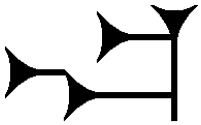
Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright © 2019. All rights reserved. Draft of October 16, 2019.

CHAPTER

27 Phonetics

The characters or letters that are the basis of all the text-based methods we've seen so far in this book aren't just random symbols. They are also an amazing scientific invention: a theoretical model of the elements that make up human speech.

The earliest independently invented writing systems (Sumerian, Chinese, Mayan) were mainly logographic, which means one symbol representing a whole word. But from the earliest stages we can find, some of the symbols also represent the sounds that make up the words. Thus, the cuneiform sign to the right pronounced *ba* and meaning "ration" in Sumerian could also function purely as the sound /ba/ in languages that used cuneiform. Chinese writing, from its early instantiations on oracle bones, also assigns phonetic meaning to many character elements. Purely sound-based writing systems, whether syllabic (like Japanese *hiragana* or *katakana*), alphabetic (like the Roman alphabet used in this book), or consonantal (like Semitic writing systems), can generally be traced back to these early logo-syllabic systems, often as two cultures came together. Thus, the Arabic, Aramaic, Hebrew, Greek, and Roman systems all derive from a West Semitic script that is presumed to have been modified by Western Semitic mercenaries from a cursive form of Egyptian hieroglyphs. The Japanese syllabaries were modified from a cursive form of a set of Chinese characters that represented sounds. These Chinese characters themselves were used in Chinese to phonetically represent the Sanskrit in the Buddhist scriptures that were brought to China in the Tang dynasty.



Whatever its origins, the idea implicit in a sound-based writing system—that the spoken word is composed of smaller units of speech—underlies the modern algorithms for **speech recognition** (transcribing acoustic waveforms into strings of text words) and **speech synthesis** or **text-to-speech** (converting strings of text words into acoustic waveforms).

phonetics In this chapter we introduce **phonetics** from a computational perspective. Phonetics is the study of the speech sounds used in the languages of the world, how they are produced by the articulators of the human vocal tract, how they are realized acoustically, and how this acoustic realization can be digitized and processed.

27.1 Speech Sounds and Phonetic Transcription

Although a letter like 'p' or 'a' is a useful rough model of the sounds of human speech, in speech processing we often model the pronunciation of a word instead as a string of **phones**. A phone is a speech sound, represented with phonetic symbols

phone

modeled on letters in the Roman alphabet.

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[p]	[p]	parsley	[p aa r s l iy]
[t]	[t]	tea	[t iy]
[k]	[k]	cook	[k uh k]
[b]	[b]	bay	[b ey]
[d]	[d]	dill	[d ih l]
[g]	[g]	garlic	[g aa r l ix k]
[m]	[m]	mint	[m ih n t]
[n]	[n]	nutmeg	[n ah t m eh g]
[ng]	[ŋ]	baking	[b ey k ix ng]
[f]	[f]	flour	[f l aw axr]
[v]	[v]	clove	[k l ow v]
[th]	[θ]	thick	[th ih k]
[dh]	[ð]	those	[dh ow z]
[s]	[s]	soup	[s uw p]
[z]	[z]	eggs	[eh g z]
[sh]	[ʃ]	squash	[s k w aa sh]
[zh]	[ʒ]	ambrosia	[ae m b r ow zh ax]
[ch]	[tʃ]	cherry	[ch eh r iy]
[jh]	[dʒ]	jar	[jh aa r]
[l]	[l]	licorice	[l ih k axr ix sh]
[w]	[w]	kiwi	[k iy w iy]
[r]	[r]	rice	[r ay s]
[y]	[j]	yellow	[y eh l ow]
[h]	[h]	honey	[h ah n iy]

Figure 27.1 ARPAbet symbols for transcribing English consonants, with IPA equivalents.

This section surveys the different phones of English, focusing on American English. The **International Phonetic Alphabet** (**IPA**) is an evolving standard originally developed by the International Phonetic Association in 1888 with the goal of transcribing the sounds of all human languages. The ARPAbet (Shoup, 1980) is a phonetic alphabet designed for American English that uses ASCII symbols; it can be thought of as a convenient ASCII representation of an American-English subset of the IPA. Because the ARPAbet is common for computational modeling, we rely on it here. Figures 27.1 and 27.2 show the ARPAbet symbols for transcribing consonants and vowels, respectively, together with their IPA equivalents.

IPA

Many of the IPA and ARPAbet symbols are equivalent to familiar Roman letters. So, for example, the ARPAbet phone [p] represents the consonant sound at the beginning of *platypus*, *puma*, and *plantain*, the middle of *leopard*, or the end of *antelope*. In general, however, the mapping between the letters of English orthography and phones is relatively **opaque**; a single letter can represent very different sounds in different contexts. The English letter *c* corresponds to phone [k] in *cougar* [k uw g axr], but phone [s] in *cell* [s eh l]. Besides appearing as *c* and *k*, the phone [k] can appear as part of *x* (*fox* [f aa k s]), as *ck* (*jackal* [jh ae k el]) and as *cc* (*raccoon* [r ae k uw n]). Many other languages, for example, Spanish, are much more **transparent** in their sound-orthography mapping than English.

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[ae s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oʊ]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɝ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[oɪ]	soil	[s oy l]

Figure 27.2 ARPAbet symbols for American English vowels, with IPA equivalents.

27.2 Articulatory Phonetics

articulatory phonetics

Articulatory phonetics is the study of how these phones are produced as the various organs in the mouth, throat, and nose modify the airflow from the lungs.

27.2.1 The Vocal Organs

Figure 27.3 shows the organs of speech. Sound is produced by the rapid movement of air. Humans produce most sounds in spoken languages by expelling air from the lungs through the windpipe (technically, the **trachea**) and then out the mouth or nose. As it passes through the trachea, the air passes through the **larynx**, commonly known as the Adam's apple or voice box. The larynx contains two small folds of muscle, the **vocal folds** (often referred to non-technically as the **vocal cords**), which can be moved together or apart. The space between these two folds is called the **glottis**. If the folds are close together (but not tightly closed), they will vibrate as air passes through them; if they are far apart, they won't vibrate. Sounds made with the vocal folds together and vibrating are called **voiced**; sounds made without this vocal cord vibration are called **unvoiced** or **voiceless**. Voiced sounds include [b], [d], [g], [v], [z], and all the English vowels, among others. Unvoiced sounds include [p], [t], [k], [f], [s], and others.

glottis

voiced sound

unvoiced sound

The area above the trachea is called the **vocal tract**; it consists of the **oral tract** and the **nasal tract**. After the air leaves the trachea, it can exit the body through the mouth or the nose. Most sounds are made by air passing through the mouth. Sounds made by air passing through the nose are called **nasal sounds**; nasal sounds use both the oral and nasal tracts as resonating cavities; English nasal sounds include [m], [n], and [ng].

nasal

Phones are divided into two main classes: **consonants** and **vowels**. Both kinds of sounds are formed by the motion of air through the mouth, throat or nose. Consonants are made by restriction or blocking of the airflow in some way, and can be voiced or unvoiced. Vowels have less obstruction, are usually voiced, and are generally louder and longer-lasting than consonants. The technical use of these terms is

consonant

vowel

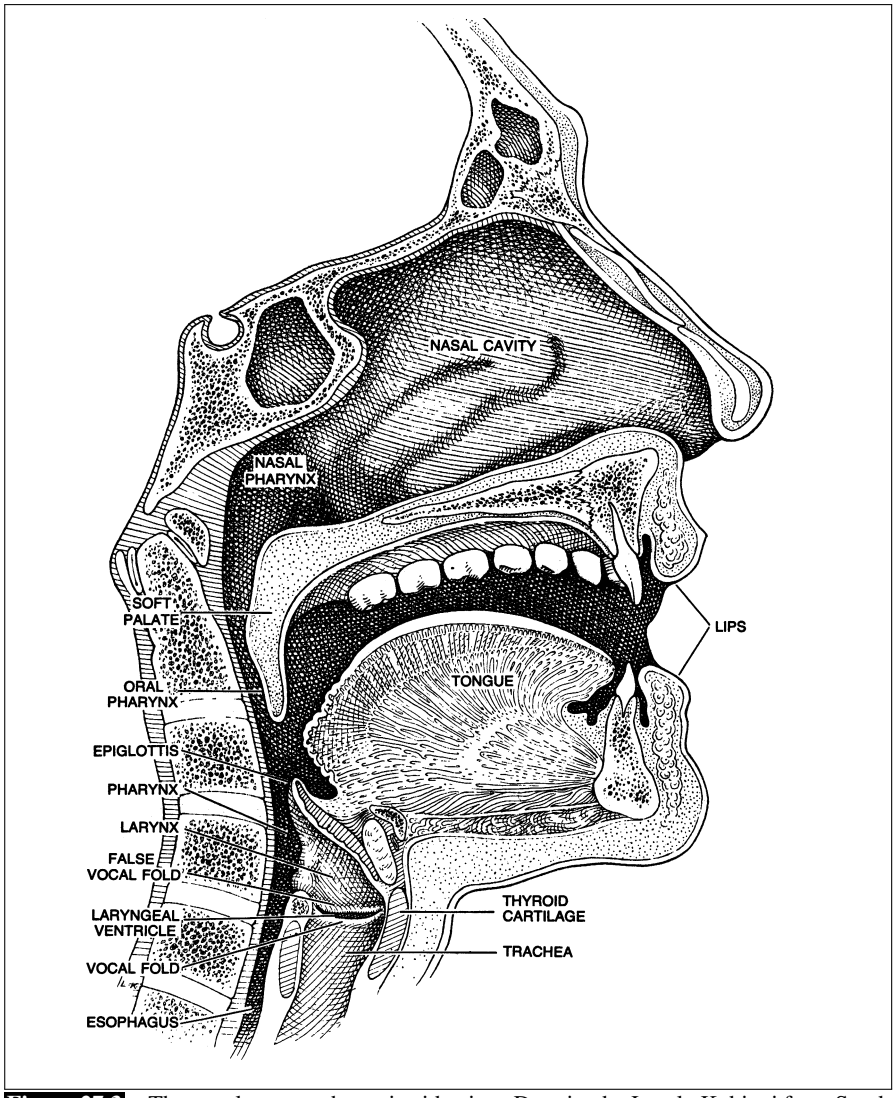


Figure 27.3 The vocal organs, shown in side view. Drawing by Laszlo Kubinyi from Sundberg (1977), ©Scientific American, used by permission.

much like the common usage; [p], [b], [t], [d], [k], [g], [f], [v], [s], [z], [r], [l], etc., are consonants; [aa], [ae], [ao], [ih], [aw], [ow], [uw], etc., are vowels. **Semivowels** (such as [y] and [w]) have some of the properties of both; they are voiced like vowels, but they are short and less syllabic like consonants.

