chapter 9, and the error rate of humans is based on a range of subjects tested on the similar task. We can see the error rate of humans is at least 5 times smaller than machines except for the sentences that are generated from a trigram language model, where the sentences have the perfect match between humans and machines so humans cannot use high-level knowledge that is not used in machines.[1]

Table 1.1 Word error rate comparisons between human and machines on similar tasks.

Tasks	Vocabulary	Humans	Machines
Connected digits	10	0.009%	0.72%
Alphabet letters	26	1%	5%
Spontaneous telephone speech	2000	3.8%	36.7%
WSJ with clean speech	5000	0.9%	4.5%
WSJ with noisy speech (10-db SNR)	5000	1.1%	8.6%
Clean speech based on trigram sentences	20,000	7.6%	4.4%

We can see that humans are far more robust than machines for normal tasks. The error rate for machine spontaneous conversational telephone speech recognition is above 35%, more than a factor 10 higher than humans on the similar task. In addition, the error rate of humans does not increase as dramatically as machines when the environment becomes noisy (from quiet to 10-db SNR environments on the WSJ task). The relative error rate of humans increases from 0.9% to 1.1% (1.2 times), while the error rate of CSR systems increases from 4.5% to 8.6% (1.9 times). One interesting experiment is that when we generated sentences using the WSJ trigram language model (cf. Chapter 11), the difference between humans and machines disappears (the last row in Table 1.1). In fact, the error rate of humans is even higher than machines. This is because both humans and machines have the same high-level syntactic and semantic models. The test sentences are somewhat random to humans but perfect to machines that used the same trigram model for decoding. This experiment indicates humans make more effective use of semantic and syntactic constraints for improved speech recognition in meaningful conversation. In addition, machines don't have attention problems as humans do on random sentences.

Fant [7] gave an excellent introduction to speech production. Early reviews of text-to-speech synthesis can be found in [3, 8, 9]. Sagisaka [26] and Carlson [6] provide more recent reviews of progress in speech synthesis. A more detailed treatment can be found in [19, 30].

Where does the state of the art text to speech system stand today? Unfortunately, like speech recognition, this is not a solved problem either. Although machine storage capabilities are improving, the quality remains a challenge for many researchers if we want to pass the Turing test.

[1] Some of these experiments were conducted at Microsoft with only a small number of human subjects (3-5 people), which is not statistically significant. Nevertheless, the experiments give some interesting insight on the performance of humans and machines.

Spoken language understanding is deeply rooted in speech recognition research. There are a number of good books on spoken language understanding [2, 5, 16]. Manning and Schutze [20] focuses on statistical methods for language understanding. Like Waibel and Lee, Grosz et al. assembled many foundational papers in *Readings in Natural Language Processing* [11]. More recent reviews of progress in spoken language understanding can be found in [25, 28]. Related spoken language interface design issues can be found in [4, 21, 27, 32].

In comparison to speech recognition and text to speech, spoken language understanding is further away from approaching the level of humans, especially for general-purpose spoken language applications.

A number of good conference proceedings and journals report the latest progress in the field. Major results on spoken language processing are presented at the *International Conference on Acoustics, Speech and Signal Processing (ICASSP), International Conference on Spoken Language Processing (ICSLP), Eurospeech Conference,* the *DARPA Speech and Human Language Technology Workshops,* and many workshops organized by the *European Speech Communications Associations (ESCA)* and *IEEE Signal Processing Society.* Journals include *IEEE Transactions on Speech and Audio Processing, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Computer Speech and Language, Speech Communication,* and *Journal of Acoustical Society of America (JASA).* Research results can also be found at computational linguistics conferences such as the *Association for Computational Linguistics (ACL), International Conference on Computational Linguistics (COLING),* and *Applied Natural Language Processing (ANLP).* The journals *Computational Linguistics* and *Natural Language Engineering* cover both theoretical and practical applications of language research. *Speech Recognition Update* published by TMA Associates is an excellent industry newsletter on spoken language applications.

REFERENCES

[1] Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, 1993, Boston, MA, Kluwer Academic Publishers.

[2] Allen, J., *Natural Language Understanding*, 2nd ed., 1995, Menlo Park, CA, The Benjamin/Cummings Publishing Company.

[3] Allen, J., M.S. Hunnicutt, and D.H. Klatt, *From Text to Speech: The MITalk System*, 1987, Cambridge, UK, University Press.

[4] Balentine, B., and D. Morgan, *How to Build a Speech Recognition Application*, 1999, Enterprise Integration Group.

[5] Bernsen, N., H. Dybkjar, and L. Dybkjar, *Designing Interactive Speech Systems*, 1998, Springer.

[6] Carlson, R., "Models of Speech Synthesis" in *Voice Communications Between Humans and Machines. National Academy of Sciences*, D.B. Roe and J.G. Wilpon, eds., 1994, Washington, D.C., National Academy of Sciences.

[7] Fant, G., *Acoustic Theory of Speech Production*, 1970, The Hague, NL, Mouton.

[8] Flanagan, J., *Speech Analysis Synthesis and Perception*, 1972, New York, Springer-Verlag.

[9] Flanagan, J., "Voices Of Men And Machines," *Journal of Acoustical Society of America*, 1972, **51**, p. 1375.

[10] Gold, B. and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, 2000, John Wiley and Sons.

[11] Grosz, B., F.S. Jones, and B.L. Webber, *Readings in Natural Language Processing*, 1986, Los Altos, CA, Morgan Kaufmann.

[12] Huang, X., *et al.*, "From Sphinx-II to Whisper -- Make Speech Recognition Usable" in *Automatic Speech and Speaker Recognition*, C.H. Lee, F.K. Soong, and K.K. Paliwal, eds. 1996, Norwell, MA, Kluwer Academic Publishers.

[13] Huang, X.D., Y. Ariki, and M.A. Jack, *Hidden Markov Models for Speech Recognition*, 1990, Edinburgh, U.K., Edinburgh University Press.

[14] Hyde, S.R., "Automatic Speech Recognition: Literature, Survey, and Discussion" in *Human Communication, A Unified Approach*, E.E. David and P.B. Denes, eds. 1972, New York, McGraw Hill.

[15] Jelinek, F., *Statistical Methods for Speech Recognition*, Language, Speech, and Communication, 1998, Cambridge, MA, MIT Press.

[16] Jurafsky, D. and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2000, Upper Saddle River, NJ, Prentice Hall.

[17] Klatt, D., "Review of the ARPA Speech Understanding Project," *Journal of Acoustical Society of America*, 1977, **62**(6), pp. 1324-1366.

[18] Klatt, D., "Review of Text-to-Speech Conversion for English," *Journal of Acoustical Society of America*, 1987, **82**, pp. 737-793.

[19] Kleijn, W.B. and K.K. Paliwal, *Speech Coding and Synthesis*, 1995, Amsterdam, Netherlands, Elsevier.

[20] Manning, C. and H. Schutze, *Foundations of Statistical Natural Language Processing*, 1999, MIT Press, Cambridge, USA.

[21] Markowitz, J., *Using Speech Recognition*, 1996, Prentice Hall.

[22] Mori, R.D., *Spoken Dialogues with Computers*, 1998, London, UK, Academic Press.

[23] Rabiner, L.R. and B.H. Juang, *Fundamentals of Speech Recognition*, May, 1993, Prentice-Hall.

[24] Reddy, D.R., "Speech Recognition by Machine: A Review," *IEEE Proc.*, 1976, **64**(4), pp. 502-531.

[25] Sadek, D. and R.D. Mori, "Dialogue Systems" in *Spoken Dialogues with Computers*, R.D. Mori, Editor 1998, London, UK, pp. 523-561, Academic Press.

[26] Sagisaka, Y., "Speech Synthesis from Text," *IEEE Communication Magazine*, 1990(1).

[27] Schmandt, C., *Voice Communication with Computers*, 1994, New York, NY, Van Nostrand Reinhold.

[28] Seneff, S., "The Use of Linguistic Hierarchies in Speech Understanding," *Int. Conf. on Spoken Language Processing*, 1998, Sydney, Australia.

[29] Turing, A.M., "Computing Machinery and Intelligence," *Mind*, 1950, **LIX**(236), pp. 433-460.

[30] van Santen, J., *et al.*, *Progress in Speech Synthesis*, 1997, New York, Springer-Verlag.

[31] Waibel, A.H. and K.F. Lee, *Readings in Speech Recognition*, 1990, San Mateo, CA, Morgan Kaufman Publishers.

[32] Weinschenk, S. and D. Barker, *Designing Effective Speech Interfaces*, 2000, John Wiley & Sons, Inc.

PART I

FUNDAMENTAL THEORY

CHAPTER 2

Spoken Language Structure

Spoken language is used to communicate information from a speaker to a listener. Speech production and perception are both important components of the speech chain. Speech begins with a thought and intent to communicate in the brain, which activates muscular movements to produce speech sounds. A listener receives it in the auditory system, processing it for conversion to neurological signals the brain can understand. The speaker continuously monitors and controls the vocal organs by receiving his or her own speech as feedback.

