MAS.632 Speech Interfaces and Mobile Devices
Fall 2008

1

Speech as Communication

Speech can be viewed in many ways. Although chapters of this book focus on specific aspects of speech and the computer technologies that utilize speech, the reader should begin with a broad perspective on the role of speech in our daily lives. It is essential to appreciate the range of capabilities that conversational systems must possess before attempting to build them. This chapter lays the groundwork for the entire book by presenting several perspectives on speech communication.

The first section of this chapter emphasizes the *interactive* and *expressive* role of voice communication. Except in formal circumstances such as lectures and dramatic performances, speech occurs in the context of a *conversation,* wherein participants take turns speaking, interrupt each other, nod in agreement, or try to change the topic. Computer systems that talk or listen may ultimately be judged by their ability to converse in like manner simply because conversation permeates human experience. The second section discusses the various components or *layers* of a conversation. Although the distinctions between these layers are somewhat contrived, they provide a means of analyzing the communication process; research disciplines have evolved for the study of each of these components. Finally, the last section introduces the *representations* of speech and conversation, corresponding in part to the layers identified in the second section. These representations provide abstractions that a computer program may employ to engage in a conversation with a human.

SPEECH AS CONVERSATION

Conversation is a process involving multiple participants, shared knowledge, and a protocol for taking turns and providing mutual feedback. Voice is our primary channel of interaction in conversation, and speech evolved in humans in response to the need among its members to communicate. It is hard to imagine many uses of speech that do not involve some interchange between multiple participants in a conversation; if we are discovered talking to ourselves, we usually feel embarrassed.

For people of normal physical and mental ability, speech is both rich in expressiveness and easy to use. We learn it without much apparent effort as children and employ it spontaneously on a daily basis.[1] People employ many layers of knowledge and sophisticated protocols while having a conversation; until we attempt to analyze dialogues, we are unaware of the complexity of this interplay between parties.

Although much is known about language, study of interactive speech communication has begun only recently. Considerable research has been done on natural language processing systems, but much of this is based on keyboard input. It is important to note the contrast between written and spoken language and between read or rehearsed speech and spontaneous utterances. Spoken language is less formal than written language, and errors in construction of spoken sentences are less objectionable. Spontaneous speech shows much evidence of the real-time processes associated with its production, including false starts, nonspeech noises such as mouth clicks and breath sounds, and pauses either silent or filled (". . . um . . .") [Zue *et al.* 1989b]. In addition, speech naturally conveys intonational and emotional information that fiction writers and playwrights must struggle to impart to written language.

Speech is rich in interactive techniques to guarantee that the listener understands what is being expressed, including facial expressions, physical and vocal gestures, "uh-huhs," and the like. At certain points in a conversation, it is appropriate for the listener to begin speaking; these points are often indicated by longer pauses and lengthened final syllables or marked decreases in pitch at the end of a sentence. Each round of speech by one person is called a **turn; interruption** occurs when a participant speaks before a break point offered by the talker. Instead of taking a turn, the listener may quickly indicate agreement with a word or two, a nonverbal sound ("uh-huh"), or a facial gesture. Such responses, called **back channels,** speed the exchange and result in more effective conversations [Kraut *et al.* 1982].[2]

Because of these interactive characteristics, speech is used for immediate communication needs, while writing often implies a distance, either in time or space,

[1]For a person with normal speech and hearing to spend a day without speaking is quite a novel experience.

[2]We will return to these topics in Chapter 9.

between the author and reader. Speech is used in transitory interactions or situations in which the process of the interaction may be as important as its result. For example, the agenda for a meeting is likely to be written, and a written summary or minutes may be issued "for the record," but the actual decisions are made during a conversation. Chapanis and his colleagues arranged a series of experiments to compare the effectiveness of several communication media, i.e., voice, video, handwriting, and typewriting, either alone or in combination, for problem-solving tasks [Ochsman and Chapanis 1974]. Their findings indicated an overwhelming contribution of voice for such interactions. Any experimental condition that included voice was superior to any excluding voice; the inclusion of other media with voice resulted in only a small additional effectiveness. Although these experiments were simplistic in their use of student subjects and invented tasks and more recent work by others [Minneman and Bly 1991] clarifies a role for video interaction, the dominance of voice seems unassailable.