27.2.2 Consonants: Place of Articulation

Because consonants are made by restricting the airflow in some way, consonants can be distinguished by where this restriction is made: the point of maximum restriction is called the **o**f a consonant. Places of articulation, shown in Fig. 27.4, can be a useful way of grouping phones into equivalence classes, described below.

place of articulation

labial

Labial: Consonants whose main restriction is formed by the two lips coming together have a **bilabial** place of articulation. In English these include [p] as in *possum*, [b] as in *bear*, and [m] as in *marmot*. The English **labiodental**

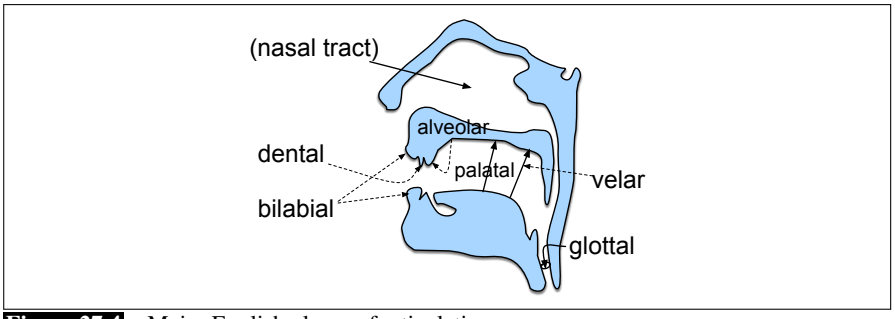


Figure 27.4 Major English places of articulation.

consonants [v] and [f] are made by pressing the bottom lip against the upper row of teeth and letting the air flow through the space in the upper teeth.

dental **Dental:** Sounds that are made by placing the tongue against the teeth are dentals. The main dentals in English are the [th] of *thing* and the [dh] of *though*, which are made by placing the tongue behind the teeth with the tip slightly between the teeth.

alveolar **Alveolar:** The alveolar ridge is the portion of the roof of the mouth just behind the upper teeth. Most speakers of American English make the phones [s], [z], [t], and [d] by placing the tip of the tongue against the alveolar ridge. The word **coronal** is often used to refer to both dental and alveolar.

palatal palate **Palatal:** The roof of the mouth (the **palate**) rises sharply from the back of the alveolar ridge. The **palato-alveolar** sounds [sh] (*shrimp*), [ch] (*china*), [zh] (*Asian*), and [jh] (*jar*) are made with the blade of the tongue against the rising back of the alveolar ridge. The palatal sound [y] of *yak* is made by placing the front of the tongue up close to the palate.

velar **Velar:** The **velum**, or soft palate, is a movable muscular flap at the very back of the roof of the mouth. The sounds [k] (*cuckoo*), [g] (*goose*), and [ŋ] (*kingfisher*) are made by pressing the back of the tongue up against the velum.

glottal **Glottal:** The glottal stop [q] (IPA [ʔ]) is made by closing the glottis (by bringing the vocal folds together).

27.2.3 Consonants: Manner of Articulation

Consonants are also distinguished by *how* the restriction in airflow is made, for example, by a complete stoppage of air or by a partial blockage. This feature is called the **manner of articulation** of a consonant. The combination of place and manner of articulation is usually sufficient to uniquely identify a consonant. Following are the major manners of articulation for English consonants:

manner of articulation

stop A **stop** is a consonant in which airflow is completely blocked for a short time. This blockage is followed by an explosive sound as the air is released. The period of blockage is called the **closure**, and the explosion is called the **release**. English has voiced stops like [b], [d], and [g] as well as unvoiced stops like [p], [t], and [k]. Stops are also called **plosives**.

nasal The **nasal** sounds [n], [m], and [ng] are made by lowering the velum and allowing air to pass into the nasal cavity.

fricatives In **fricatives**, airflow is constricted but not cut off completely. The turbulent airflow that results from the constriction produces a characteristic "hissing" sound. The English labiodental fricatives [f] and [v] are produced by pressing the lower lip against the upper teeth, allowing a restricted airflow between the upper teeth.

The dental fricatives [th] and [dh] allow air to flow around the tongue between the teeth. The alveolar fricatives [s] and [z] are produced with the tongue against the alveolar ridge, forcing air over the edge of the teeth. In the palato-alveolar fricatives [sh] and [zh], the tongue is at the back of the alveolar ridge, forcing air through a groove formed in the tongue. The higher-pitched fricatives (in English [s], [z], [sh] and [zh]) are called **sibilants**. Stops that are followed immediately by fricatives are called **affricates**; these include English [ch] (*chicken*) and [jh] (*giraffe*).

sibilants

approximant

In **approximants**, the two articulators are close together but not close enough to cause turbulent airflow. In English [y] (*yellow*), the tongue moves close to the roof of the mouth but not close enough to cause the turbulence that would characterize a fricative. In English [w] (*wood*), the back of the tongue comes close to the velum. American [r] can be formed in at least two ways; with just the tip of the tongue extended and close to the palate or with the whole tongue bunched up near the palate. [l] is formed with the tip of the tongue up against the alveolar ridge or the teeth, with one or both sides of the tongue lowered to allow air to flow over it. [l] is called a **lateral** sound because of the drop in the sides of the tongue.

tap

A **tap** or **flap** [dx] (or IPA [ɾ]) is a quick motion of the tongue against the alveolar ridge. The consonant in the middle of the word *lotus* ([l ow dx ax s]) is a tap in most dialects of American English; speakers of many U.K. dialects would use a [t] instead of a tap in this word.

27.2.4 Vowels

Like consonants, vowels can be characterized by the position of the articulators as they are made. The three most relevant parameters for vowels are what is called vowel **height**, which correlates roughly with the height of the highest part of the tongue, vowel **frontness** or **backness**, indicating whether this high point is toward the front or back of the oral tract and whether the shape of the lips is **rounded** or not. Figure 27.5 shows the position of the tongue for different vowels.

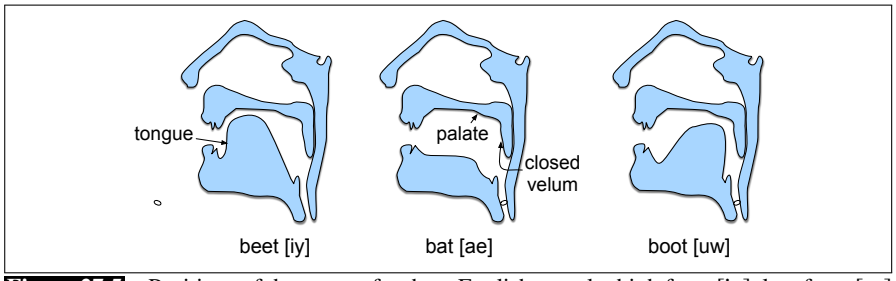


Figure 27.5 Positions of the tongue for three English vowels: high front [iy], low front [ae] and high back [uw].

In the vowel [iy], for example, the highest point of the tongue is toward the front of the mouth. In the vowel [uw], by contrast, the high-point of the tongue is located toward the back of the mouth. Vowels in which the tongue is raised toward the front are called **front vowels**; those in which the tongue is raised toward the back are called **back vowels**. Note that while both [ih] and [eh] are front vowels, the tongue is higher for [ih] than for [eh]. Vowels in which the highest point of the tongue is comparatively high are called **high vowels**; vowels with mid or low values of maximum tongue height are called **mid vowels** or **low vowels**, respectively.

Front vowel

Back vowel

High vowel

Figure 27.6 shows a schematic characterization of the height of different vowels. It is schematic because the abstract property **height** correlates only roughly with ac-

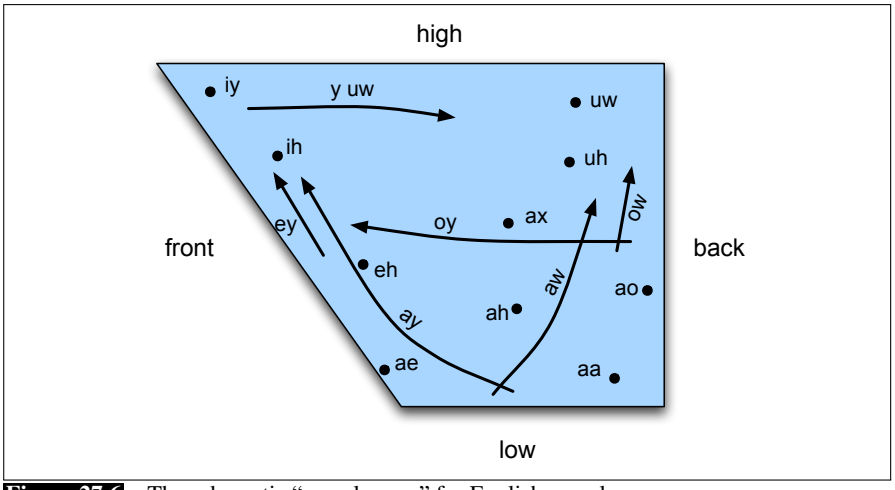


Figure 27.6 The schematic "vowel space" for English vowels.

tual tongue positions; it is, in fact, a more accurate reflection of acoustic facts. Note that the chart has two kinds of vowels: those in which tongue height is represented as a point and those in which it is represented as a path. A vowel in which the tongue position changes markedly during the production of the vowel is a **diphthong**. English is particularly rich in diphthongs.

diphthong

The second important articulatory dimension for vowels is the shape of the lips. Certain vowels are pronounced with the lips rounded (the same lip shape used for whistling). These **rounded** vowels include [uw], [ao], and [ow].

rounded vowel

27.2.5 Syllables

syllable

Consonants and vowels combine to make a **syllable**. A syllable is a vowel-like (or **sonorant**) sound together with some of the surrounding consonants that are most closely associated with it. The word *dog* has one syllable, [d aa g] (in our dialect); the word *catnip* has two syllables, [k ae t] and [n ih p]. We call the vowel at the core of a syllable the **nucleus**. The optional initial consonant or set of consonants is called the **onset**. If the onset has more than one consonant (as in the word *strike* [s t r ay k]), we say it has a **complex onset**. The **coda** is the optional consonant or sequence of consonants following the nucleus. Thus [d] is the onset of *dog*, and [g] is the coda. The **rime**, or **rhyme**, is the nucleus plus coda. Figure 27.7 shows some sample syllable structures.

nucleus
onset
coda
rime

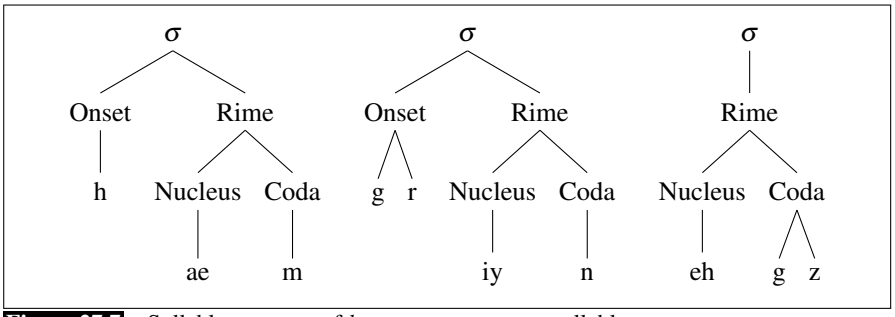


Figure 27.7 Syllable structure of *ham*, *green*, *eggs*. σ=syllable.

syllabification
phonotactics

The task of automatically breaking up a word into syllables is called **syllabification**. Syllable structure is also closely related to the **phonotactics** of a language. The term **phonotactics** means the constraints on which phones can follow each other in a language. For example, English has strong constraints on what kinds of consonants can appear together in an onset; the sequence [zdr], for example, cannot be a legal English syllable onset. Phonotactics can be represented by a language model or finite-state model of phone sequences.