Considering the universal components of speech communication as shown in Figure 2.1, the fabric of spoken interaction is woven from many distinct elements. The speech production process starts with the semantic message in a person's mind to be transmitted to the listener via speech. The computer counterpart to the process of message formulation is the application semantics that creates the concept to be expressed. After the message is created,

the next step is to convert the message into a sequence of words. Each word consists of a sequence of phonemes that corresponds to the pronunciation of the words. Each sentence also contains a prosodic pattern that denotes the duration of each phoneme, intonation of the sentence, and loudness of the sounds. Once the language system finishes the mapping, the talker executes a series of neuromuscular signals. The neuromuscular commands perform articulatory mapping to control the vocal cords, lips, jaw, tongue, and velum, thereby producing the sound sequence as the final output. The speech understanding process works in reverse order. First the signal is passed to the cochlea in the inner ear, which performs frequency analysis as a filter bank. A neural transduction process follows and converts the spectral signal into activity signals on the auditory nerve, corresponding roughly to a feature extraction component. Currently, it is unclear how neural activity is mapped into the language system and how message comprehension is achieved in the brain.

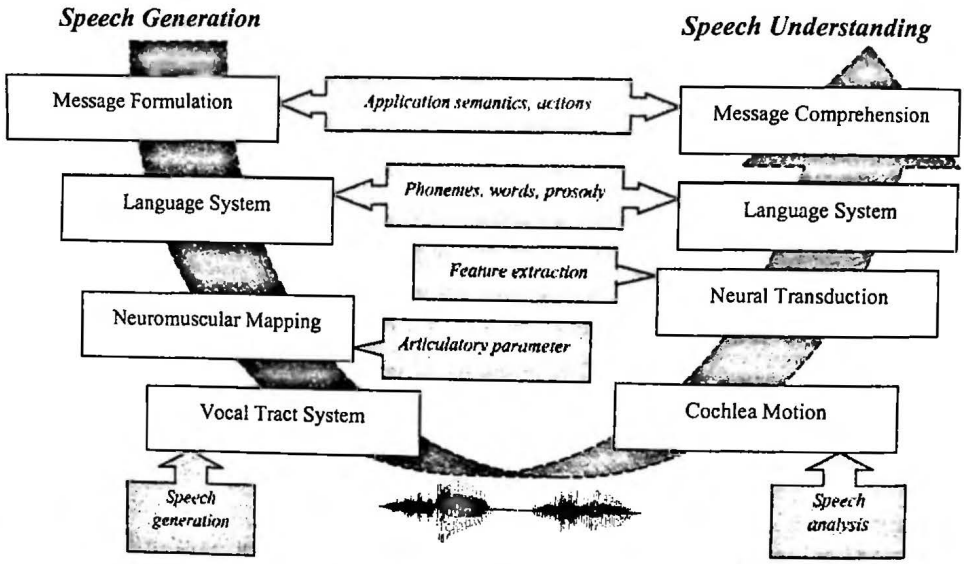

Figure 2.1 The underlying determinants of speech generation and understanding. The gray boxes indicate the corresponding computer system components for spoken language processing.

Speech signals are composed of analog sound patterns that serve as the basis for a discrete, symbolic representation of the spoken language – phonemes, syllables, and words. The production and interpretation of these sounds are governed by the syntax and semantics of the language spoken. In this chapter, we take a bottom up approach to introduce the basic concepts from sound to phonetics and phonology. Syllables and words are followed by syntax and semantics, which form the structure of spoken language processing. The examples in this book are drawn primarily from English, though they are relevant to other languages.

2.1. SOUND AND HUMAN SPEECH SYSTEMS

In this section, we briefly review human speech production and perception systems. We hope spoken language research will enable us to build a computer system that is as good as or better than our own speech production and understanding system.

2.1.1. Sound

Sound is a longitudinal pressure wave formed of compressions and rarefactions of air molecules, in a direction parallel to that of the application of energy. Compressions are zones where air molecules have been forced by the application of energy into a tighter-than-usual configuration, and rarefactions are zones where air molecules are less tightly packed. The alternating configurations of compression and rarefaction of air molecules along the path of an energy source are sometimes described by the graph of a sine wave as shown in Figure 2.2. In this representation, crests of the sine curve correspond to moments of maximal compression and troughs to moments of maximal rarefaction.

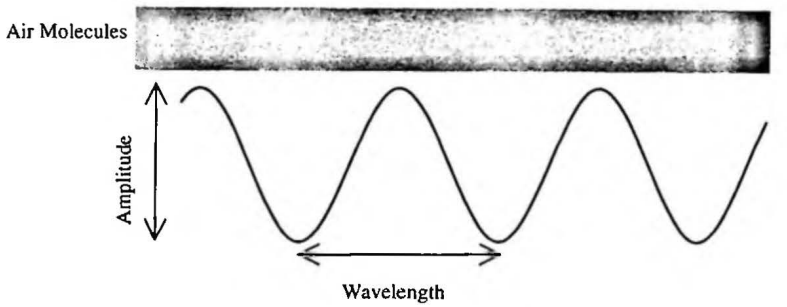

Figure 2.2 Application of sound energy causes alternating compression/rarefaction of air molecules, described by a sine wave. There are two important parameters, amplitude and wavelength, to describe a sine wave. Frequency [cycles/second measured in Hertz (Hz)] is also used to measure of the waveform.

The use of the sine graph in Figure 2.2 is only a notational convenience for charting local pressure variations over time, since sound does not form a transverse wave, and the air particles are just *oscillating in place* along the line of application of energy. The speed of a sound pressure wave in air is approximately $331.5 + 0.6T_c m/s$, where T_c is the Celsius temperature.

The amount of work done to generate the energy that sets the air molecules in motion is reflected in the amount of displacement of the molecules from their resting position. This *degree of displacement* is measured as the amplitude of a sound as shown in Figure 2.2. Because of the wide range, it is convenient to measure sound amplitude on a logarithmic scale in *decibels* (dB). A decibel scale is a means for comparing the intensity of two sounds:

$$10 \log_{10} (I / I_0) \tag{2.1}$$

where I and I_0 are the two intensity levels, with intensity being proportional to the square of the sound pressure P.

Sound pressure level (SPL) is a measure of absolute sound pressure P in dB:

$$SPL(dB) = 20 \log_{10} \left(\frac{P}{P_0} \right) \tag{2.2}$$

where the reference 0 dB corresponds to the threshold of hearing, which is $P_0 = 0.0002 \mu bar$ for a tone of 1kHz. The speech conversation level at 3 feet is about 60 dB SPL, and a jackhammer's level is about 120 dB SPL. Alternatively, watts/meter^2 units are often used to indicate intensity. We can bracket the limits of human hearing as shown in Table 2.1. On the low end, the human ear is quite sensitive. A typical person can detect sound waves having an intensity of 10^{-12} W/m^2 (the *threshold of hearing* or TOH). This intensity corresponds to a pressure wave affecting a given region by only one-billionth of a centimeter of molecular motion. On the other end, the most intense sound that can be safely detected without suffering physical damage is one trillion times more intense than the TOH. 0 dB begins with the TOH and advances logarithmically. The faintest audible sound is arbitrarily assigned a value of 0 dB, and the loudest sounds that the human ear can tolerate are about 120 dB.

Table 2.1 Intensity and decibel levels of various sounds.

Sound	dB Level	Times > TOH
Threshold of hearing (TOH: $10^{-12} W/m^2$)	0	10^0
Light whisper	10	10^1
Quiet living room	20	10^2
Quiet conversation	40	10^4
Average office	50	10^5
Normal conversation	60	10^6
Busy city street	70	10^7
Acoustic guitar – 1 ft. away	80	10^8
Heavy truck traffic	90	10^9
Subway from platform	100	10^{10}
Power tools	110	10^{11}
Pain threshold of ear	120	10^{12}
Airport runway	130	10^{13}
Sonic boom	140	10^{14}
Permanent damage to hearing	150	10^{15}
Jet engine, close up	160	10^{16}
Rocket engine	180	10^{18}
Twelve ft. from artillery cannon muzzle ($10^{10} W/m^2$)	220	10^{22}

The absolute threshold of hearing is the maximum amount of energy of a pure tone that cannot be detected by a listener in a noise free environment. The absolute threshold of hearing is a function of frequency that can be approximated by

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \quad (dB\ SPL) \tag{2.3}$$

and is plotted in Figure 2.3.

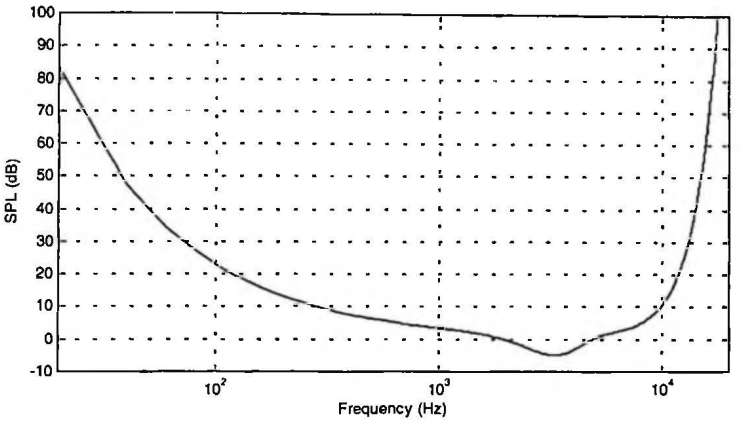

Figure 2.3 The sound pressure level (SPL) level in dB of the absolute threshold of hearing as a function of frequency. Sounds below this level are inaudible. Note that below 100 Hz and above 10 kHz this level rises very rapidly. Frequency goes from 20 Hz to 20 kHz and is plotted in a logarithmic scale from Eq. (2.3).