But conversation is more than mere interaction; communication often serves a purpose of changing or influencing the parties speaking to each other. I tell you something I have learned with the intention that you share my knowledge and hence enhance your view of the world. Or I wish to obtain some information from you so I ask you a question, hoping to elicit a reply. Or perhaps I seek to convince you to perform some activity for me; this may be satisfied either by your physical performance of the requested action or by your spoken promise to perform the act at a later time. "Speech Act" theories (to be discussed in more detail in Chapter 9) attempt to explain language as action, e.g., to request, command, query, and promise, as well as to inform.

The intention behind an utterance may not be explicit. For example, "Can you pass the salt?" is not a query about one's ability; it is a request. Many actual conversations resist such purposeful classifications. Some utterances ("go ahead," "uh-huh," "just a moment") exist only to guide the flow of the conversation or comment on the state of the discourse, rather than to convey information. Directly purposeful requests are often phrased in a manner allowing flexibility of interpretation and response. This looseness is important to the process of people defining and maintaining their work roles with respect to each other and establishing socially comfortable relationships in a hierarchical organization. The richness of speech allows a wide range of "acceptance" and "agreement" from wholehearted to skeptical to incredulous.

Speech also serves a strong social function among individuals and is often used just to pass the time, tell jokes, or talk about the weather. Indeed, extended periods of silence among a group may be associated with interpersonal awkwardness or discomfort. Sometimes the actual occurrence of the conversation serves a more significant purpose than any of the topics under discussion. Speech may be used to call attention to oneself in a social setting or as an exclamation of surprise or dismay in which an utterance has little meaning with respect to any preceding conversation. [Goffman 1981]

The expressiveness of speech and robustness of conversation strongly support the use of speech in computer systems, both for stored voice as a data type as well as speech as a medium of interaction. Unfortunately, current computers are

capable of uttering only short sentences of marginal intelligibility and occasionally recognizing single words. Engaging a computer in a conversation can be like an interaction in a foreign country. One studies the phrase book, utters a request, and in return receives either a blank stare (wrong pronunciation, try again) or a torrent of fluent speech in which one cannot perceive even the word boundaries.

However, limitations in technology only reinforce the need to take advantage of conversational techniques to ensure that the user is understood. Users will judge the performance of computer systems employing speech on the basis of their expectations about conversation developed from years of experience speaking with fellow humans. Users may expect computers to be either deaf and dumb, or once they realize the system can talk and listen, expect it to speak fluently like you and me. Since the capabilities of current speech technology lie between these extremes, building effective conversational computer systems can be very frustrating.

HIERARCHICAL STRUCTURE OF CONVERSATION

A more analytic approach to speech communication reveals a number of different ways of describing what actually occurs when we speak. The hierarchical structure of such analysis suggests goals to be attained at various stages in computer-based speech communication.

Conversation requires apparatus both for listening and speaking. Effective communication invokes mental processes employing the mouth and ears to convey a message thoroughly and reliably. There are many layers at which we can analyze the communication process, from the lower layers where speech is considered primarily acoustically to higher layers that express meaning and intention. Each layer involves increased knowledge and potential for intelligence and interactivity.

From the point of view of the speaker, we may look at speech from at least eight layers of processing as shown in Figure 1.1.

Layers of Speech Processing

discourse The regulation of conversation for pragmatic ends. This includes taking turns talking, the history of referents in a conversation so pronouns can refer to words spoken earlier, and the process of introducing new topics.

pragmatics The intent or motivation for an utterance. This is the underlying reason the utterance was spoken.

semantics The meaning of the words individually and their meaning as combined in a particular sentence.

syntax The rules governing the combination of words in a sentence, their parts of speech, and their forms, such as case and number.