27.3 Prosodic Prominence: Accent, Stress and Schwa

prominence

In a natural utterance of American English, some words sound more **prominent** than others, and certain syllables in these words are also more **prominent** than others. What we mean by prominence is that these words or syllables are perceptually more salient to the listener; speakers make a word or syllable more salient in English by saying it louder, saying it slower (so it has a longer duration), or by varying F0 during the word, making it higher or more variable.

pitch accent

We capture the core notion of prominence by associating a linguistic marker with prominent words and syllables, a marker called **pitch accent**. Words or syllables that are prominent are said to **bear** (be associated with) a pitch accent. Pitch accent is thus part of the phonological description of a word in context in a spoken utterance.

Thus this utterance might be pronounced by **accenting** the underlined words:

(27.1) I'm a little <u>surprised</u> to hear it <u>characterized</u> as <u>happy</u>.

Nuclear Accent

emphatic accent

We generally need more fine-grained distinctions than just a binary distinction between accented and unaccented words. For example, the last accent in a phrase generally is perceived as being more prominent than the other accents. This prominent last accent is called the **nuclear** or **emphatic accent**. Emphatic accents are generally used for semantic purposes, such as marking a word as the focus of the sentence or as contrastive or otherwise important in some way. Such emphatic words are often written IN CAPITAL LETTERS or with **stars** around them in texts or email or *Alice in Wonderland*; here's an example from the latter:

(27.2) "I know SOMETHING interesting is sure to happen," she said to herself.

Lexical Stress

lexical stress

The syllables that bear pitch accent are called **accented** syllables, but not every syllable of a word can be accented. Pitch accent has to be realized on the syllable that has **lexical stress**. Lexical stress is a property of the words' pronunciation in dictionaries; the syllable that has lexical stress is the one that will be louder or longer if the word is accented. For example, the word *surprised* is stressed on its second syllable, not its first. (try stressing the other syllable by saying SURprised; hopefully that sounds wrong to you). Thus, if the word *surprised* receives a pitch accent in a sentence, it is the second syllable that will be stronger. The following example shows underlined accented words with the stressed syllable bearing the accent (the louder, longer syllable) in boldface:

(27.3) I'm a little <u>sur**prised**</u> to hear it <u>**char**acterized</u> as <u>**hap**py</u>.

IPR2026-00274
Krisp Technologies EX1029 Page 8

Stress can be marked in dictionaries in various ways. The CMU dictionary (CMU, 1993), for example, marks each vowel with the number 0 (unstressed), 1 (stressed), or 2 (secondary stress). Thus, the word *counter* is listed as [K AW1 N T ER0] and the word *table* as [T EY1 B AH0 L]. **secondary stress** is defined as a level of stress lower than primary stress but higher than an unstressed vowel, as in the word *dictionary* [D IH1 K SH AH0 N EH2 R IY0]. Difference in lexical stress can affect word meaning. For example the word *content* can be a noun or an adjective, but have different stressed syllables (the noun is pronounced [K AA1 N T EH0 N T] and the adjective [K AA0 N T EH1 N T]). In IPA, on the other hand, the symbol [ˈ] before a syllable indicates that it has lexical stress (e.g., [ˈpar.sli]).

secondary stress

Reduced Vowels and Schwa

reduced vowel
schwa

Vowels that are unstressed can be weakened even further to **reduced vowels**. The most common reduced vowel is **schwa** ([ax]). Reduced vowels in English don't have their full form; the articulatory gesture isn't as complete as for a full vowel. As a result, the shape of the mouth is somewhat neutral; the tongue is neither particularly high nor low. The second vowel in *parakeet* is a schwa: [p ae r ax k iy t].

While schwa is the most common reduced vowel, it is not the only one, at least not in some dialects (Bolinger, 1981). Besides [ax], the ARPAbet also includes a reduced front vowel [ix] (IPA [ɨ]), as well as [axr], which is an r-colored schwa (often called **schwar**).[1] Fig. 27.8 shows these reduced vowels.

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[ax]	[ə]	lotus	[l ow dx ax s]
[axr]	[ɚ]	heather	[h eh dh axr]
[ix]	[ɨ]	tulip	[t uw l ix p]

Figure 27.8 Reduced vowels in American English, ARPAbet and IPA. [ax] is the reduced vowel schwa, [ix] is the reduced vowel corresponding to [ih], and [axr] is the reduced vowel corresponding to [er].

Not all unstressed vowels are reduced; any vowel, and diphthongs in particular, can retain its full quality even in unstressed position. For example, the vowel [iy] can appear in stressed position as in the word *eat* [iy t] or in unstressed position as in the word *carry* [k ae r iy].

prominence

We have mentioned a number of potential levels of **prominence**: accented, stressed, secondary stress, full vowel, and reduced vowel. It is still an open research question exactly how many levels are appropriate. Very few computational systems make use of all five of these levels, most using between one and three.

27.4 Prosodic Structure and Tune

prosody

In poetry, the word **prosody** refers to the study of the metrical structure of verse. In language processing, however, we use the term **prosody** to mean the study of the intonational and rhythmic aspects of language. More technically, prosody has been defined by Ladd (1996) as the "use of suprasegmental features to convey sentence-level pragmatic meanings". The term **suprasegmental** means above and beyond the

suprasegmental

[1] [ix] is generally dropped in computational applications (Miller, 1998), and [ax] and [ix] are falling together in many dialects of English (Wells, 1982, p. 167–168).

level of the segment or phone. The term refers especially to the uses of acoustic features like **F0**, **duration**, and **energy** independently of the phone string.

By **sentence-level pragmatic meaning**, Ladd is referring to a number of kinds of meaning that have to do with the relation between a sentence and its discourse or external context. For example, prosody can be used to mark **discourse structure or function**, like the difference between statements and questions, or the way that a conversation is structured into segments or subdialogs. Prosody is also used to mark **saliency**, such as indicating that a particular word or phrase is important or salient. Finally, prosody is heavily used for affective and emotional meaning, such as expressing happiness, surprise, or anger.

The kind of prosodic prominence, that we saw in the prior section is one of the most computational studied aspects of prosody, but there are two others that we introduce in this section: **prosodic structure** and **tune**.

27.4.1 Prosodic Structure

Spoken sentences have prosodic structure in the sense that some words seem to group naturally together and some words seem to have a noticeable break or disjuncture between them. Prosodic structure is often described in terms of **prosodic phrasing**, meaning that an utterance has a prosodic phrase structure in a similar way to it having a syntactic phrase structure. For example, in the sentence *I wanted to go to London, but could only get tickets for France* there seem to be two main **intonation phrases**, their boundary occurring at the comma. Furthermore, in the first phrase, there seems to be another set of lesser prosodic phrase boundaries (often called **intermediate phrase**s) that split up the words as *I wanted | to go | to London*.

prosodic phrasing

intonation phrase

intermediate phrase

There is also a correlation between prosodic structure and **syntactic structure** (Price et al. 1991, Ostendorf and Veilleux 1994, Koehn et al. 2000).

27.4.2 Tune

Two utterances with the same prominence and phrasing patterns can still differ prosodically by having different **tunes**. The **tune** of an utterance is the rise and fall of its F0 over time. A very obvious example of tune is the difference between statements and yes-no questions in English. The same sentence can be said with a final rise in F0 to indicate a yes-no question, or a final fall in F0 to indicate a declarative intonation. Figure 27.9 shows the F0 track of the same words spoken as a question or a statement. Note that the question rises at the end; this is often called a **question rise**. The falling intonation of the statement is called a **final fall**.

tune

question rise

final fall

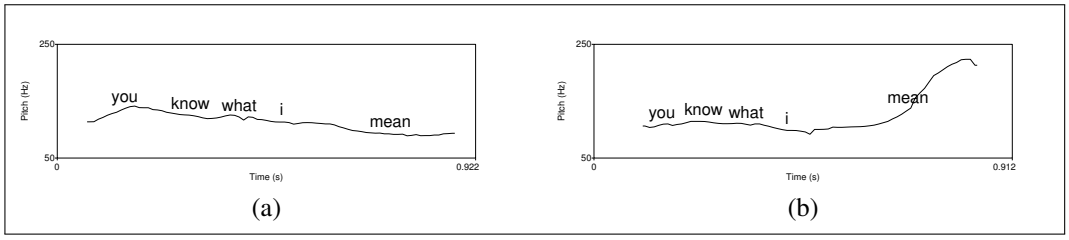


Figure 27.9 The same text read as the statement *You know what I mean* (on the left) and as a question *You know what I mean?* (on the right). Notice that yes-no question intonation in English has a sharp final rise in F0.

It turns out that English makes wide use of tune to express meaning. Besides this well-known rise for yes-no questions, an English phrase containing a list of nouns

continuation rise

separated by commas often has a short rise called a **continuation rise** after each noun. Other examples include the characteristic English contours for expressing **contradiction** and expressing **surprise**.

The mapping between meaning and tune in English is extremely complex. Consider the utterance *oh, really*. Without varying the phrasing or stress, it is still possible to have many variants of this by varying the intonational tune. For example, we might have an excited version *oh, really!* (in the context of a reply to a statement that you've just won the lottery); a sceptical version *oh, really?*—in the context of not being sure that the speaker is being honest; to an angry *oh, really!* indicating displeasure.

Linking Tune with Prominence: ToBI

It is also possible to link models of prominence with models of tune, allowing us to model differences between pitch accents according to the **tune** associated with them.

ToBI

One of the most widely used linguistic models of prosody that enables this association is the **ToBI** (Tone and Break Indices) model (Silverman et al. 1992, Beckman and Hirschberg 1994, Pierrehumbert 1980, Pitrelli et al. 1994). ToBI is a phonological theory of intonation that models prominence, tune, and boundaries. ToBI's model of prominence and tunes is based on the five **pitch accents** and four **boundary tones** shown in Fig. 27.10.