Let's compute how the pressure level varies with distance for a sound wave emitted by a point source located a distance r away. Assuming no energy absorption or reflection, the sound wave of a point source is propagated in a spherical front, such that the energy is the same for the sphere's surface at all radius r. Since the surface of a sphere of radius r is $4\pi r^2$, the sound's energy is inversely proportional to r^2, so that every time the distance is doubled, the sound pressure level decreases by 6 dB. For the point sound source, the energy (E) transported by a wave is proportional to the square of the amplitude (A) of the wave and the distance (r) between the sound source and the listener:

$$E \propto \frac{A^2}{r^2} \tag{2.4}$$

The typical sound intensity of a speech signal one inch away (close-talking microphone) from the talker is 1 Pascal = 10μbar, which corresponds to 94 dB SPL. The typical sound intensity 10 inches away from a talker is 0.1 Pascal = 1μbar, which corresponds to 74 dB SPL.

2.1.2. Speech Production

We review here basic human speech production systems, which have influenced research on speech coding, synthesis, and recognition.

2.1.2.1. Articulators

Speech is produced by air-pressure waves emanating from the mouth and the nostrils of a speaker. In most of the world's languages, the inventory of *phonemes,* as discussed in Section 2.2.1, can be split into two basic classes:

- consonants – articulated in the presence of constrictions in the throat or obstructions in the mouth (tongue, teeth, lips) as we speak.
- vowels – articulated without major constrictions and obstructions.

The sounds can be further partitioned into subgroups based on certain articulatory properties. These properties derive from the anatomy of a handful of important articulators and the places where they touch the boundaries of the human vocal tract. Additionally, a large number of muscles contribute to articulator positioning and motion. We restrict ourselves to a schematic view of only the major articulators, as diagrammed in Figure 2.4. The

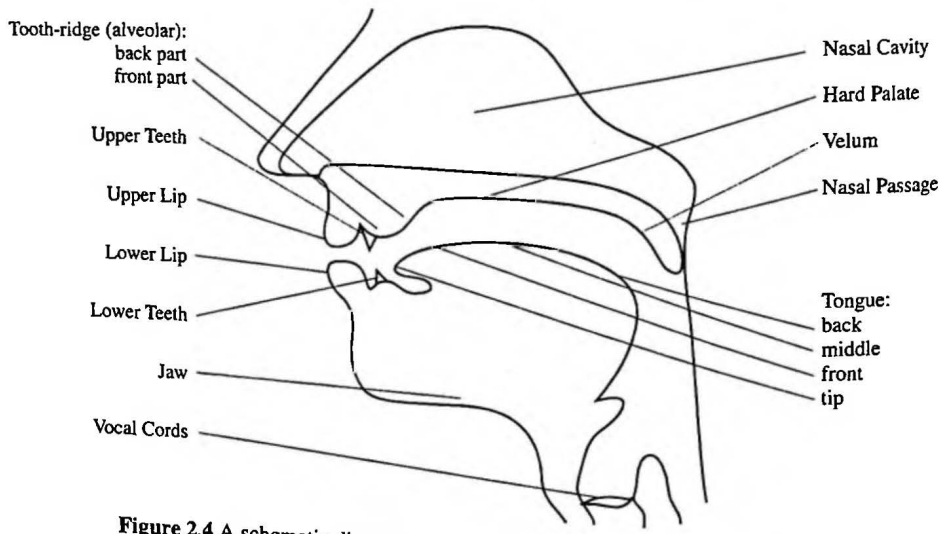

Figure 2.4 A schematic diagram of the human speech production apparatus.

gross components of the speech production apparatus are the lungs, trachea, larynx (organ of voice production), pharyngeal cavity (throat), oral and nasal cavity. The pharyngeal and oral cavities are typically referred to as the vocal tract, and the nasal cavity as the nasal tract. As illustrated in Figure 2.4, the human speech production apparatus consists of:

- *Lungs*: source of air during speech.
- *Vocal cords (larynx):* when the vocal folds are held close together and oscillate against one another during a speech sound, the sound is said to be *voiced.* When the folds are too slack or tense to vibrate periodically, the sound is said to be *unvoiced.* The place where the vocal folds come together is called the *glottis.*
- *Velum (soft palate)*: operates as a *valve*, opening to allow passage of air (and thus resonance) through the nasal cavity. Sounds produced with the flap open include *m* and *n*.
- *Hard palate*: a long relatively hard surface at the roof inside the mouth, which, when the tongue is placed against it, enables consonant articulation.
- *Tongue*: flexible articulator, shaped away from the palate for vowels, placed close to or on the palate or other hard surfaces for consonant articulation.
- *Teeth*: another place of articulation used to brace the tongue for certain consonants.
- *Lips*: can be rounded or spread to affect vowel quality, and closed completely to stop the oral air flow in certain consonants (*p, b, m*).

2.1.2.2. The Voicing Mechanism

The most fundamental distinction between sound types in speech is the voiced/voiceless distinction. Voiced sounds, including vowels, have in their time and frequency structure a roughly regular pattern that voiceless sounds, such as consonants like *s*, lack. Voiced sounds typically have more energy as shown in Figure 2.5. We see here the waveform of the word *sees*, which consists of three phonemes: an unvoiced consonant */s/*, a vowel */iy/*, and a voiced consonant */z/*.

What in the speech production mechanism creates this fundamental distinction? When the vocal folds vibrate during phoneme articulation, the phoneme is considered voiced; otherwise it is unvoiced. Vowels are voiced throughout their duration. The distinct vowel *timbres* are created by using the tongue and lips to shape the main oral resonance cavity in different ways. The vocal folds vibrate at slower or faster rates, from as low as 60 cycles per second (Hz) for a large man, to as high as 300 Hz or higher for a small woman or child. The rate of cycling (opening and closing) of the vocal folds in the larynx during phonation of voiced sounds is called the *fundamental frequency*. This is because it sets the periodic baseline for all higher-frequency harmonics contributed by the pharyngeal and oral resonance

cavities above. The fundamental frequency also contributes more than any other single factor to the perception of *pitch* (the semi-musical rising and falling of voice tones) in speech.

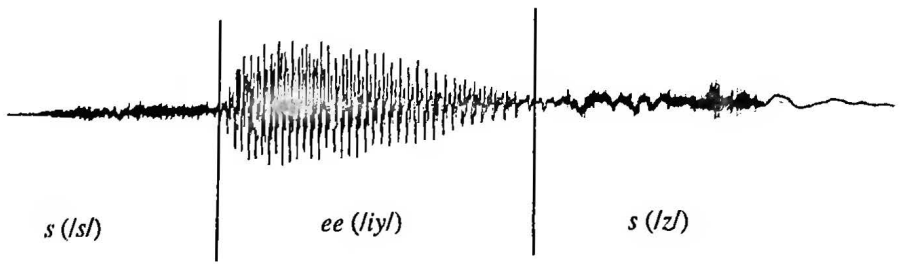

Figure 2.5 Waveform of *sees*, showing a voiceless phoneme /s/, followed by a voiced sound, the vowel /iy/. The final sound, /z/, is a type of voiced consonant.

The glottal cycle is illustrated in Figure 2.6. At stage (a), the vocal folds are closed and the air stream from the lungs is indicated by the arrow. At some point, the air pressure on the underside of the barrier formed by the vocal folds increases until it overcomes the resistance of the vocal fold closure and the higher air pressure below blows them apart (b). However, the tissues and muscles of the larynx and the vocal folds have a natural elasticity which tends to make them fall back into place rapidly, once air pressure is temporarily equalized (c). The successive airbursts resulting from this process are the source of energy for all voiced sounds. The time for a single open-close cycle depends on the stiffness and size of the vocal folds and the amount of subglottal air pressure. These factors can be controlled by a speaker to raise and lower the perceived frequency or pitch of a voiced sound.

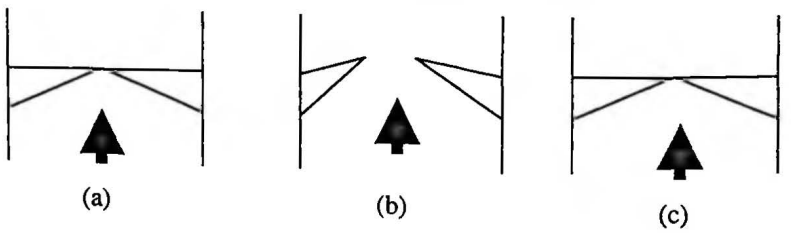

Figure 2.6 Vocal fold cycling at the larynx. (a) Closed with sub-glottal pressure buildup; (b) trans-glottal pressure differential causing folds to blow apart; (c) pressure equalization and tissue elasticity forcing temporary reclosure of vocal folds, ready to begin next cycle.