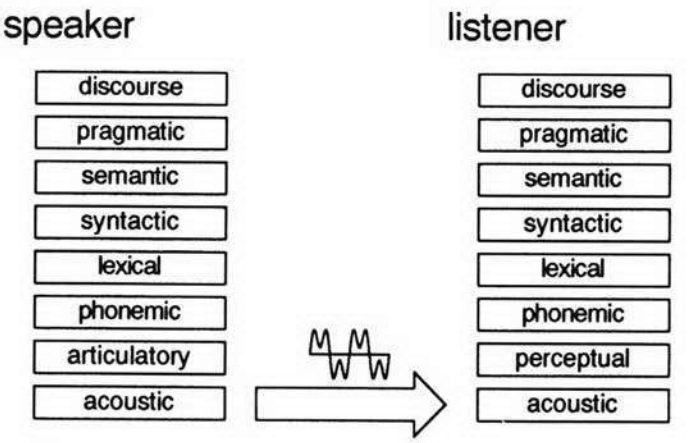

Figure 1.1. A layered view of speech communication.

lexical The set of words in a language, the rules for forming new words from affixes (prefixes and suffixes), and the stress ("accent") of syllables within the words.

phonetics The series of sounds that uniquely convey the series of words in the sentence.

articulation The motions or configurations of the vocal tract that produce the sounds, e.g., the tongue touching the lips or the vocal cords vibrating.

acoustics The realization of the string of phonemes in the sentence as vibrations of air molecules to produce pressure waves, i.e., sound.

Consider two hikers walking through the forest when one hiker's shoelace becomes untied. The other hiker sees this and says, "Hey, you're going to trip on your shoelace." The listener then ties the shoelace. We can consider this utterance at each layer of description.

Discourse analysis reveals that "Hey" serves to call attention to the urgency of the message and probably indicates the introduction of a new topic of conversation. It is probably spoken in a raised tone and an observer would reasonably expect the listener to acknowledge this utterance, either with a vocal response or by tying the shoe. Experience with discourse indicates that this is an appropriate interruption or initiation of a conversation at least under some circumstances. Discourse structure may help the listener understand that subsequent utterances refer to the shoelace instead of the difficulty of the terrain on which the conversants are traveling.

In terms of **pragmatics,** the speaker's intent is to warn the listener against tripping; presumably the speaker does not wish the listener to fall. But this utterance might also have been a ruse intended to get the listener to look down for the sake of playing a trick. We cannot differentiate these possibilities without know-

ing more about the context in which the sentence was spoken and the relationship between the conversants.

From a **semantics** standpoint, the sentence is about certain objects in the world: the listener, hiking, an article of clothing worn on the foot, and especially the string by which the boot is held on. The concern at the semantic layer is how the words refer to the world and what states of affairs they describe or predict. In this case, the meaning has to do with an animate entity ("you") performing some physical action ("tripping"), and the use of future tense indicates that the talker is making a prediction of something not currently taking place. Not all words refer to specific subjects; in the example, "Hey" serves to attract attention, but has no innate meaning.

Syntax is concerned with how the words fit together into the structure of the sentence. This includes the ordering of parts of speech (nouns, verbs, adjectives) and relations between words and the words that modify them. Syntax indicates that the correct word order is subject followed by verb, and syntax forces agreement of number, person, and case of the various words in the sentence. "You *is* going to . . ." is syntactically ill formed. Because the subject of the example is "you," the associated form of the verb "to be" is "are." The chosen verb form also indicates a future tense.

Lexical analysis tells us that "shoelace" comes from the root words "shoe" and "lace" and that the first syllable is stressed. Lexical analysis also identifies a set of definitions for each word taken in isolation. "Trip," for example, could be the act of falling or it could refer to a journey. Syntax reveals which definition is appropriate as each is associated with the word "trip" used as a different part of speech. In the example, "trip" is used as a verb and so refers to falling.