Pitch Accents		Boundary Tones	
H*	peak accent	**L-L%**	"final fall": "declarative contour" of American English
L*	low accent	**L-H%**	continuation rise
L*+H	scooped accent	**H-H%**	"question rise": cantonical yes-no question contour
L+H*	rising peak accent	**H-L%**	final level plateau (plateau because H- causes "upstep" of following)
H+!H*	step down		

Figure 27.10 The accent and boundary tones labels from the ToBI transcription system for American English intonation (Beckman and Ayers 1997, Beckman and Hirschberg 1994).

boundary tone

An utterance in ToBI consists of a sequence of intonational phrases, each of which ends in one of the four **boundary tones**. The boundary tones represent the utterance final aspects of tune. Each word in the utterances can optionally be associated with one of the five types of pitch accents.

Each intonational phrase consists of one or more **intermediate phrase**. These phrases can also be marked with kinds of boundary tone, including the **%H** high initial boundary tone, which marks a phrase that is particularly high in the speaker's pitch range, as well as final phrase accents **H-** and **L-**.

Break index

In addition to accents and boundary tones, ToBI distinguishes four levels of phrasing, labeled on a separate **break index** tier. The largest phrasal breaks are the intonational phrase (break index **4**) and the intermediate phrase (break index **3**), discussed above. Break index **2** is used to mark a disjuncture or pause between words that is smaller than an intermediate phrase, and **1** is used for normal phrase-medial word boundaries.

Tier

Figure 27.11 shows the tone, orthographic, and phrasing **tiers** of a ToBI transcription, using the Praat program. The same sentence is read with two different tunes. In (a), the word *Marianna* is spoken with a high H* accent, and the sentence has the declarative boundary tone L-L%. In (b), the word *Marianna* is spoken with a

low L* accent and the yes-no question boundary tone H-H%. One goal of ToBI is to express different meanings to the different type of accents. Here, the L* accent adds a meaning of *surprise* to the sentence (i.e., with a connotation like 'Are you really saying it was Marianna?') (Hirschberg and Pierrehumbert 1986, Steedman 2007).

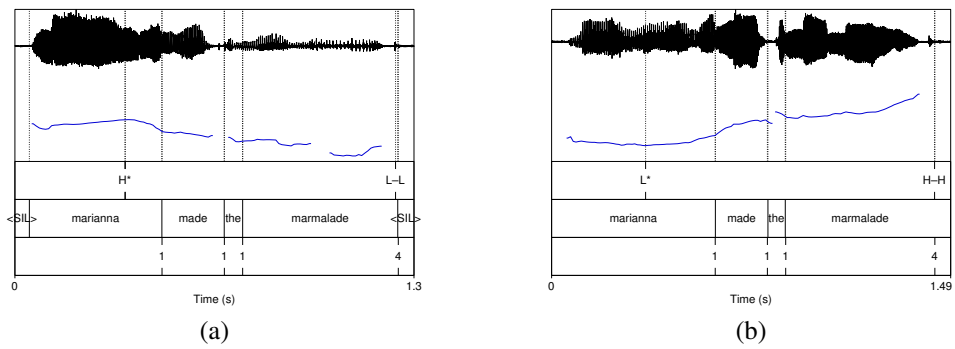


Figure 27.11 The same sentence read by Mary Beckman with two different intonation patterns and transcribed in ToBI. (a) Shows an H* accent and the typical American English declarative final fall L-L%. (b) Shows the L* accent, with the typical American English yes-no question rise H-H%.

ToBI models have been proposed for many languages (Jun, 2005), such as the J_TOBI system for Japanese (Venditti, 2005).

27.5 Acoustic Phonetics and Signals

We begin with a brief introduction to the acoustic waveform and how it is digitized and summarize the idea of frequency analysis and spectra. This is an extremely brief overview; the interested reader is encouraged to consult the references at the end of the chapter.

27.5.1 Waves

Acoustic analysis is based on the sine and cosine functions. Figure 27.12 shows a plot of a sine wave, in particular the function

$$y = A * sin(2\pi f t) \tag{27.4}$$

where we have set the amplitude A to 1 and the frequency f to 10 cycles per second.

Recall from basic mathematics that two important characteristics of a wave are its **frequency** and **amplitude**. The frequency is the number of times a second that a wave repeats itself, that is, the number of **cycles**. We usually measure frequency in **cycles per second**. The signal in Fig. 27.12 repeats itself 5 times in .5 seconds, hence 10 cycles per second. Cycles per second are usually called **hertz** (shortened to **Hz**), so the frequency in Fig. 27.12 would be described as 10 Hz. The **amplitude** A of a sine wave is the maximum value on the Y axis.

frequency
amplitude
Hertz

The **period** T of the wave is defined as the time it takes for one cycle to complete, defined as

period

$$T = \frac{1}{f} \tag{27.5}$$

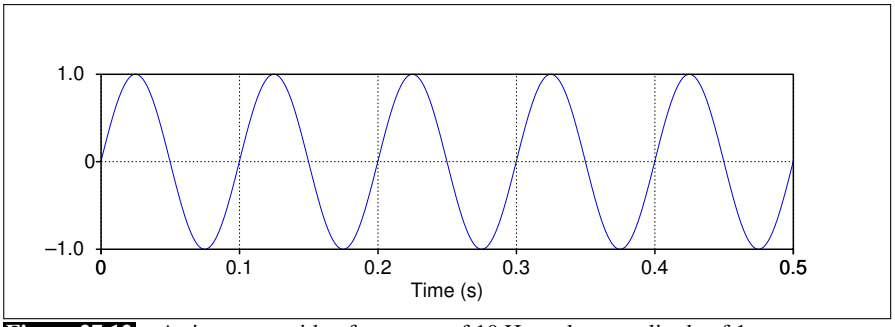


Figure 27.12 A sine wave with a frequency of 10 Hz and an amplitude of 1.

In Fig. 27.12 we can see that each cycle lasts a tenth of a second; hence $T = .1$ seconds.

27.5.2 Speech Sound Waves

Let's turn from hypothetical waves to sound waves. The input to a speech recognizer, like the input to the human ear, is a complex series of changes in air pressure. These changes in air pressure obviously originate with the speaker and are caused by the specific way that air passes through the glottis and out the oral or nasal cavities. We represent sound waves by plotting the change in air pressure over time. One metaphor which sometimes helps in understanding these graphs is that of a vertical plate blocking the air pressure waves (perhaps in a microphone in front of a speaker's mouth, or the eardrum in a hearer's ear). The graph measures the amount of **compression** or **rarefaction** (uncompression) of the air molecules at this plate. Figure 27.13 shows a short segment of a waveform taken from the Switchboard corpus of telephone speech of the vowel [iy] from someone saying "she just had a baby".

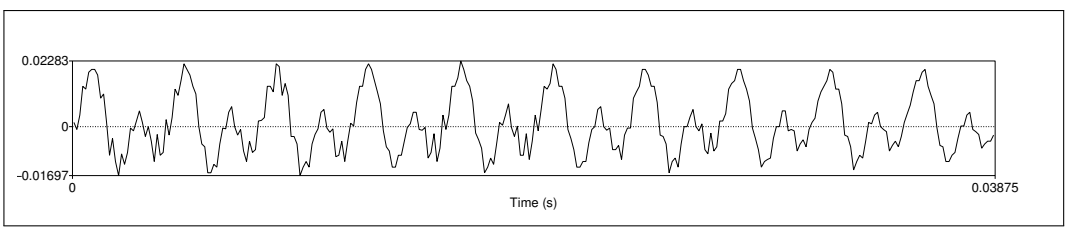


Figure 27.13 A waveform of the vowel [iy] from an utterance shown later in Fig. 27.17 on page 534. The y-axis shows the level of air pressure above and below normal atmospheric pressure. The x-axis shows time. Notice that the wave repeats regularly.

Let's explore how the digital representation of the sound wave shown in Fig. 27.13 would be constructed. The first step in processing speech is to convert the analog representations (first air pressure and then analog electric signals in a microphone) into a digital signal. This process of **analog-to-digital conversion** has two steps: **sampling** and **quantization**. To sample a signal, we measure its amplitude at a particular time; the **sampling rate** is the number of samples taken per second. To accurately measure a wave, we must have at least two samples in each cycle: one measuring the positive part of the wave and one measuring the negative part. More than two samples per cycle increases the amplitude accuracy, but fewer than two samples causes the frequency of the wave to be completely missed. Thus, the maxi-

sampling

mum frequency wave that can be measured is one whose frequency is half the sample rate (since every cycle needs two samples). This maximum frequency for a given sampling rate is called the **Nyquist frequency**. Most information in human speech is in frequencies below 10,000 Hz; thus, a 20,000 Hz sampling rate would be necessary for complete accuracy. But telephone speech is filtered by the switching network, and only frequencies less than 4,000 Hz are transmitted by telephones. Thus, an 8,000 Hz sampling rate is sufficient for **telephone-bandwidth** speech like the Switchboard corpus. A 16,000 Hz sampling rate (sometimes called **wideband**) is often used for microphone speech.

Nyquist frequency

Even an 8,000 Hz sampling rate requires 8000 amplitude measurements for each second of speech, so it is important to store amplitude measurements efficiently. They are usually stored as integers, either 8 bit (values from -128–127) or 16 bit (values from -32768–32767). This process of representing real-valued numbers as integers is called **quantization** because the difference between two integers acts as a minimum granularity (a quantum size) and all values that are closer together than this quantum size are represented identically.

quantization

Once data is quantized, it is stored in various formats. One parameter of these formats is the sample rate and sample size discussed above; telephone speech is often sampled at 8 kHz and stored as 8-bit samples, and microphone data is often sampled at 16 kHz and stored as 16-bit samples. Another parameter of these formats is the number of **channels**. For stereo data or for two-party conversations, we can store both channels in the same file or we can store them in separate files. A final parameter is individual sample storage—linearly or compressed. One common compression format used for telephone speech is μ-law (often written u-law but still pronounced mu-law). The intuition of log compression algorithms like μ-law is that human hearing is more sensitive at small intensities than large ones; the log represents small values with more faithfulness at the expense of more error on large values. The linear (unlogged) values are generally referred to as **linear PCM** values (PCM stands for pulse code modulation, but never mind that). Here's the equation for compressing a linear PCM sample value x to 8-bit μ-law, (where μ=255 for 8 bits):

channel

PCM

$$F(x) = \frac{sgn(s)\log(1+\mu|s|)}{\log(1+\mu)} \tag{27.6}$$

There are a number of standard file formats for storing the resulting digitized wavefile, such as Microsoft's .wav, Apple's AIFF and Sun's AU, all of which have special headers; simple headerless "raw" files are also used. For example, the .wav format is a subset of Microsoft's RIFF format for multimedia files; RIFF is a general format that can represent a series of nested chunks of data and control information. Figure 27.14 shows a simple .wav file with a single data chunk together with its format chunk.