The waveform of air pressure variations created by this process can be described as a periodic flow, in cubic centimeters per second (after [15]). As shown in Figure 2.7, during the time bracketed as *one cycle*, there is no air flow during the initial closed portion. Then as

the glottis opens (open phase), the volume of air flow becomes greater. After a short peak, the folds begin to resume their original position and the air flow declines until complete closure is attained, beginning the next cycle. A common measure is the number of such cycles per second (Hz), or the fundamental frequency (*F0*). Thus the fundamental frequency for the waveform in Figure 2.7 is about 120 Hz.

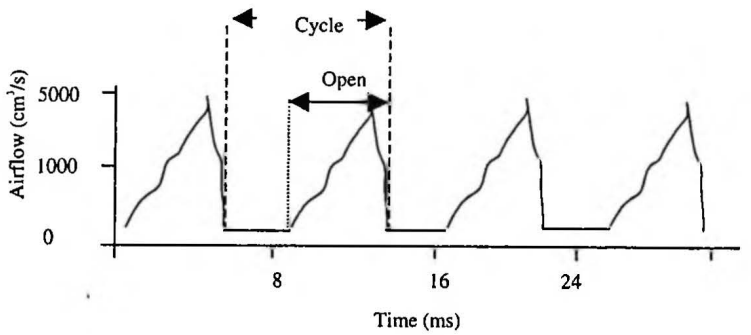

Figure 2.7 Waveform showing air flow during laryngeal cycle.

2.1.2.3. Spectrograms and Formants

Since the glottal wave is periodic, consisting of fundamental frequency (*F0*) and a number of harmonics (integral multiples of *F0*), it can be analyzed as a sum of sine waves as discussed in Chapter 5. The resonances of the vocal tract (above the glottis) are excited by the glottal energy. Suppose, for simplicity, we regard the vocal tract as a straight tube of uniform cross-sectional area, closed at the glottal end, open at the lips. When the shape of the vocal tract changes, the resonances change also. Harmonics near the resonances are emphasized, and, in speech, the resonances of the cavities that are typical of particular articulator configurations (e.g., the different vowel timbres) are called *formants*. The vowels in an actual speech waveform can be viewed from a number of different perspectives, emphasizing either a *cross-sectional* view of the harmonic responses at a single moment, or a longer-term view of the formant track evolution over time. The actual spectral analysis of a vowel at a single time-point, as shown in Figure 2.8, gives an idea of the uneven distribution of energy in resonances for the vowel */iy/* in the waveform for *see*, which is shown in Figure 2.5.

Another view of *sees* of Figure 2.5, called a spectrogram, is displayed in the lower part of Figure 2.9. It shows a long-term frequency analysis, comparable to a complete series of single time-point *cross sections* (such as that in Figure 2.8) ranged alongside one another in time and viewed from *above*.

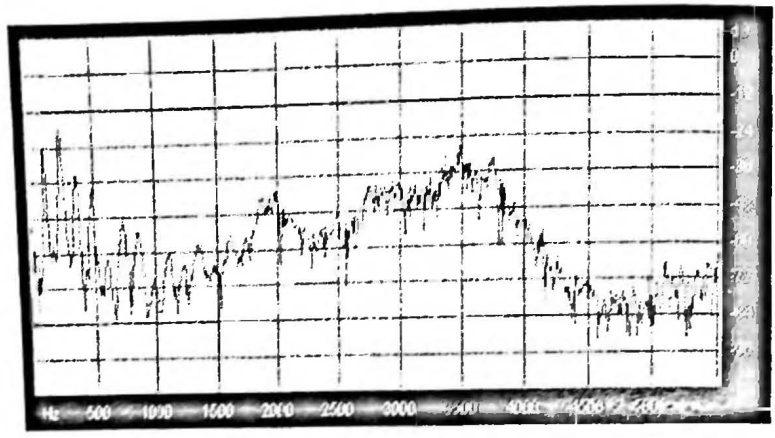

Figure 2.8 A spectral analysis of the vowel */iy/*, showing characteristically uneven distribution of energy at different frequencies.

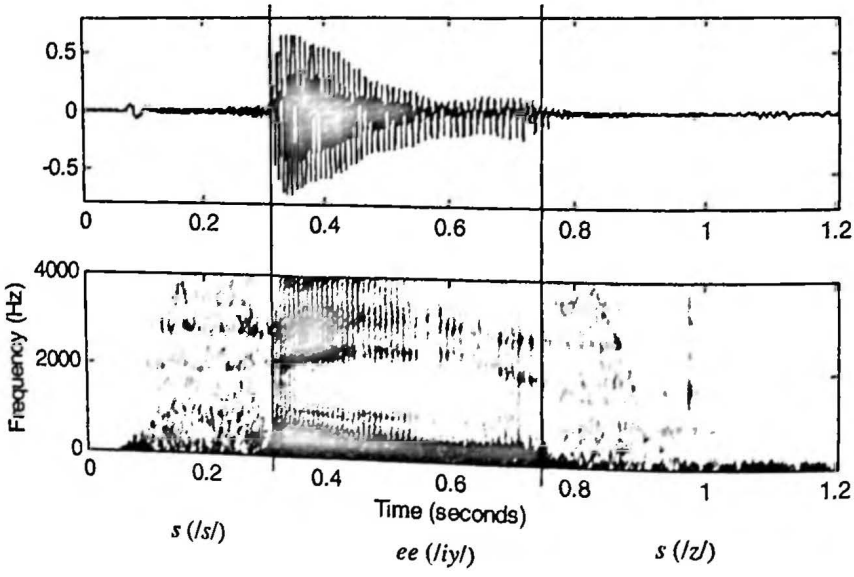

Figure 2.9 The spectrogram representation of the speech waveform *sees* (approximate phone boundaries are indicated with heavy vertical lines).

In the spectrogram of Figure 2.9, the darkness or lightness of a band indicates the relative amplitude or energy present at a given frequency. The dark horizontal bands show the formants, which are harmonics of the fundamental at natural resonances of the vocal tract cavity position for the vowel */iy/* in *see*. The mathematical methods for deriving analyses and representations such as those illustrated above are covered in Chapters 5 and 6.

2.1.3. Speech Perception

There are two major components in the auditory perception system: the peripheral auditory organs (ears) and the auditory nervous system (brain). The ear processes an acoustic pressure signal by first transforming it into a mechanical vibration pattern on the basilar membrane, and then representing the pattern by a series of pulses to be transmitted by the auditory nerve. Perceptual information is extracted at various stages of the auditory nervous system. In this section we focus mainly on the auditory organs.

2.1.3.1. Physiology of the Ear

The human ear, as shown in Figure 2.10, has three sections: the outer ear, the middle ear, and the inner ear. The outer ear consists of the external visible part and the external auditory canal that forms a tube along which sound travels. This tube is about 2.5 cm long and is covered by the eardrum at the far end. When air pressure variations reach the eardrum from the outside, it vibrates, and transmits the vibrations to bones adjacent to its opposite side. The vibration of the eardrum is at the same frequency (alternating compression and rarefaction) as the incoming sound pressure wave. The middle ear is an air-filled space or cavity about 1.3 cm across, and about 6 cm^3 volume. The air travels to the middle ear cavity along the tube (when opened) that connects the cavity with the nose and throat. The oval window shown in Figure 2.10 is a small membrane at the bony interface to the inner ear (cochlea). Since the cochlear walls are bony, the energy is transferred by mechanical action of the stapes into an impression on the membrane stretching over the oval window.

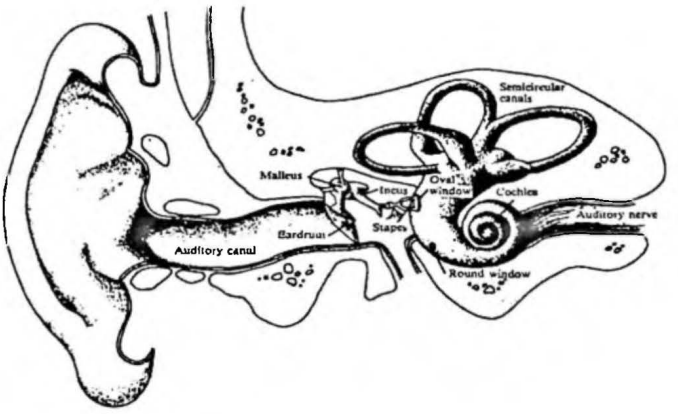

Figure 2.10 The structure of the peripheral auditory system with the outer, middle, and inner ear.

The relevant structure of the inner ear for sound perception is the cochlea, which communicates directly with the auditory nerve, conducting a representation of sound to the brain. The cochlea is a spiral tube about 3.5 cm long, which coils about 2.6 times. The spiral is divided, primarily by the basilar membrane running lengthwise, into two fluid-filled chambers. The cochlea can be roughly regarded as a filter bank, whose outputs are ordered by location, so that a frequency-to-place transformation is accomplished. The filters closest to the cochlear base respond to the higher frequencies, and those closest to its apex respond to the lower.