The **phonemic** layer is concerned with the string of phonemes of which the words are composed. Phonemes are the speech sounds that form the words of any language.[3] Phonemes include all the sounds associated with vowels and consonants. A grunt, growl, hiss, or gargling sound is not a phoneme in English, so it cannot be part of a word; such sounds are not referred to as speech. At the phoneme layer, while talking we are either continuously producing speech sounds or are silent. We are not silent at word boundaries; the phonemes all run together.

At the **articulatory** layer, the speaker makes a series of vocal gestures to produce the sounds that make up the phonemes. These sounds are created by a noise source at some location in the vocal tract, which is then modified by the configuration of the rest of the vocal tract. For example, to produce a "b" sound, the lips are first closed and air pressure from the lungs is built up behind them. A sudden release of air between the lips accompanied by vibration of the vocal cords produces the "b" sound. An "s" sound, by comparison, is produced by turbulence caused as a stream of air rushes through a constriction formed by the tongue and the roof of the mouth. The mouth can also be used to create nonspeech sounds, such as sighs and grunts.

[3]A more rigorous definition will be given in the next section.

Finally, the **acoustics** of the utterance is its nature as sound. Sound is transmitted as variations in air pressure over time; sound can be converted to an electrical signal by a microphone and represented as an electrical waveform. We can also analyze sound by converting the waveform to a spectrogram, which displays the various frequency components present in the sound. At the acoustic layer, speech is just another sound like the wind in the trees or a jet plane flying overhead.

From the perspective of the listener, the articulatory layer is replaced by a **perceptual** layer, which comprises the processes whereby sound (variations in air pressure over time) is converted to neural signals in the ear and ultimately interpreted as speech sounds in the brain. It is important to keep in mind that the hearer can directly sense only the acoustic layer of speech. If we send an electric signal representing the speech waveform over a telephone line and convert this signal to sound at the other end, the listening party can understand the speech. Therefore, the acoustic layer alone must contain all the information necessary to understand the speaker's intent, but it can be represented at the various layers as part of the process of understanding.

This layered approach is actually more descriptive than analytic in terms of human cognitive processes. The distinctions between the layers are fuzzy, and there is little evidence that humans actually organize discourse production into such layers. Intonation is interpreted in parallel at all these layers and thus illustrates the lack of sharp boundaries or sequential processing among them. At the pragmatic layer, intonation differentiates the simple question from exaggerated disbelief; the same words spoken with different intonation can have totally different meaning. At the syntactic layer, intonation is a cue to phrase boundaries. Intonation can differentiate the noun and verb forms of some words (e.g., conduct, convict) at the syntactic layer by conveying lexical stress. Intonation is not phonemic in English, but in some other languages a change in pitch does indicate a different word for the otherwise identical articulation. And intonation is articulated and realized acoustically in part as the fundamental frequency at which the vocal cords vibrate.

Dissecting the communication process into layers offers several benefits, both in terms of understanding as well as for practical implementations. Understanding this layering helps us appreciate the complexity and richness of speech. Research disciplines have evolved around each layer. A layered approach to representing conversation is essential for modular software development; a clean architecture isolates each module from the specialized knowledge of the others with information passed over well-defined interfaces. Because each layer consists of a different perspective on speech communication, each is likely to employ its own representation of speech for analysis and generation.

As a cautionary note, it needs to be recognized from the start that there is little evidence that humans actually function by invoking each of these layers during conversation. The model is descriptive without attempting to explain or identify components of our cognitive processes. The model is incomplete in that there are some aspects of speech communication that do not fit it, but it can serve as a framework for much of our discussion of conversational computer systems.

REPRESENTATIONS OF SPEECH

We need a means of describing and manipulating speech at each of these layers. Representations for the lower layers, such as acoustic waveforms or phonemes, are simpler and more complete and also more closely correspond to directly observable phenomena. Higher-layer representations, such as semantic or discourse structure, are subject to a great deal more argument and interpretation and are usually abstractions convenient for a computer program or a linguistic comparison of several languages. Any single representation is capable of conveying particular aspects of speech; different representations are suitable for discussion of the different layers of the communication process.