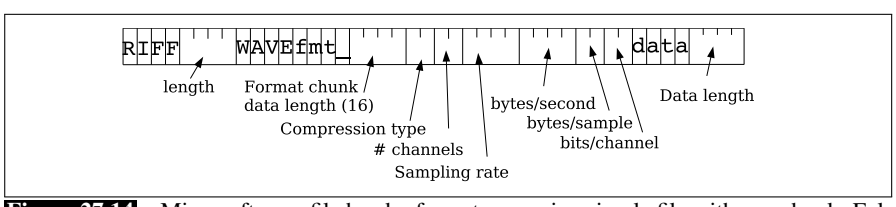


Figure 27.14 Microsoft wavefile header format, assuming simple file with one chunk. Following this 44-byte header would be the data chunk.

27.5.3 Frequency and Amplitude; Pitch and Loudness

Sound waves, like all waves, can be described in terms of frequency, amplitude, and the other characteristics that we introduced earlier for pure sine waves. In sound waves, these are not quite as simple to measure as they were for sine waves. Let's consider frequency. Note in Fig. 27.13 that although not exactly a sine, the wave is nonetheless periodic, repeating 10 times in the 38.75 milliseconds (.03875 seconds) captured in the figure. Thus, the frequency of this segment of the wave is 10/.03875 or 258 Hz.

Where does this periodic 258 Hz wave come from? It comes from the speed of vibration of the vocal folds; since the waveform in Fig. 27.13 is from the vowel [iy], it is voiced. Recall that voicing is caused by regular openings and closing of the vocal folds. When the vocal folds are open, air is pushing up through the lungs, creating a region of high pressure. When the folds are closed, there is no pressure from the lungs. Thus, when the vocal folds are vibrating, we expect to see regular peaks in amplitude of the kind we see in Fig. 27.13, each major peak corresponding to an opening of the vocal folds. The frequency of the vocal fold vibration, or the frequency of the complex wave, is called the **fundamental frequency** of the waveform, often abbreviated **F0**. We can plot F0 over time in a **pitch track**. Figure 27.15 shows the pitch track of a short question, "Three o'clock?" represented below the waveform. Note the rise in F0 at the end of the question.

fundamental frequency
F0
pitch track

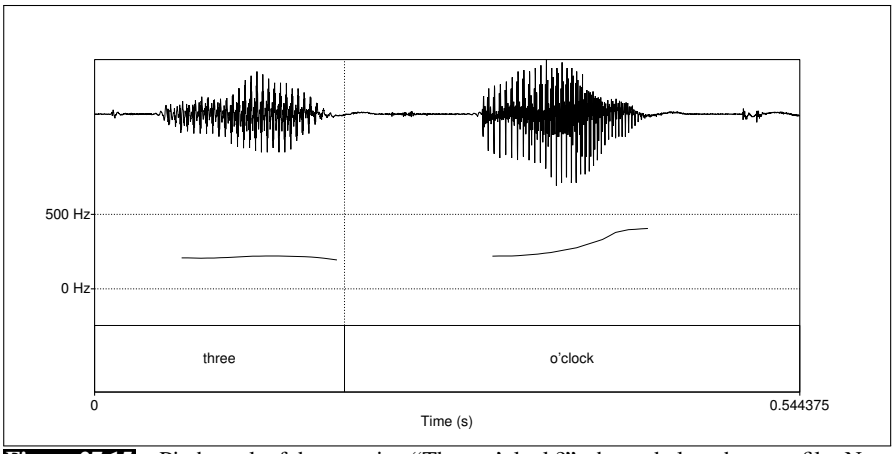


Figure 27.15 Pitch track of the question "Three o'clock?", shown below the wavefile. Note the rise in F0 at the end of the question. Note the lack of pitch trace during the very quiet part (the "o'" of "o'clock"; automatic pitch tracking is based on counting the pulses in the voiced regions, and doesn't work if there is no voicing (or insufficient sound).

The vertical axis in Fig. 27.13 measures the amount of air pressure variation; pressure is force per unit area, measured in Pascals (Pa). A high value on the vertical axis (a high amplitude) indicates that there is more air pressure at that point in time, a zero value means there is normal (atmospheric) air pressure, and a negative value means there is lower than normal air pressure (rarefaction).

In addition to this value of the amplitude at any point in time, we also often need to know the average amplitude over some time range, to give us some idea of how great the average displacement of air pressure is. But we can't just take the average of the amplitude values over a range; the positive and negative values would (mostly) cancel out, leaving us with a number close to zero. Instead, we generally use the RMS (root-mean-square) amplitude, which squares each number

before averaging (making it positive), and then takes the square root at the end.

$$\text{RMS amplitude}_{i=1}^{N} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{27.7}$$

Power The **power** of the signal is related to the square of the amplitude. If the number of samples of a sound is N, the power is

$$\text{Power} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{27.8}$$

Intensity Rather than power, we more often refer to the **intensity** of the sound, which normalizes the power to the human auditory threshold and is measured in dB. If P_0 is the auditory threshold pressure = 2×10^{-5} Pa, then intensity is defined as follows:

$$\text{Intensity} = 10\log_{10}\frac{1}{NP_0}\sum_{i=1}^{N} x_i^2 \tag{27.9}$$

Figure 27.16 shows an intensity plot for the sentence "Is it a long movie?" from the CallHome corpus, again shown below the waveform plot.

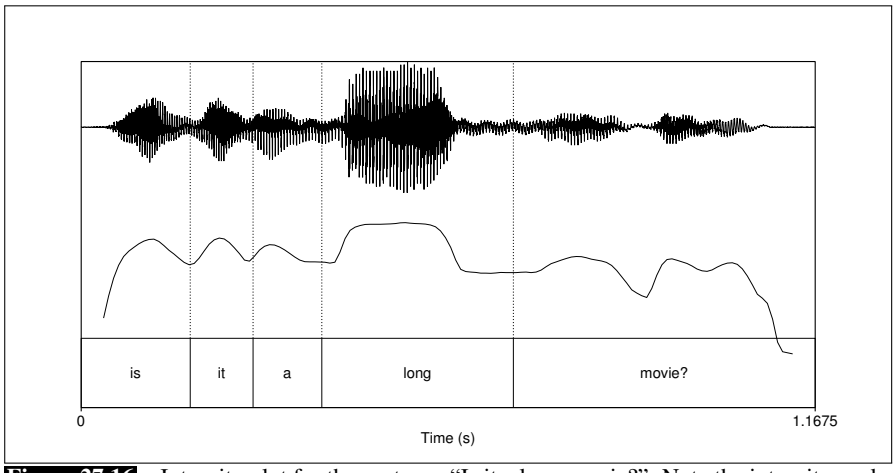


Figure 27.16 Intensity plot for the sentence "Is it a long movie?". Note the intensity peaks at each vowel and the especially high peak for the word *long*.

pitch Two important perceptual properties, **pitch** and **loudness**, are related to frequency and intensity. The **pitch** of a sound is the mental sensation, or perceptual correlate, of fundamental frequency; in general, if a sound has a higher fundamental frequency we perceive it as having a higher pitch. We say "in general" because the relationship is not linear, since human hearing has different acuities for different frequencies. Roughly speaking, human pitch perception is most accurate between 100 Hz and 1000 Hz and in this range pitch correlates linearly with frequency. Human hearing represents frequencies above 1000 Hz less accurately, and above this range, pitch correlates logarithmically with frequency. Logarithmic representation means that the differences between high frequencies are compressed and hence not as accurately perceived. There are various psychoacoustic models of pitch perception scales. One common model is the **mel** scale (Stevens et al. 1937, Stevens and

Mel

Volkmann 1940). A mel is a unit of pitch defined such that pairs of sounds which are perceptually equidistant in pitch are separated by an equal number of mels. The mel frequency m can be computed from the raw acoustic frequency as follows:

$$m = 1127 \ln(1 + \frac{f}{700}) \tag{27.10}$$

As we'll see in Chapter 28, the mel scale plays an important role in speech recognition.

The **loudness** of a sound is the perceptual correlate of the **power**. So sounds with higher amplitudes are perceived as louder, but again the relationship is not linear. First of all, as we mentioned above when we defined μ-law compression, humans have greater resolution in the low-power range; the ear is more sensitive to small power differences. Second, it turns out that there is a complex relationship between power, frequency, and perceived loudness; sounds in certain frequency ranges are perceived as being louder than those in other frequency ranges.

pitch extraction

Various algorithms exist for automatically extracting F0. In a slight abuse of terminology, these are called **pitch extraction** algorithms. The autocorrelation method of pitch extraction, for example, correlates the signal with itself at various offsets. The offset that gives the highest correlation gives the period of the signal. Other methods for pitch extraction are based on the cepstral features we introduce in Chapter 28. There are various publicly available pitch extraction toolkits; for example, an augmented autocorrelation pitch tracker is provided with Praat (Boersma and Weenink, 2005).

27.5.4 Interpretation of Phones from a Waveform

Much can be learned from a visual inspection of a waveform. For example, vowels are pretty easy to spot. Recall that vowels are voiced; another property of vowels is that they tend to be long and are relatively loud (as we can see in the intensity plot in Fig. 27.16). Length in time manifests itself directly on the x-axis, and loudness is related to (the square of) amplitude on the y-axis. We saw in the previous section that voicing is realized by regular peaks in amplitude of the kind we saw in Fig. 27.13, each major peak corresponding to an opening of the vocal folds. Figure 27.17 shows the waveform of the short sentence "she just had a baby". We have labeled this waveform with word and phone labels. Notice that each of the six vowels in Fig. 27.17, [iy], [ax], [ae], [ax], [ey], [iy], all have regular amplitude peaks indicating voicing.

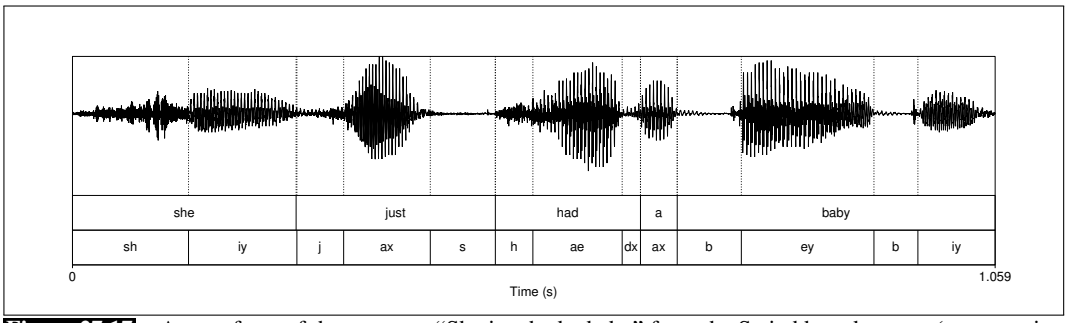


Figure 27.17 A waveform of the sentence "She just had a baby" from the Switchboard corpus (conversation 4325). The speaker is female, was 20 years old in 1991, which is approximately when the recording was made, and speaks the South Midlands dialect of American English.

For a stop consonant, which consists of a closure followed by a release, we can often see a period of silence or near silence followed by a slight burst of amplitude. We can see this for both of the [b]'s in *baby* in Fig. 27.17.