2.1.3.2. Physical vs. Perceptual Attributes

In psychoacoustics, a basic distinction is made between the perceptual attributes of a sound, especially a speech sound, and the measurable physical properties that characterize it. Each of the perceptual attributes, as listed in Table 2.2, seems to have a strong correlation with one main physical property, but the connection is complex, because other physical properties of the sound may affect perception in complex ways.

Table 2.2 Relation between perceptual and physical attributes of sound.

Physical Quantity	Perceptual Quality
Intensity	Loudness
Fundamental frequency	Pitch
Spectral shape	Timbre
Onset/offset time	Timing
Phase difference in binaural hearing	Location

Although sounds with a greater intensity level usually sound louder, the sensitivity of the ear varies with the frequency and the quality of the sound. One fundamental divergence between physical and perceptual qualities is the phenomenon of non-uniform *equal loudness* perception of tones of varying frequencies. In general, tones of differing pitch have different inherent *perceived loudness*. The sensitivity of the ear varies with the frequency and the quality of the sound. The graph of equal loudness contours adopted by ISO is shown in Figure 2.11. These curves demonstrate the relative insensitivity of the ear to sounds of low frequency at moderate to low intensity levels. Hearing sensitivity reaches a maximum around 4000 Hz, which is near the first resonance frequency of the outer ear canal, and peaks again around 13 kHz, the frequency of the second resonance [38].

Pitch is indeed most closely related to the fundamental frequency. The higher the fundamental frequency, the higher the pitch we perceive. However, discrimination between two pitches depends on the frequency of the lower pitch. Perceived pitch will change as intensity is increased and frequency is kept constant.

In another example of the non-identity of acoustic and perceptual effects, it has been observed experimentally that when the ear is exposed to two or more different tones, it is a common experience that one tone may *mask* the others. Masking is probably best explained

as an upward shift in the hearing threshold of the weaker tone by the louder tone. Pure tones, complex sounds, narrow and broad bands of noise all show differences in their ability to mask other sounds. In general, pure tones close together in frequency mask each other more than tones widely separated in frequency. A pure tone masks tones of higher frequency more effectively than tones of lower frequency. The greater the intensity of the masking tone, the broader the range of the frequencies it can mask [18, 31].

Binaural listening greatly enhances our ability to sense the direction of the sound source. The sense of localization attention is mostly focused on side-to-side discrimination or *lateralization*. Time and intensity cues have different impacts for low frequency and high frequency, respectively. Low-frequency sounds are lateralized mainly on the basis of interaural time difference, whereas high-frequency sounds are localized mainly on the basis of interaural intensity differences [5].

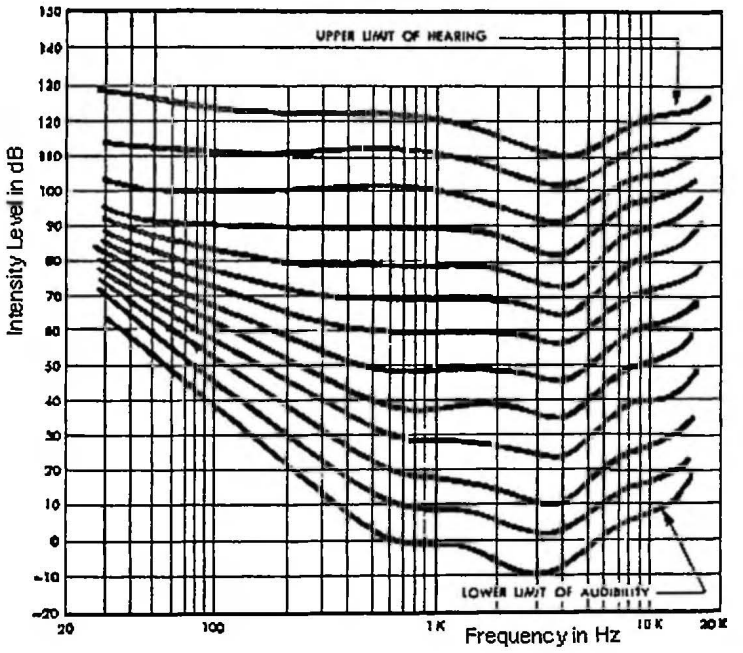

Figure 2.11 Equal-loudness curves indicate that the response of the human hearing mechanism is a function of frequency and loudness levels. This relationship again illustrates the difference between physical dimensions and psychological experience (after ISO 226).

Finally, an interesting perceptual issue is the question of distinctive voice quality. Speech from different people sounds different. Partially this is due to obvious factors, such as differences in characteristic fundamental frequency caused by, for example, the greater mass and length of adult male vocal folds as opposed to female. But there are more subtle

effects as well. In psychoacoustics, the concept of *timbre* (of a sound or instrument) is defined as that attribute of auditory sensation by which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar. In other words, when all the easily measured differences are controlled, the remaining perception of difference is ascribed to timbre. This is heard most easily in music, where the same note in the same octave played for the same duration on a violin sounds different from a flute. The timbre of a sound depends on many physical variables including a sound's spectral power distribution, its temporal envelope, rate and depth of amplitude or frequency modulation, and the degree of inharmonicity of its harmonics.

2.1.3.3. Frequency Analysis

Researchers have undertaken psychoacoustic experimental work to derive frequency scales that attempt to model the natural response of the human perceptual system, since the cochlea of the inner ear acts as a spectrum analyzer. The complex mechanism of the inner ear and auditory nerve implies that the perceptual attributes of sounds at different frequencies may not be entirely simple or linear in nature. It is well known that the western musical pitch is described in *octaves*[1] and *semi-tones.*[2] The perceived musical pitch of complex tones is basically proportional to the logarithm of frequency. For complex tones, the just noticeable difference for frequency is essentially constant on the octave/semi-tone scale. Musical pitch scales are used in prosodic research (on speech intonation contour generation).

AT&T Bell Labs has contributed many influential discoveries in hearing, such as critical band and articulation index, since the turn of the 20th century [3]. Fletcher's work [14] pointed to the existence of critical bands in the cochlear response. Critical bands are of great importance in understanding many auditory phenomena such as perception of loudness, pitch, and timbre. The auditory system performs frequency analysis of sounds into their component frequencies. The cochlea acts as if it were made up of overlapping filters having bandwidths equal to the critical bandwidth. One class of critical band scales is called *Bark frequency scale*. It is hoped that by treating spectral energy over the Bark scale, a more natural fit with spectral information processing in the ear can be achieved. The Bark scale ranges from 1 to 24 Barks, corresponding to 24 critical bands of hearing as shown in Table 2.3. As shown in Figure 2.12, the perceptual resolution is finer in the lower frequencies. It should be noted that the ear's critical bands are continuous, and a tone of any audible frequency always finds a critical band centered on it. The Bark frequency b can be expressed in terms of the linear frequency (in Hz) by

$$b(f) = 13\arctan(0.00076f) + 3.5 * \arctan\left((f/7500)^2\right) \quad (Bark) \tag{2.5}$$

[1] A tone of frequency f_1 is said to be an octave above a tone with frequency f_2 if and only if $f_1 = 2f_2$.
[2] There are 12 semitones in one octave, so a tone of frequency f_1 is said to be a semitone above a tone with frequency f_2 if and only if $f_1 = 2^{1/12} f_2 = 1.05946 f_2$.

Table 2.3 The Bark frequency scale.

Bark Band #	Edge (Hz)	Center (Hz)
1	100	50
2	200	150
3	300	250
4	400	350
5	510	450
6	630	570
7	770	700
8	920	840
9	1080	1000
10	1270	1170
11	1480	1370
12	1720	1600
13	2000	1850
14	2320	2150
15	2700	2500
16	3150	2900
17	3700	3400
18	4400	4000
19	5300	4800
20	6400	5800
21	7700	7000
22	9500	8500
23	12000	10500
24	15500	13500

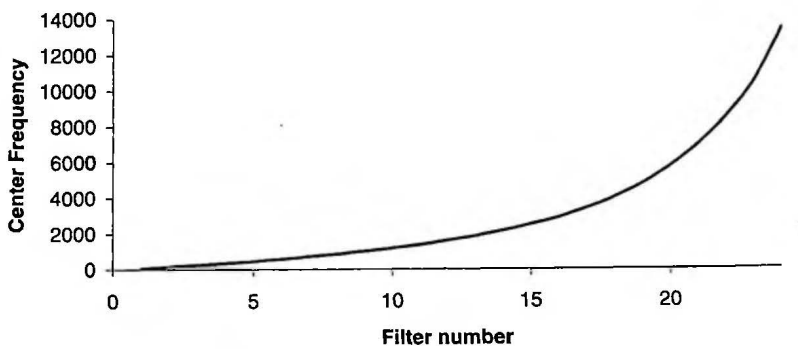

Figure 2.12 The center frequency of 24 Bark frequency filters as illustrated in Table 2.3.