The representation chosen for any layer should contain all the information required for analysis at that layer. Since higher layers possess a greater degree of abstraction than lower layers, higher-layer representations extract features from lower-layer representations and hence lose the ability to recreate the original information completely. For example, numerous cues at the acoustic layer may indicate that the talker is female, but if we represent the utterance as a string of phones or of words we have lost those cues. In terms of computer software, the representation is the data type by which speech is described. One must match the representation and the particular knowledge about speech that it conveys to the algorithms employing it at a particular layer of speech understanding.

Acoustic Representations

Sounds consist of variations in air pressure over time at frequencies that we can hear. Speech consists of a subset of the sounds generated by the human vocal tract. If we wish to analyze a sound or save it to hear again later, we need to capture the variations in air pressure. We can convert air pressure to electric voltage with a microphone and then convert the voltage to magnetic flux on an audiocassette tape using a recording head, for example.

We can plot the speech signal in any of these media (air pressure, voltage, or magnetic flux) over time as a **waveform** as illustrated in Figure 1.2. This representation exhibits positive and negative values over time because the speech radiating from our mouths causes air pressure to be temporarily greater or less than that of the ambient air.

A waveform describing sound pressure in air is continuous, while the waveforms employed by computers are **digital,** or **sampled,** and have discrete values for each sample; these concepts are described in detail in Chapter 3. Tape recorders store analog waveforms; a compact audio disc holds a digital waveform. A digitized waveform can be made to very closely represent the original sound, and it can be captured easily with inexpensive equipment. A digitized sound stored in computer memory allows for fast random access. Once digitized, the sound may be further processed or compressed using digital signal processing techniques. The analog audiotape supports only sequential access (it must be rewound or fast-forwarded to jump to a different part of the tape) and is prone to mechanical breakdown.

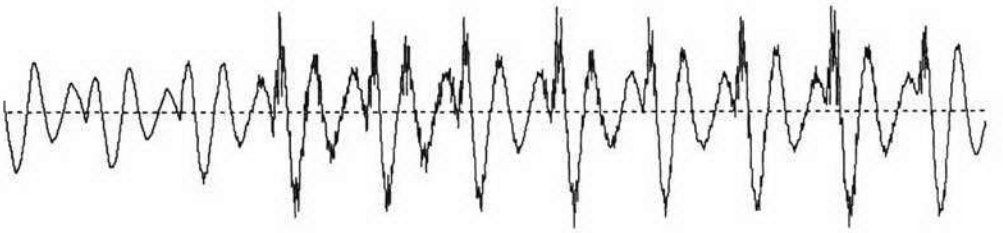

Figure 1.2. A waveform, showing 100 milliseconds of the word "me." The vertical axis depicts amplitude, and the horizontal axis represents time. The display depicts the transition from "m" to "e."

A waveform can effectively represent the original signal visually when plotted on a piece of paper or a computer screen. But to make observations about what a waveform sounds like, we must analyze it across a span of time not just at a single point. For example, in Figure 1.2 we can determine the amplitude ("volume") of the signal by looking at the differences between its highest and lowest points. We can also see that it is periodic: The signal repeats a pattern over and over. Since the horizontal axis represents time, we can determine the frequency of the signal by counting the number of periods in one second. A periodic sound with a higher frequency has a higher pitch than a periodic sound with a lower frequency.

One disadvantage of working directly with waveforms is that they require considerable storage space, making them bulky; Figure 1.2 shows only 100 milliseconds of speech. A variety of schemes for compressing speech to minimize storage are discussed in Chapter 3. A more crucial limitation is that a waveform simply shows the signal as a function of time. A waveform is in no way speech specific and can represent any acoustical phenomenon equally well. As a general-purpose representation, it contains all the acoustic information but does not explicitly describe its content in terms of properties of speech signals.