Another phone that is often quite recognizable in a waveform is a fricative. Recall that fricatives, especially very strident fricatives like [sh], are made when a narrow channel for airflow causes noisy, turbulent air. The resulting hissy sounds have a noisy, irregular waveform. This can be seen somewhat in Fig. 27.17; it's even clearer in Fig. 27.18, where we've magnified just the first word *she*.

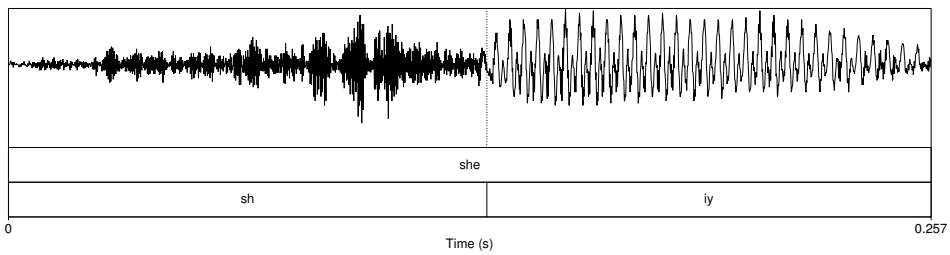


Figure 27.18 A more detailed view of the first word "she" extracted from the wavefile in Fig. 27.17. Notice the difference between the random noise of the fricative [sh] and the regular voicing of the vowel [iy].

27.5.5 Spectra and the Frequency Domain

While some broad phonetic features (such as energy, pitch, and the presence of voicing, stop closures, or fricatives) can be interpreted directly from the waveform, most computational applications such as speech recognition (as well as human auditory processing) are based on a different representation of the sound in terms of its component frequencies. The insight of **Fourier analysis** is that every complex wave can be represented as a sum of many sine waves of different frequencies. Consider the waveform in Fig. 27.19. This waveform was created (in Praat) by summing two sine waveforms, one of frequency 10 Hz and one of frequency 100 Hz.

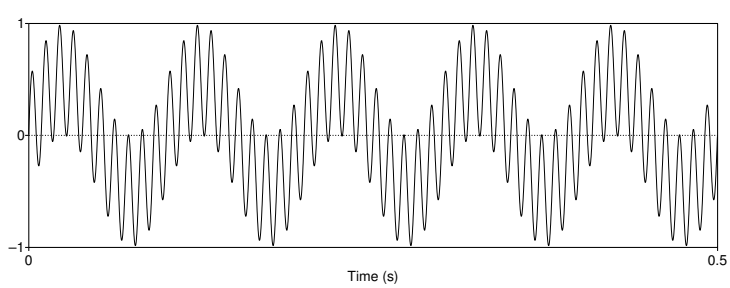


Figure 27.19 A waveform that is the sum of two sine waveforms, one of frequency 10 Hz (note five repetitions in the half-second window) and one of frequency 100 Hz, both of amplitude 1.

spectrum We can represent these two component frequencies with a **spectrum**. The spectrum of a signal is a representation of each of its frequency components and their amplitudes. Figure 27.20 shows the spectrum of Fig. 27.19. Frequency in Hz is on the x-axis and amplitude on the y-axis. Note the two spikes in the figure, one

at 10 Hz and one at 100 Hz. Thus, the spectrum is an alternative representation of the original waveform, and we use the spectrum as a tool to study the component frequencies of a sound wave at a particular time point.

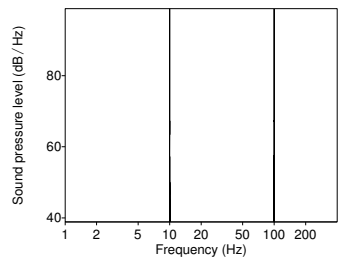


Figure 27.20 The spectrum of the waveform in Fig. 27.19.

Let's look now at the frequency components of a speech waveform. Figure 27.21 shows part of the waveform for the vowel [ae] of the word *had*, cut out from the sentence shown in Fig. 27.17.

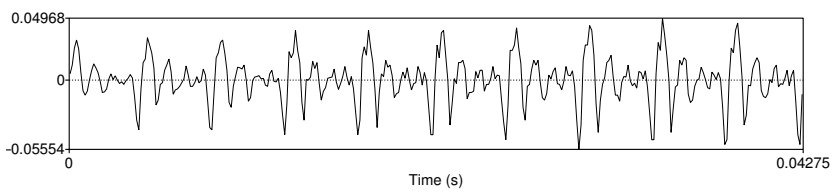


Figure 27.21 The waveform of part of the vowel [ae] from the word *had* cut out from the waveform shown in Fig. 27.17.

Note that there is a complex wave that repeats about ten times in the figure; but there is also a smaller repeated wave that repeats four times for every larger pattern (notice the four small peaks inside each repeated wave). The complex wave has a frequency of about 234 Hz (we can figure this out since it repeats roughly 10 times in .0427 seconds, and 10 cycles/.0427 seconds = 234 Hz).

The smaller wave then should have a frequency of roughly four times the frequency of the larger wave, or roughly 936 Hz. Then, if you look carefully, you can see two little waves on the peak of many of the 936 Hz waves. The frequency of this tiniest wave must be roughly twice that of the 936 Hz wave, hence 1872 Hz.

Figure 27.22 shows a smoothed spectrum for the waveform in Fig. 27.21, computed with a discrete Fourier transform (DFT).

The x-axis of a spectrum shows frequency, and the y-axis shows some measure of the magnitude of each frequency component (in decibels (dB), a logarithmic measure of amplitude that we saw earlier). Thus, Fig. 27.22 shows significant frequency components at around 930 Hz, 1860 Hz, and 3020 Hz, along with many other lower-magnitude frequency components. These first two components are just what we noticed in the time domain by looking at the wave in Fig. 27.21!

Why is a spectrum useful? It turns out that these spectral peaks that are easily visible in a spectrum are characteristic of different phones; phones have characteristic spectral "signatures". Just as chemical elements give off different wavelengths of light when they burn, allowing us to detect elements in stars by looking at the spec-

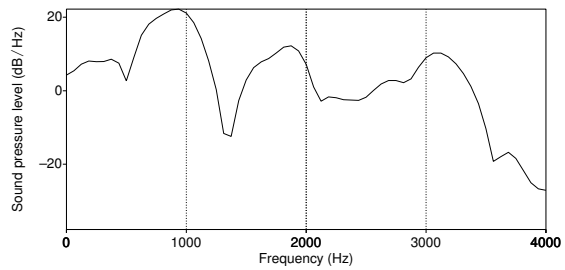


Figure 27.22 A spectrum for the vowel [ae] from the word *had* in the waveform of *She just had a baby* in Fig. 27.17.

trum of the light, we can detect the characteristic signature of the different phones by looking at the spectrum of a waveform. This use of spectral information is essential to both human and machine speech recognition. In human audition, the function of the **cochlea**, or **inner ear**, is to compute a spectrum of the incoming waveform. Similarly, the various kinds of acoustic features used in speech recognition as the HMM observation are all different representations of spectral information.

cochlea

Let's look at the spectrum of different vowels. Since some vowels change over time, we'll use a different kind of plot called a **spectrogram**. While a spectrum shows the frequency components of a wave at one point in time, a **spectrogram** is a way of envisioning how the different frequencies that make up a waveform change over time. The x-axis shows time, as it did for the waveform, but the y-axis now shows frequencies in hertz. The darkness of a point on a spectrogram corresponds to the amplitude of the frequency component. Very dark points have high amplitude, light points have low amplitude. Thus, the spectrogram is a useful way of visualizing the three dimensions (time x frequency x amplitude).

spectrogram

Figure 27.23 shows spectrograms of three American English vowels, [ih], [ae], and [ah]. Note that each vowel has a set of dark bars at various frequency bands, slightly different bands for each vowel. Each of these represents the same kind of spectral peak that we saw in Fig. 27.21.

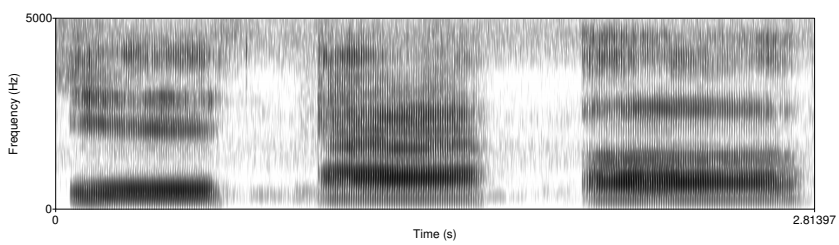


Figure 27.23 Spectrograms for three American English vowels, [ih], [ae], and [uh], spoken by the first author.

Each dark bar (or spectral peak) is called a **formant**. As we discuss below, a formant is a frequency band that is particularly amplified by the vocal tract. Since different vowels are produced with the vocal tract in different positions, they will produce different kinds of amplifications or resonances. Let's look at the first two formants, called F1 and F2. Note that F1, the dark bar closest to the bottom, is in a different position for the three vowels; it's low for [ih] (centered at about 470 Hz)

formant

and somewhat higher for [ae] and [ah] (somewhere around 800 Hz). By contrast, F2, the second dark bar from the bottom, is highest for [ih], in the middle for [ae], and lowest for [ah].

We can see the same formants in running speech, although the reduction and coarticulation processes make them somewhat harder to see. Figure 27.24 shows the spectrogram of "she just had a baby", whose waveform was shown in Fig. 27.17. F1 and F2 (and also F3) are pretty clear for the [ax] of *just*, the [ae] of *had*, and the [ey] of *baby*.

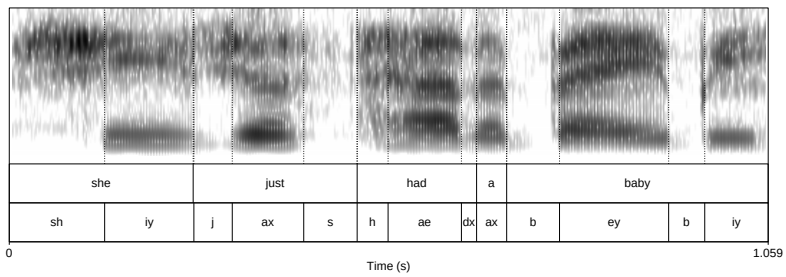


Figure 27.24 A spectrogram of the sentence "she just had a baby" whose waveform was shown in Fig. 27.17. We can think of a spectrogram as a collection of spectra (time slices), like Fig. 27.22 placed end to end.

What specific clues can spectral representations give for phone identification? First, since different vowels have their formants at characteristic places, the spectrum can distinguish vowels from each other. We've seen that [ae] in the sample waveform had formants at 930 Hz, 1860 Hz, and 3020 Hz. Consider the vowel [iy] at the beginning of the utterance in Fig. 27.17. The spectrum for this vowel is shown in Fig. 27.25. The first formant of [iy] is 540 Hz, much lower than the first formant for [ae], and the second formant (2581 Hz) is much higher than the second formant for [ae]. If you look carefully, you can see these formants as dark bars in Fig. 27.24 just around 0.5 seconds.