Another such perceptually motivated scale is the mel frequency scale [41], which is linear below 1 kHz, and logarithmic above, with equal numbers of samples taken below and above 1 kHz. The mel scale is based on experiments with simple tones (sinusoids) in which subjects were required to divide given frequency ranges into four perceptually equal intervals or to adjust the frequency of a stimulus tone to be half as high as that of a comparison tone. One mel is defined as one thousandth of the pitch of a 1 kHz tone. As with all such attempts, it is hoped that the mel scale more closely models the sensitivity of the human ear than a purely linear scale and provides for greater discriminatory capability between speech segments. Mel-scale frequency analysis has been widely used in modern speech recognition systems. It can be approximated by:

$$B(f) = 1125 \ln(1 + f/700) \tag{2.6}$$

The mel scale is plotted in Figure 2.13 together with the Bark scale and the bilinear transform (see Chapter 6).

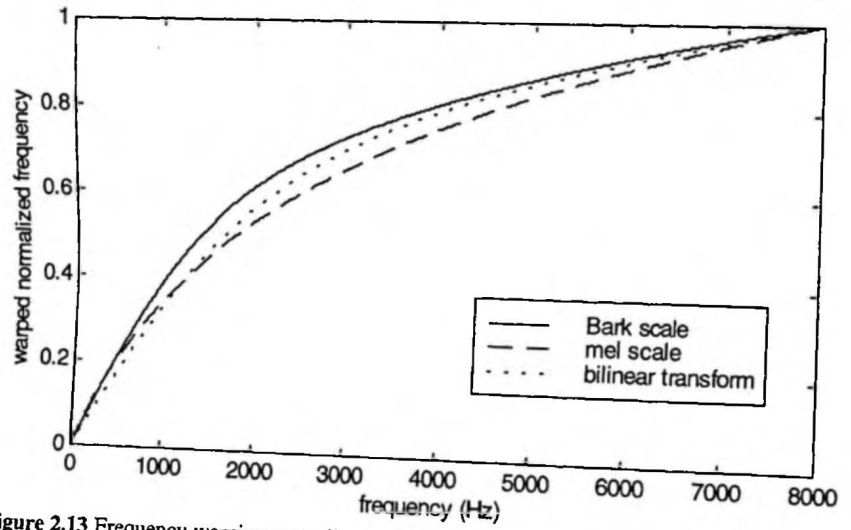

Figure 2.13 Frequency warping according to the Bark scale, ERB scale, mel-scale, and bilinear transform for $\alpha = 0.6$: linear frequency in the x-axis and normalized frequency in the y-axis.

A number of techniques in the modern spoken language system, such as cepstral analysis, and dynamic feature, have benefited tremendously from perceptual research as discussed throughout this book.

2.1.3.4. Masking

Frequency masking is a phenomenon under which one sound cannot be perceived if another sound close in frequency has a high enough level. The first sound *masks* the other one. Fre-

quency-masking levels have been determined empirically, with complicated models that take into account whether the masker is a tone or noise, the masker's level, and other considerations.

We now describe a phenomenon known as *tone-masking noise.* It has been determined empirically that noise with energy E_N (dB) at Bark frequency g masks a tone at Bark frequency b if the tone's energy is below the threshold

$$T_T(b) = E_N - 6.025 - 0.275g + S_m(b-g) \quad (dB\ SPL) \tag{2.7}$$

where the *spread-of-masking* function $S_m(b)$ is given by

$$S_m(b) = 15.81 + 7.5(b+0.474) - 17.5\sqrt{1+(b+0.474)^2} \quad (dB) \tag{2.8}$$

We now describe a phenomenon known as *noise-masking tone.* It has been determined empirically that a tone at Bark frequency g with energy E_T (dB) masks noise at Bark frequency b if the noise energy is below the threshold

$$T_N(b) = E_T - 2.025 - 0.175g + S_m(b-g) \quad (dB\ SPL) \tag{2.9}$$

Masking thresholds are commonly referred to in the literature as Bark scale functions of *just noticeable distortion* (JND). Equation (2.8) can be approximated by a triangular spreading function that has slopes of +25 and –10 dB per Bark, as shown in Figure 2.14.

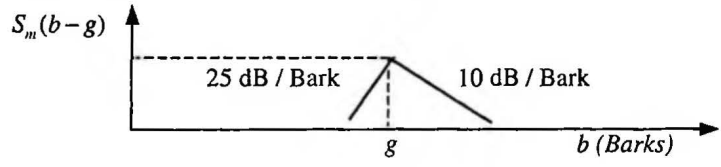

Figure 2.14 Contribution of Bark frequency g to the masked threshold $S_m(b)$.

In Figure 2.15 we show both the threshold of hearing and the masked threshold of a tone at 1 kHz with a 69 dB SPL. The combined masked threshold is the sum of the two in the linear domain

$$T(f) = 10\log_{10}\left(10^{0.1T_h(f)} + 10^{0.1T_T(f)}\right) \tag{2.10}$$

which is approximately the largest of the two.

In addition to frequency masking, there is a phenomenon called temporal masking by which a sound too close in time to another sound cannot be perceived. Whereas premasking tends to last about 5 ms, postmasking can last from 50 to 300 ms. Temporal masking level of a masker with a uniform level starting at 0 ms and lasting 200 ms is shown in Figure 2.16.

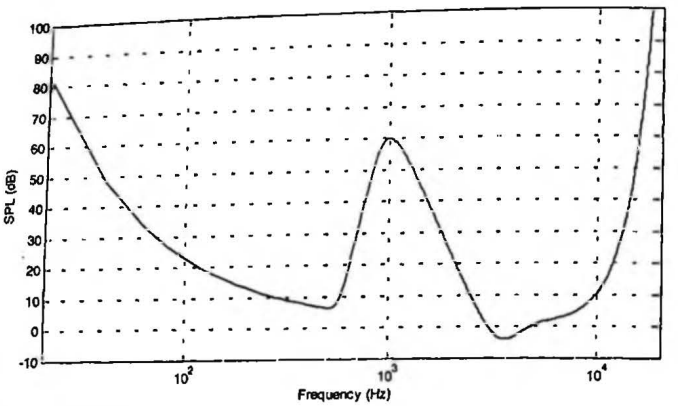

Figure 2.15 Absolute threshold of hearing and spread of masking threshold for a 1 kHz sine-wave masker with a 69 dB SPL. The overall masked threshold is approximately the largest of the two thresholds.

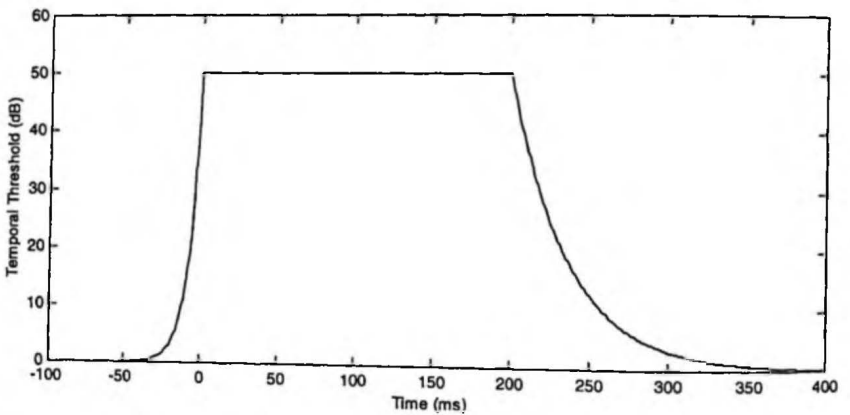

Figure 2.16 Temporal masking level of a masker with a uniform level starting at 0 ms and lasting 200 ms.

2.2. PHONETICS AND PHONOLOGY

We now discuss basic phonetics and phonology needed for spoken language processing. *Phonetics* refers to the study of speech sounds and their production, classification, and transcription. *Phonology* is the study of the distribution and patterning of speech sounds in a language and of the tacit rules governing pronunciation.

2.2.1. Phonemes

Linguist Ferdinand de Saussere (1857-1913) is credited with the observation that the relation between a sign and the object signified by it is arbitrary. The same concept, a certain yellow and black flying social insect, has the sign *honeybee* in English and *mitsubachi* in Japanese.

There is no particular relation between the various pronunciations and the meaning, nor do these pronunciations per se describe the bee's characteristics in any detail. For phonetics, this means that the speech sounds described in this chapter have no inherent meaning, and should be randomly distributed across the lexicon, except as affected by extraneous historical or etymological considerations. The sounds are just a set of arbitrary effects made available by human vocal anatomy. You might wonder about this theory when you observe, for example, the number of words beginning with *sn* that have to do with nasal functions in English: *sneeze, snort, sniff, snot, snore, snuffle*, etc. But Saussere's observation is generally true, except for obvious onomatopoetic (sound) words like *buzz*.

Like fingerprints, every speaker's vocal anatomy is unique, and this makes for unique vocalizations of speech sounds. Yet language communication is based on commonality of form at the perceptual level. To allow discussion of the commonalities, researchers have identified certain gross characteristics of speech sounds that are adequate for description and classification of words in dictionaries. They have also adopted various systems of notation to represent the subset of phonetic phenomena that are crucial for meaning.