A speech-specific representation more succinctly conveys those features salient to speech and phonemes, such as syllable boundaries, fundamental frequency, and the higher-energy frequencies in the sound. A **spectrogram** is a transformation of the waveform into the frequency domain. As seen in Figure 1.3, the spectrogram reveals the distribution of various frequency components of the signal as a function of time indicating the energy at each frequency. The horizontal axis represents time, the vertical axis represents frequency, and the intensity or blackness at a point indicates the acoustic energy at that frequency and time.

A spectrogram still consists of a large amount of data but usually requires much less storage than the original waveform and reveals acoustic features specific to speech. Because of this, spectral analysis[4] is often employed to process

[4]To be precise, spectral or Fourier analysis uses mathematical techniques to derive the values of energy at particular frequencies. We can plot these as described above; this visual representation is the spectrogram.

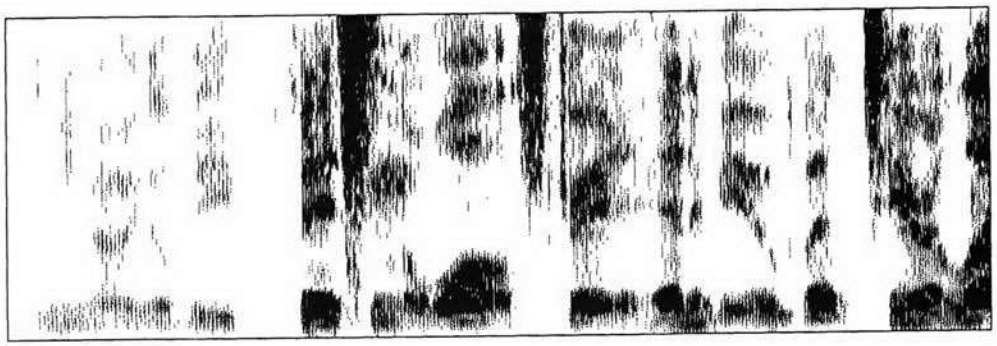

Figure 1.3. A spectrogram of 2.5 seconds of speech. The vertical axis is frequency, the horizontal axis is time, and energy maps to darkness.

speech for analysis by a human or a computer. People have been trained to read spectrograms and determine the words that were spoken. Although the spectrogram conveys salient features of a sound, the original acoustic signal cannot be reconstructed from it without some difficulty. As a result, the spectrogram is more useful for analysis of the speech signal than as a means of storing it for later playback.

Other acoustic representations in the frequency domain are even more succinct, though they are more difficult for a human to process visually than a spectrogram. **Linear Prediction Coefficients** and **Cepstral Analysis,** for example, are two such techniques that rely heavily on digital signal processing.[5] Both of these techniques reveal the resonances of the vocal tract and separate information about how the sound was produced at the noise source and how it was modified by various parts of the vocal tract. Because these two techniques extract salient information about how the sound was articulated they are frequently used as representations for computer analysis of speech.

PHONEMES AND SYLLABLES

Two representations of speech which are more closely related to its lexical structure are phonemes and syllables. Phonemes are important small units, several of which make up most syllables.

Phonemes

A **phoneme** is a unit of speech, the set of which defines all the sounds from which words can be constructed in a particular language. There is at least one pair of

[5]Linear prediction will be explained in Chapter 3. Cepstral analysis is beyond the scope of this book.

words in a language for which replacing one phoneme with another will change what is spoken into a different word. Of course, not every combination of phonemes results in a word; many combinations are nonsense.

For example, in English, the words "bit" and "bid" have different meanings, indicating that the "t" and "d" are different phonemes. Two words that differ in only a single phoneme are called a **minimal pair.** "Bit" and "bif" also vary by one sound, but this example does not prove that "t" and "f" are distinct phonemes as "bif" is not a word. But "tan" and "fan" are different words; this proves the phonemic difference between "t" and "f".