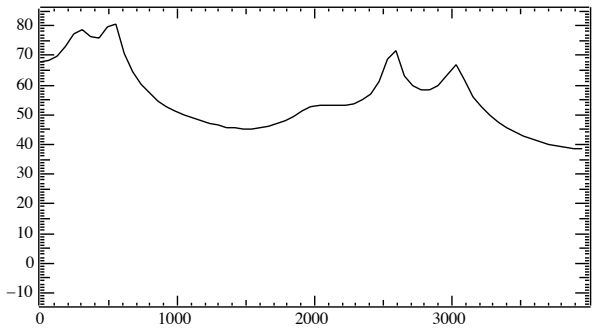


Figure 27.25 A smoothed (LPC) spectrum for the vowel [iy] at the start of *She just had a baby*. Note that the first formant (540 Hz) is much lower than the first formant for [ae] shown in Fig. 27.22, and the second formant (2581 Hz) is much higher than the second formant for [ae].

The location of the first two formants (called F1 and F2) plays a large role in determining vowel identity, although the formants still differ from speaker to speaker.

Higher formants tend to be caused more by general characteristics of a speaker's vocal tract rather than by individual vowels. Formants also can be used to identify the nasal phones [n], [m], and [ng] and the liquids [l] and [r].

27.5.6 The Source-Filter Model

Why do different vowels have different spectral signatures? As we briefly mentioned above, the formants are caused by the resonant cavities of the mouth. The **source-filter model** is a way of explaining the acoustics of a sound by modeling how the pulses produced by the glottis (the **source**) are shaped by the vocal tract (the **filter**).

source-filter model

Let's see how this works. Whenever we have a wave such as the vibration in air caused by the glottal pulse, the wave also has **harmonics**. A harmonic is another wave whose frequency is a multiple of the fundamental wave. Thus, for example, a 115 Hz glottal fold vibration leads to harmonics (other waves) of 230 Hz, 345 Hz, 460 Hz, and so on on. In general, each of these waves will be weaker, that is, will have much less amplitude than the wave at the fundamental frequency.

harmonic

It turns out, however, that the vocal tract acts as a kind of filter or amplifier; indeed any cavity, such as a tube, causes waves of certain frequencies to be amplified and others to be damped. This amplification process is caused by the shape of the cavity; a given shape will cause sounds of a certain frequency to resonate and hence be amplified. Thus, by changing the shape of the cavity, we can cause different frequencies to be amplified.

When we produce particular vowels, we are essentially changing the shape of the vocal tract cavity by placing the tongue and the other articulators in particular positions. The result is that different vowels cause different harmonics to be amplified. So a wave of the same fundamental frequency passed through different vocal tract positions will result in different harmonics being amplified.

We can see the result of this amplification by looking at the relationship between the shape of the vocal tract and the corresponding spectrum. Figure 27.26 shows the vocal tract position for three vowels and a typical resulting spectrum. The formants are places in the spectrum where the vocal tract happens to amplify particular harmonic frequencies.

27.6 Phonetic Resources

A wide variety of phonetic resources can be drawn on for computational work. One key set of resources are **pronunciation dictionaries**. Such on-line phonetic dictionaries give phonetic transcriptions for each word. Three commonly used on-line dictionaries for English are the CELEX, CMUdict, and PRONLEX lexicons; for other languages, the LDC has released pronunciation dictionaries for Egyptian Arabic, German, Japanese, Korean, Mandarin, and Spanish. All these dictionaries can be used for both speech recognition and synthesis work.

Pronunciation dictionary

The CELEX dictionary (Baayen et al., 1995) is the most richly annotated of the dictionaries. It includes all the words in the 1974 Oxford Advanced Learner's Dictionary (41,000 lemmata) and the 1978 Longman Dictionary of Contemporary English (53,000 lemmata); in total it has pronunciations for 160,595 wordforms. Its (British rather than American) pronunciations are transcribed with an ASCII version of the IPA called SAM. In addition to basic phonetic information like phone strings, syllabification, and stress level for each syllable, each word is also annotated with

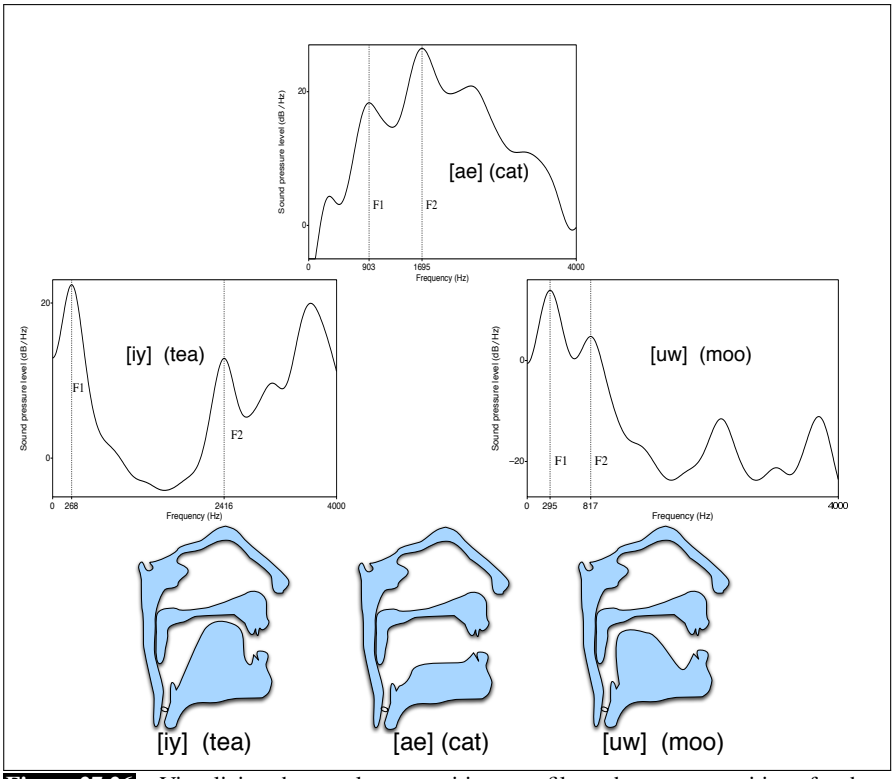


Figure 27.26 Visualizing the vocal tract position as a filter: the tongue positions for three English vowels and the resulting smoothed spectra showing F1 and F2.

morphological, part-of-speech, syntactic, and frequency information. CELEX (as well as CMU and PRONLEX) represent three levels of stress: primary stress, secondary stress, and no stress. For example, some of the CELEX information for the word *dictionary* includes multiple pronunciations ('dIk-S@n-rI and 'dIk-S@-n@-rI, corresponding to ARPAbet [d ih k sh ax n r ih] and [d ih k sh ax n ax r ih], respectively), together with the CV skelata for each one ([CVC][CVC][CV] and [CVC][CV][CV][CV]), the frequency of the word, the fact that it is a noun, and its morphological structure (diction+ary).

The free CMU Pronouncing Dictionary (CMU, 1993) has pronunciations for about 125,000 wordforms. It uses a 39-phone ARPAbet-derived phoneme set. Transcriptions are phonemic, and thus instead of marking any kind of surface reduction like flapping or reduced vowels, it marks each vowel with the number 0 (unstressed), 1 (stressed), or 2 (secondary stress). Thus, the word *tiger* is listed as [T AY1 G ER0], the word *table* as [T EY1 B AH0 L], and the word *dictionary* as [D IH1 K SH AH0 N EH2 R IY0]. The dictionary is not syllabified, although the nucleus is implicitly marked by the (numbered) vowel. Figure 27.27 shows some sample pronunciations.

ANTECEDENTS	AE2 N T IH0 S IY1 D AH0 N T S	*PAKISTANI*	P AE2 K IH0 S T AE1 N IY0
CHANG	CH AE1 NG	*TABLE*	T EY1 B AH0 L
DICTIONARY	D IH1 K SH AH0 N EH2 R IY0	*TROTSKY*	T R AA1 T S K IY2
DINNER	D IH1 N ER0	*WALTER*	W AO1 L T ER0
LUNCH	L AH1 N CH	*WALTZING*	W AO1 L T S IH0 NG
MCFARLAND	M AH0 K F AA1 R L AH0 N D	*WALTZING(2)*	W AO1 L S IH0 NG

Figure 27.27 Some sample pronunciations from the CMU Pronouncing Dictionary.

The PRONLEX dictionary (LDC, 1995) was designed for speech recognition and contains pronunciations for 90,694 wordforms. It covers all the words used in many years of the *Wall Street Journal*, as well as the Switchboard Corpus. PRONLEX has the advantage that it includes many proper names (20,000, whereas CELEX only has about 1000). Names are important for practical applications, and they are both frequent and difficult; we return to a discussion of deriving name pronunciations in Chapter 28.

The CMU dictionary was designed for speech recognition rather than synthesis uses; thus, it does not specify which of the multiple pronunciations to use for synthesis, does not mark syllable boundaries, and because it capitalizes the dictionary headwords, does not distinguish between, for example, *US* and *us* (the form *US* has the two pronunciations [AH1 S] and [Y UW1 EH1 S]).

The 110,000 word UNISYN dictionary, freely available for research purposes, resolves many of these issues as it was designed specifically for synthesis (Fitt, 2002). UNISYN gives syllabifications, stress, and some morphological boundaries. Furthermore, pronunciations in UNISYN can also be read off in any of dozens of dialects of English, including General American, RP British, Australia, and so on. The UNISYN uses a slightly different phone set; here are some examples:

```
going:       { g * ou }.> i ng >
antecedents:   { * a n . t^ i . s ~ ii . d n! t }> s >
dictionary:    { d * i k . sh @ . n ~ e . r ii }
```

Another useful resource is a **phonetically annotated corpus**, in which a collection of waveforms is hand-labeled with the corresponding string of phones. Three important phonetic corpora in English are the TIMIT corpus, the Switchboard corpus, and the Buckeye corpus.