As an analogy, consider the system of computer coding of text characters. In such systems, the *character* is an abstraction, e.g. the Unicode character U+0041. The identifying property of this character is its Unicode name *LATIN CAPITAL LETTER A*. This is a genuine abstraction; no particular realization is necessarily specified. As the Unicode 2.1 standard [1] states:

The Unicode Standard does not define glyph images. The standard defines how characters are interpreted, not how glyphs are rendered. The software or hardware-rendering engine of a computer is responsible for the appearance of the characters on the screen. The Unicode Standard does not specify the size, shape, nor orientation of on-screen characters.

Thus, the U+0041 character can be realized differently for different purposes, and in different sizes with different fonts:

U+0041 → A, *A*, A, A, A, ...

The realizations of the character U+0041 are called glyphs, and there is no distinguished uniquely correct glyph for U+0041. In speech science, the term *phoneme* is used to denote any of the minimal units of speech sound in a language that can serve to distinguish one word from another. We conventionally use the term *phone* to denote a phoneme's acoustic realization. In the example given above, U+0041 corresponds to a phoneme and the various fonts correspond to the phone. For example, English phoneme */t/* have two very different acoustic realizations in the words *sat* and *meter*. You had better treat them as two different phones if you want to build a spoken language system. We will use the terms *phone* or *phoneme* interchangeably to refer to the speaker-independent and context-independent units of meaningful sound contrast. Table 2.4 shows a complete list of phonemes used in American English. The set of phonemes will differ in realization across individual speakers. But phonemes will always function systematically to differentiate meaning in words, just as the phoneme */p/* signals the word *pat* as opposed to the similar-sounding but distinct *bat*. The important contrast distinguishing this pair of words is */p/* vs. */b/*.

In this section we concentrate on the basic qualities that *define and differentiate abstract phonemes*. In Section 2.2.1.3 below we consider *why and how phonemes vary* in their actual realizations by different speakers and in different contexts.

Table 2.4 English phonemes used for typical spoken language systems.

Phonemes	Word Examples	Description
iy	*feel, eve, me*	front close unrounded
ih	*fill, hit, lid*	front close unrounded (lax)
ae	*at, carry, gas*	front open unrounded (tense)
aa	*father, ah, car*	back open unrounded
ah	*cut, bud, up*	open-mid back unrounded
ao	*dog, lawn, caught*	open-mid back round
ay	*tie, ice, bite*	diphthong with quality: aa + ih
ax	*ago, comply*	central close mid (schwa)
ey	*ate, day, tape*	front close-mid unrounded (tense)
eh	*pet, berry, ten*	front open-mid unrounded
er	*turn, fur, meter*	central open-mid unrounded rhoti-
ow	*go, own, tone*	back close-mid rounded
aw	*foul, how, our*	diphthong with quality: aa + uh
oy	*toy, coin, oil*	diphthong with quality: ao + ih
uh	*book, pull, good*	back close-mid unrounded (lax)
uw	*tool, crew, moo*	back close round
b	*big, able, tab*	voiced bilabial plosive
p	*put, open, tap*	voiceless bilabial plosive
d	*dig, idea, wad*	voiced alveolar plosive
t	*talk, sat*	voiceless alveolar plosive &
t	*meter*	alveolar flap
g	*gut, angle, tag*	voiced velar plosive
k	*cut, ken, take*	voiceless velar plosive
f	*fork, after, if*	voiceless labiodental fricative
v	*vat, over, have*	voiced labiodental fricative
s	*sit, cast, toss*	voiceless alveolar fricative
z	*zap, lazy, haze*	voiced alveolar fricative
th	*thin, nothing, truth*	voiceless dental fricative
dh	*then, father, scythe*	voiced dental fricative
sh	*she, cushion, wash*	voiceless postalveolar fricative
zh	*genre, azure*	voiced postalveolar fricative
l	*lid*	alveolar lateral approximant
l	*elbow, sail*	velar lateral approximant
r	*red, part, far*	retroflex approximant
y	*yacht, yard*	palatal sonorant glide
w	*with, away*	labiovelar sonorant glide
hh	*help, ahead, hotel*	voiceless glottal fricative
m	*mat, amid, aim*	bilabial nasal
n	*no, end, pan*	alveolar nasal
ng	*sing, anger*	velar nasal
ch	*chin, archer, march*	voiceless alveolar affricate: t + sh
jh	*joy, agile, edge*	voiced alveolar affricate: d + zh

2.2.1.1. Vowels

The tongue shape and positioning in the oral cavity do not form a major constriction of air flow during vowel articulation. However, variations of tongue placement give each vowel its distinct character by changing the resonance, just as different sizes and shapes of bottles give rise to different acoustic effects when struck. The primary energy entering the pharyngeal and oral cavities in vowel production vibrates at the fundamental frequency. The major resonances of the oral and pharyngeal cavities for vowels are called F1 and F2 – the first and second formants, respectively. They are determined by tongue placement and oral tract shape in vowels, and they determine the characteristic timbre or quality of the vowel.

The relationship of F1 and F2 to one another can be used to describe the English vowels. While the shape of the complete vocal tract determines the spectral outcome in a complex, nonlinear fashion, generally F1 corresponds to the back or pharyngeal portion of the cavity, while F2 is determined more by the size and shape of the oral portion, forward of the major tongue extrusion. This makes intuitive sense – the cavity from the glottis to the tongue extrusion is longer than the forward part of the oral cavity, thus we would expect its resonance to be lower. In the vowel of *see*, for example, the tongue extrusion is far forward in the mouth, creating an exceptionally long rear cavity, and correspondingly low F1. The forward part of the oral cavity, at the same time, is extremely short, contributing to higher F2. This accounts for the wide separation of the two lowest dark horizontal bands in Figure 2.9, corresponding to F1 and F2, respectively. Rounding the lips has the effect of extending the front-of-tongue cavity, thus lowering F2. Typical values of F1 and F2 of American English vowels are listed in Table 2.5.

Table 2.5 Phoneme labels and typical formant values for vowels of English.

Vowel Labels	Mean F1 (Hz)	Mean F2 (Hz)
iy (feel)	300	2300
ih (fill)	360	2100
ae (gas)	750	1750
aa (father)	680	1100
ah (cut)	720	1240
ao (dog)	600	900
ax (comply)	720	1240
eh (pet)	570	1970
er (turn)	580	1380
ow (tone)	600	900
uh (good)	380	950
uw (tool)	300	940

The characteristic F1 and F2 values for vowels are sometimes called formant targets, which are ideal locations for perception. Sometimes, due to fast speaking or other limitations on performance, the speaker cannot quite attain an ideal target before the articulators begin shifting to targets for the following phoneme, which is phonetic context dependent. Additionally, there is a special class of vowels that combine two distinct sets of F1/F2 targets.

These are called *diphthongs*. As the articulators move, the initial vowel targets glide smoothly to the final configuration. Since the articulators are working faster in production of a diphthong, sometimes the *ideal* formant target values of the component values are not fully attained. Typical diphthongs of American English are listed in Table 2.6.

Table 2.6 The diphthongs of English.

Diphthong Labels	Components
ay (tie)	/aa/ → /iy/
ey (ate)	/eh/ → /iy/
oy (coin)	/ao/ → /iy/
aw (foul)	/aa/ → /uw/

Figure 2.17 shows the first two formants for a number of typical vowels.

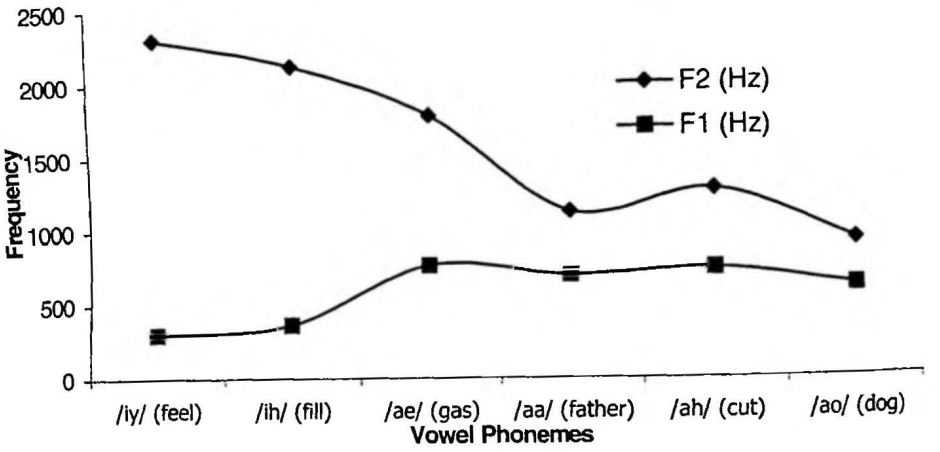

Figure 2.17 F1 and F2 values for articulations of some English vowels.