Vowels are also phonemes; the words "heed," "had," "hid," "hide," "howed," and "hood" each differ only by one sound, showing us that English has at least six vowel phonemes. It is simple to construct a minimal pair for any two vowels in English, while it may not be as simple to find a pair for two consonants.

An **allophone** is one of a number of different ways of pronouncing the same phoneme. Replacing one allophone of a phoneme with another does not change the meaning of a sentence, although the speaker will sound unnatural, stilted, or like a non-native. For example, consider the "t" sound in "sit" and "sitter." The "t" in "sit" is somewhat aspirated; a puff of air is released with the consonant. You can feel this if you put your hand in front of your mouth as you say the word. But in "sitter" the same phoneme is not aspirated; we say the aspiration is not phonemic for "t" and conclude that we have identified two allophones. If you aspirate the "t" in "sitter," it sounds somewhat forced but does not change the meaning of the word.

In contrast, aspiration of stop consonants is phonemic in Nepali. For an example of English phonemes that are allophones in another language, consider the difficulties Japanese speakers have distinguishing our "l" and "r." The reason is simply that while these are two phonemes in English, they are allophonic variants on the same phoneme in Japanese. Each language has its own set of phonemes and associated allophones. Sounds that are allophonic in one language may be phonemic in another and may not even exist in a third. When you learn a language, you learn its phonemes and how to employ the permissible allophonic variations on them. But learning phonemes is much more difficult as an adult than as a child.

Because phonemes are language specific, we can not rely on judgments based solely on our native languages to classify speech sounds. An individual speech sound is a **phone,** or **segment.** For any particular language, a given phoneme will have a set of allophones, each of which is a segment. Segments are properties of the human vocal mechanism, and phonemes are properties of languages. For most practical purposes, phone and phoneme may be considered to be synonyms.

Linguists use a notation for phones called the **International Phonetic Alphabet,** or **IPA.** IPA has a symbol for almost every possible phone; some of these symbols are shown in Figure 1.4. Since there are far more than 26 such phones, it is not possible to represent them all with the letters of the English alphabet. IPA borrows symbols from the Greek alphabet and elsewhere. For example, the "th" sound in "thin" is represented as "θ" in IPA, and the sound

Phoneme	Example Word	Phoneme	Example Word	Phoneme	Example Word
i	beet	p	put	č	chin
I	bit	t	tap	ǰ	judge
ε	bet	k	cat	m	map
e	bait	b	bit	n	nap
æ	bat	d	dill	η	sing
α	cot	g	get	r	ring
ↄ	caught	f	fain	l	lip
ʌ	but	θ	thin	w	will
o	boat	s	sit	y	yell
U	foot	ʃ	shoe	h	head
u	boot	v	veal		
ɜ	bird	δ	then		
αj (αI)	bite	z	zeal		
ↄj (ↄI)	boy	ž	azure		
αw (αU)	bout				
ə	about				

Figure 1.4. The English phonemes in IPA, the International Phonetic Alphabet.

in "then" is "δ."[6] American linguists who use computers have developed the **Arpabet,** which uses ordinary alphabet characters to represent phones; some phonemes are represented as a pair of letters. Arpabet[7] was developed for the convenience of computer manipulation and representation of speech using ASCII-printable characters.

To avoid the necessity of the reader learning either IPA or Arpabet, this book indicates phones by example, such as, "the 't' in bottle." Although slightly awkward, such a notation suffices for the limited examples described. The reader will find it necessary to learn a notation (IPA is more common in textbooks) to make any serious study of phonetics or linguistics.

A phonemic transcription, although compact, has lost much of the original signal content, such as pitch, speed, and amplitude of speech. Phonemes are abstractions from the original signal that highlight the speech-specific aspects of that

[6]Are these sounds phonemic in English?

[7]The word comes from the acronym ARPA, the Advanced Research Projects Agency (sometimes called DARPA), a research branch of the U.S. Defense Department that funded much early speech research in this country and continues to be the most significant source of government support for such research.

signal; this makes a phonemic transcription a concise representation for lexical analysis as it is much more abstract than the original waveform.

Syllables

Another natural way to divide speech sounds is by the **syllable.** Almost any native speaker of English can break a word down into syllables, although some words can be more difficult, e.g., "chocolate" or "factory." A syllable consists of one or more consonants, a vowel (or diphthong[8]), followed by one or more consonants; consonants are optional, but the vowel is not. Two or more adjacent consonants are called a consonant **cluster;** examples are the initial sounds of "screw" and "sling." Acoustically, a syllable consists of a relatively high energy core (the vowel) optionally preceded or followed by periods of lower energy (consonants). Consonants have lower energy because they impose constrictions on the air flow from the lungs. Many natural languages, such as those written with the Arabic script and the northern Indian languages, are written with a syllabic system in which one symbol represents both a consonant and its associated vowel.

Other Representations

There are many other representations of speech appropriate to higher layer aspects of conversation.[9] Lexical analysis reveals an utterance as a series of **words.** A dictionary, or **lexicon** lists all the words of a language and their meanings. The **phrase,** sometimes called a "breath group" when describing intonation, is relevant both to the study of prosody (pitch, rhythm, and meter) as well as syntax, which deals with structures such as the noun phrase and verb phrase. A **parse tree,** as shown in Figure 1.5, is another useful representation of the syntactic relationships among words in a sentence.

Representations for higher layers of analysis are varied and complex. Semantics associates meaning with words, and meaning implies a relationship to other words or concepts. A **semantic network** indicates the logical relationships between words and meaning. For example, a door is a physical object, but it has specific meaning only in terms of other objects, such as walls and buildings, as it covers entrance holes in these objects.

Discourse analysis has produced a variety of models of the focus of a conversation. For example, one of these uses a **stack** to store potential topics of current focus. New topics are pushed onto the stack, and a former topic again becomes the focus when all topics above it are popped off the stack. Once removed from the stack, a topic cannot become the focus without being reintroduced.

[8]A diphthong consists of two vowel sounds spoken in sequence and is considered a single phoneme. The two vowel sounds in a diphthong cannot be separated into different syllables. The vowels in "hi" and "bay" are examples of diphthongs.

[9]Most of the speech representations mentioned in this section will be detailed in Chapter 9.

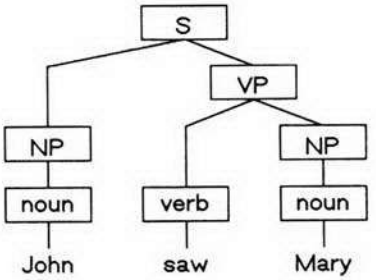

Figure 1.5. A parse tree.

SUMMARY

The perspective of speech as a conversational process constitutes the foundation and point of view of this book. Conversations employ spontaneous speech and a variety of interaction techniques to coordinate the exchange of utterances between participants. Speech is our primary medium of communication, although writing may be favored for longer-lasting or more formal messages. Computer systems would do well to exploit the richness and robustness of human conversation.

But "conversation" is simply too rich and varied a term to be amenable to analysis without being analyzed in components. This chapter identified a number of such components of a conversation: acoustic, articulatory, phonetic, lexical, syntactic, semantic, pragmatic, and discourse. Disciplines of research have been established for each of these areas, and the rest of this book will borrow heavily from them. Each discipline has its own set of representations of speech, which allow utterances to be described and analyzed.

Representations such as waveforms, spectrograms, and phonetic transcriptions provide suitable abstractions that can be embedded in the computer programs that attempt to implement various layers of speech communication. Each representation highlights particular features or characteristics of speech and may be far removed from the original speech sounds.

The rationale for this book is that the study of speech in conversations is interdisciplinary and that the designers of conversational computer systems need to understand each of these individual components in order to fully appreciate the whole. The rest of this book is organized in part as a bottom-up analysis of the layers of speech communication. Each layer interacts with the other layers, and the underlying goal for conversational communication is unification of the layers.