The major articulator for English vowels is the middle to rear portion of the tongue. The position of the tongue's surface is manipulated by large and powerful muscles in its root, which move it as a whole within the mouth. The linguistically important dimensions of movement are generally the ranges [front ⇔ back] and [high ⇔ low]. You can feel this movement easily. Say mentally, or whisper, the sound */iy/* (as in *see*) and then */aa/* (as in *father*). Do it repeatedly, and you will get a clear perception of the tongue movement from high to low. Now try */iy/* and then */uw/* (as in *blue*), repeating a few times. You will get a clear perception of place of articulation from front */iy/* to back */uw/*. Figure 2.18 shows a schematic characterization of English vowels in terms of relative tongue positions. There are two kinds of vowels: those in which tongue height is represented as a point and those in which it is represented as a vector.

Though the tongue hump is the major actor in vowel articulation, other articulators come into play as well. The most important secondary vowel mechanism for English and many other languages is lip rounding. Repeat the exercise above, moving from the */iy/* (*see*)

to the */uw/* (*blue*) position. Now rather than noticing the tongue movement, pay attention to your lip shape. When you say */iy/*, your lips will be flat, slightly open, and somewhat spread. As you move to */uw/*, they begin to *round out*, ending in a more puckered position. This lengthens the oral cavity during */uw/*, and affects the spectrum in other ways.

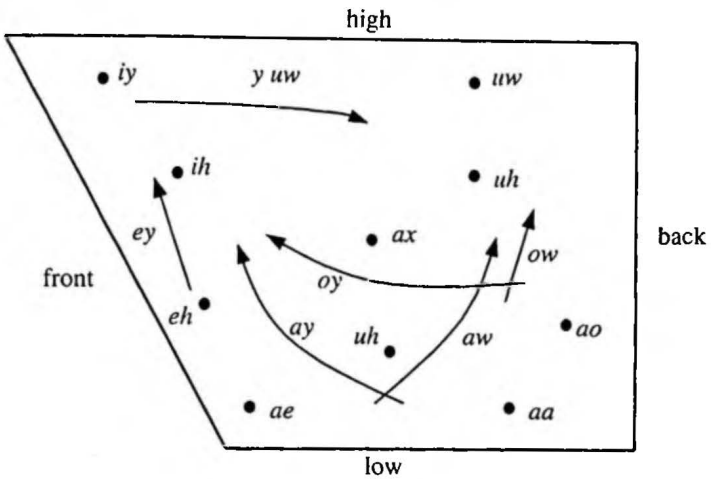

Figure 2.18 Relative tongue positions of English vowels [24].

Though there is always some controversy, linguistic study of phonetic abstractions, called *phonology*, has largely converged on the five binary features: +/- high, +/- low, +/- front, +/- back, and +/- round, plus the phonetically ambiguous but phonologically useful feature +/- tense, as adequate to uniquely characterize the major vowel distinctions of Standard English (and many other languages). Obviously, such a system is a little bit too free with logically contradictory specifications, such as [+high, +low], but these are excluded from real-world use. These features can be seen in Table 2.7.

Table 2.7 Phonological (abstract) feature decomposition of basic English vowels.

Vowel	high	low	front	back	round	tense
iy	+	-	+	-	-	+
ih	+	-	+	-	-	-
ae	-	+	+	-	-	+
aa	-	+	-	-	-	+
ah	-	-	-	-	-	+
ao	-	+	-	+	+	+
ax	-	-	-	-	-	-
eh	-	-	+	-	-	-
ow	-	-	-	+	+	+
uh	+	-	-	+	-	-
uw	+	-	-	+	-	+

This kind of abstract analysis allows researchers to make convenient statements about classes of vowels that behave similarly under certain conditions. For example, one may speak simply of the high vowels to indicate the set */iy, ih, uh, uw/*.

2.2.1.2. Consonants

Consonants, as opposed to vowels, are characterized by significant constriction or obstruction in the pharyngeal and/or oral cavities. Some consonants are voiced; others are not. Many consonants occur in pairs, that is, they share the same configuration of articulators, and one member of the pair additionally has voicing which the other lacks. One such pair is */s, z/*, and the voicing property that distinguishes them shows up in the non-periodic noise of the initial segment */s/* in Figure 2.5 as opposed to the voiced consonant end-phone, */z/*. Manner of articulation refers to the articulation mechanism of a consonant. The major distinctions in manner of articulation are listed in Table 2.8.

Table 2.8 Consonant manner of articulation.

Manner	Sample Phone	Example Words	Mechanism
Plosive	*/p/*	tat, tap	Closure in oral cavity
Nasal	*/m/*	team, meet	Closure of nasal cavity
Fricative	*/s/*	sick, kiss	Turbulent airstream noise
Retroflex liquid	*/r/*	rat, tar	Vowel-like, tongue high and curled back
Lateral liquid	*/l/*	lean, kneel	Vowel-like, tongue central, side airstream
Glide	*/y/,/w/*	yes, well	Vowel-like

The English phones that typically have voicing without complete obstruction or narrowing of the vocal tract are called *semivowels* and include */l, r/*, the *liquid* group, and */y, w/*, the *glide* group. Liquids, glides, and vowels are all *sonorant*, meaning they have continuous voicing. Liquids */l/* and */r/* are quite vowel-like and in fact may become *syllabic* or act entirely as vowels in certain positions, such as the *l* at the end of *edible*. In */l/*, the airstream flows around the sides of the tongue, leading to the descriptive term *lateral*. In */r/*, the tip of the tongue is curled back slightly, leading to the descriptive term *retroflex*. Figure 2.19 shows some semivowels.

Glides */y, w/* are basically vowels */iy, uw/* whose initial position within the syllable require them to be a little shorter and to lack the ability to be stressed, rendering them just different enough from true vowels that they are classed as a special category of consonant. Pre-vocalic glides that share the syllable-initial position with another consonant, such as the */y/* in the second syllable of *computer /k uh m . p y uw . t er/*, or the */w/* in *quick /k w ih k/*, are sometimes called *on-glides*. The semivowels, as a class, are sometimes called *approximants*, meaning that the tongue approaches the top of the oral cavity, but does not completely contact so as to obstruct the air flow.

Even the non-sonorant consonants that require complete or close-to-complete obstruction may still maintain some voicing before or during the obstruction, until the pressure dif-

ferential across the glottis starts to disappear, due to the closure. Such voiced consonants include */b,d,g, z, zh, v/*. They have a set of counterparts that differ only in their characteristic lack of voicing: */p,t,k, s, sh, fl*.

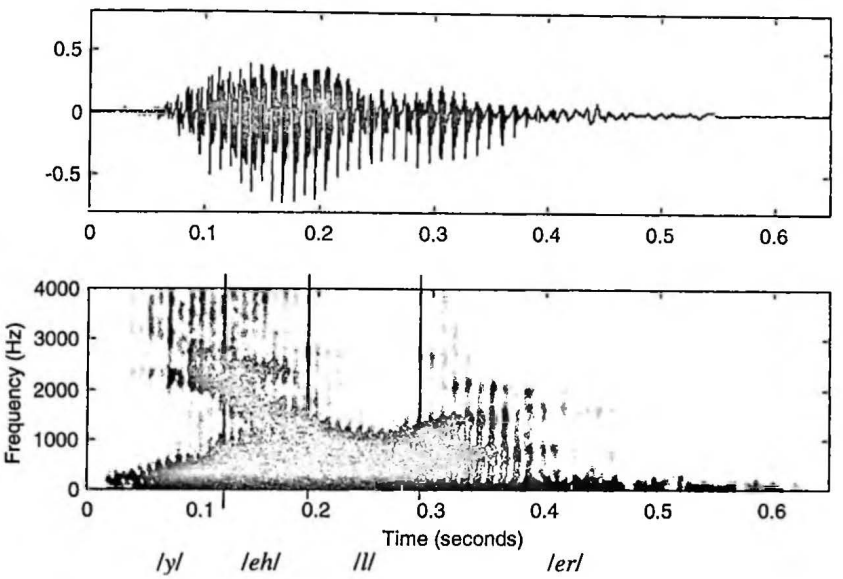

Figure 2.19 Spectrogram for the word *yeller*, showing semivowels */y/*, */l/*, */er/* (approximate phone boundaries shown with vertical lines).

Nasal consonants */m,n/* are a mixed bag: the oral cavity has significant constriction (by the tongue or lips), yet the voicing is continuous, like that of the sonorants, because, with the velar flap open, air passes freely through the nasal cavity, maintaining a pressure differential across the glottis.

A consonant that involves complete blockage of the oral cavity is called an obstruent stop, or plosive consonant. These may be voiced throughout if the trans-glottal pressure drop can be maintained long enough, perhaps through expansion of the wall of the oral cavity. In any case, there can be voicing for the early sections of stops. Voiced, unvoiced pairs of stops include: */b,p/*, */d,t/*, and */g,k/*. In viewing the waveform of a stop, a period of silence corresponding to the oral closure can generally be observed. When the closure is removed (by opening the constrictor, which may be lips or tongue), the trapped air rushes out in a more or less sudden manner. When the upper oral cavity is unimpeded, the closure of the vocal folds themselves can act as the initial blocking mechanism for a type of stop heard at the very beginning of vowel articulation in vowel-initial words like *atrophy*. This is called a *glottal stop*. Voiceless plosive consonants in particular exhibit a characteristic aperiodic *burst* of energy at the (articulatory) point of closure as shown in Figure 2.20 just prior to */i/*. By com-