The TIMIT corpus (NIST, 1990) was collected as a joint project between Texas Instruments (TI), MIT, and SRI. It is a corpus of 6300 read sentences, with 10 sentences each from 630 speakers. The 6300 sentences were drawn from a set of 2342 predesigned sentences, some selected to have particular dialect shibboleths, others to maximize phonetic diphone coverage. Each sentence in the corpus was phonetically hand-labeled, the sequence of phones was automatically aligned with the sentence wavefile, and then the automatic phone boundaries were manually hand-corrected (Seneff and Zue, 1988). The result is a **time-aligned transcription**: a transcription in which each phone is associated with a start and end time in the waveform. We showed a graphical example of a time-aligned transcription in Fig. 27.17 on page 534.

time-aligned transcription

The phoneset for TIMIT and for the Switchboard Transcription Project corpus below, is a more detailed one than the minimal phonemic version of the ARPAbet. In particular, these phonetic transcriptions make use of the various reduced and rare phones mentioned in Fig. 27.1 and Fig. 27.2: the flap [dx], glottal stop [q], reduced vowels [ax], [ix], [axr], voiced allophone of [h] ([hv]), and separate phones for stop closure ([dcl], [tcl], etc) and release ([d], [t], etc.). An example transcription is shown in Fig. 27.28.

she	had	your	dark	suit	in	greasy	wash	water	all	year
sh iy	hv ae dcl	jh axr	dcl d aa r kcl	s ux q	en	gcl g r iy s ix	w aa sh	q w aa dx axr q	aa l	y ix axr

Figure 27.28 Phonetic transcription from the TIMIT corpus. This transcription uses special features of ARPAbet for narrow transcription, such as the palatalization of [d] in *had*, unreleased final stop in *dark*, glottalization of final [t] in *suit* to [q], and flap of [t] in *water*. The TIMIT corpus also includes time-alignments for each phone (not shown).

Where TIMIT is based on read speech, the more recent Switchboard Transcription Project corpus is based on the Switchboard corpus of conversational speech. This phonetically annotated portion consists of approximately 3.5 hours of sentences extracted from various conversations (Greenberg et al., 1996). As with TIMIT, each annotated utterance contains a time-aligned transcription. The Switchboard transcripts are time aligned at the syllable level rather than at the phone level; thus, a transcript consists of a sequence of syllables with the start and end time of each syllables in the corresponding wavefile. Figure 27.29 shows an example from the Switchboard Transcription Project for the phrase *they're kind of in between right now*.

0.470	0.640	0.720	0.900	0.953	1.279	1.410	1.630
dh er	k aa	n ax	v ih m	b ix	t w iy n	r ay	n aw

Figure 27.29 Phonetic transcription of the Switchboard phrase *they're kind of in between right now*. Note vowel reduction in *they're* and *of*, coda deletion in *kind* and *right*, and re-syllabification (the [v] of *of* attaches as the onset of *in*). Time is given in number of seconds from the beginning of sentence to the start of each syllable.

The Buckeye corpus (Pitt et al. 2007, Pitt et al. 2005) is a phonetically transcribed corpus of spontaneous American speech, containing about 300,000 words from 40 talkers. Phonetically transcribed corpora are also available for other languages, including the Kiel corpus of German and Mandarin corpora transcribed by the Chinese Academy of Social Sciences (Li et al., 2000).

In addition to resources like dictionaries and corpora, there are many useful phonetic software tools. One of the most versatile is the Praat package (Boersma and Weenink, 2005), which includes spectrum and spectrogram analysis, pitch extraction and formant analysis, and an embedded scripting language for automation.

27.7 Summary

This chapter has introduced many of the important concepts of phonetics and computational phonetics.

- We can represent the pronunciation of words in terms of units called **phones**. The standard system for representing phones is the **International Phonetic Alphabet** or **IPA**. The most common computational system for transcription of English is the **ARPAbet**, which conveniently uses ASCII symbols.
- Phones can be described by how they are produced **articulatorily** by the vocal organs; consonants are defined in terms of their **place** and **manner** of articulation and **voicing**; vowels by their **height**, **backness**, and **roundness**.
- A **phoneme** is a generalization or abstraction over different phonetic realizations. **Allophonic rules** express how a phoneme is realized in a given context.
- Speech sounds can also be described **acoustically**. Sound waves can be described in terms of **frequency**, **amplitude**, or their perceptual correlates, **pitch** and **loudness**.
- The **spectrum** of a sound describes its different frequency components. While some phonetic properties are recognizable from the waveform, both humans and machines rely on spectral analysis for phone detection.
- A **spectrogram** is a plot of a spectrum over time. Vowels are described by characteristic harmonics called **formants**.

- **Pronunciation dictionaries** are widely available and used for both speech recognition and synthesis, including the CMU dictionary for English and CELEX dictionaries for English, German, and Dutch. Other dictionaries are available from the LDC.
- Phonetically transcribed corpora are a useful resource for building computational models of phone variation and reduction in natural speech.

Bibliographical and Historical Notes

The major insights of articulatory phonetics date to the linguists of 800–150 B.C. India. They invented the concepts of place and manner of articulation, worked out the glottal mechanism of voicing, and understood the concept of assimilation. European science did not catch up with the Indian phoneticians until over 2000 years later, in the late 19th century. The Greeks did have some rudimentary phonetic knowledge; by the time of Plato's *Theaetetus* and *Cratylus*, for example, they distinguished vowels from consonants, and stop consonants from continuants. The Stoics developed the idea of the syllable and were aware of phonotactic constraints on possible words. An unknown Icelandic scholar of the 12th century exploited the concept of the phoneme and proposed a phonemic writing system for Icelandic, including diacritics for length and nasality. But his text remained unpublished until 1818 and even then was largely unknown outside Scandinavia (Robins, 1967). The modern era of phonetics is usually said to have begun with Sweet, who proposed what is essentially the phoneme in his *Handbook of Phonetics* (1877). He also devised an alphabet for transcription and distinguished between *broad* and *narrow* transcription, proposing many ideas that were eventually incorporated into the IPA. Sweet was considered the best practicing phonetician of his time; he made the first scientific recordings of languages for phonetic purposes and advanced the state of the art of articulatory description. He was also infamously difficult to get along with, a trait that is well captured in Henry Higgins, the stage character that George Bernard Shaw modeled after him. The phoneme was first named by the Polish scholar Baudouin de Courtenay, who published his theories in 1894.

Students with further interest in transcription and articulatory phonetics should consult an introductory phonetics textbook such as Ladefoged (1993) or Clark and Yallop (1995). Pullum and Ladusaw (1996) is a comprehensive guide to each of the symbols and diacritics of the IPA. A good resource for details about reduction and other phonetic processes in spoken English is Shockey (2003). Wells (1982) is the definitive three-volume source on dialects of English.

Many of the classic insights in acoustic phonetics had been developed by the late 1950s or early 1960s; just a few highlights include techniques like the sound spectrograph (Koenig et al., 1946), theoretical insights like the working out of the source-filter theory and other issues in the mapping between articulation and acoustics ((Fant, 1960), Stevens et al. 1953, Stevens and House 1955, Heinz and Stevens 1961, Stevens and House 1961) the F1xF2 space of vowel formants (Peterson and Barney, 1952), the understanding of the phonetic nature of stress and the use of duration and intensity as cues (Fry, 1955), and a basic understanding of issues in phone perception (Miller and Nicely 1955,Liberman et al. 1952). Lehiste (1967) is a collection of classic papers on acoustic phonetics. Many of the seminal papers of Gunnar Fant have been collected in Fant (2004).

Excellent textbooks on acoustic phonetics include Johnson (2003) and Ladefoged (1996). Coleman (2005) includes an introduction to computational processing

of acoustics as well as other speech processing issues, from a linguistic perspective. Stevens (1998) lays out an influential theory of speech sound production. A wide variety of books address speech from a signal processing and electrical engineering perspective. The ones with the greatest coverage of computational phonetics issues include Huang et al. (2001), O'Shaughnessy (2000), and Gold and Morgan (1999). Excellent textbooks on digital signal processing are Lyons (2004) and Rabiner and Schafer (1978).

There are a number of software packages for acoustic phonetic analysis. Probably the most widely-used one is **Praat** (Boersma and Weenink, 2005).

Many phonetics papers of computational interest are to be found in the *Journal of the Acoustical Society of America (JASA)*, *Computer Speech and Language*, and *Speech Communication*.

Exercises

27.1 Find the mistakes in the ARPAbet transcriptions of the following words:

a. "three" [dh r i]
b. "sing" [s ih n g]
c. "eyes" [ay s]
d. "study" [s t uh d i]
e. "though" [th ow]
f. "planning" [p pl aa n ih ng]
g. "slight" [s l iy t]

27.2 Translate the pronunciations of the following color words from the IPA into the ARPAbet (and make a note if you think you pronounce them differently than this!):

a. [rɛd]
b. [blu]
c. [grin]
d. [ˈjɛloʊ]
e. [blæk]
f. [waɪt]
g. [ˈɔrɪndʒ]
h. [ˈpɝpl̩]
i. [pjus]
j. [toʊp]

27.3 Ira Gershwin's lyric for *Let's Call the Whole Thing Off* talks about two pronunciations (each) of the words "tomato", "potato", and "either". Transcribe into the ARPAbet both pronunciations of each of these three words.

27.4 Transcribe the following words in the ARPAbet:

1. dark
2. suit
3. greasy
4. wash
5. water

27.5 Take a wavefile of your choice. Some examples are on the textbook website. Download the Praat software, and use it to transcribe the wavefiles at the word level and into ARPAbet phones, using Praat to help you play pieces of each wavefile and to look at the wavefile and the spectrogram.

27.6 Record yourself saying five of the English vowels: [aa], [eh], [ae], [iy], [uw]. Find F1 and F2 for each of your vowels.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database (Release 2) [CD-ROM]*. Linguistic Data Consortium, University of Pennsylvania [Distributor].

Beckman, M. E. and Ayers, G. M. (1997). Guidelines for ToBI labelling. Unpublished manuscript, Ohio State University, http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/.

Beckman, M. E. and Hirschberg, J. (1994). The ToBI annotation conventions. Manuscript, Ohio State University.

Boersma, P. and Weenink, D. (2005). Praat: doing phonetics by computer (version 4.3.14). [Computer program]. Retrieved May 26, 2005, from http://www.praat.org/.

Bolinger, D. (1981). Two kinds of vowels, two kinds of rhythm. Indiana University Linguistics Club.

Clark, J. and Yallop, C. (1995). *An Introduction to Phonetics and Phonology* (2nd Ed.). Blackwell.

CMU (1993). The Carnegie Mellon Pronouncing Dictionary v0.1. Carnegie Mellon University.

Coleman, J. (2005). *Introducing Speech and Language Processing*. Cambridge University Press.

Fant, G. M. (1960). *Acoustic Theory of Speech Production*. Mouton.

Fant, G. M. (2004). *Speech Acoustics and Phonetics*. Kluwer.

Fitt, S. (2002). Unisyn lexicon. http://www.cstr.ed.ac.uk/projects/unisyn/.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *JASA*, *27*, 765–768.

Gold, B. and Morgan, N. (1999). *Speech and Audio Signal Processing*. Wiley Press.

Greenberg, S., Ellis, D., and Hollenback, J. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *ICSLP-96*, S24–27.

Heinz, J. M. and Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *JASA*, *33*, 589–596.

Hirschberg, J. and Pierrehumbert, J. B. (1986). The intonational structuring of discourse. In *ACL-86*, 136–144.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall.

Johnson, K. (2003). *Acoustic and Auditory Phonetics* (2nd Ed.). Blackwell.

Jun, S.-A. (Ed.). (2005). *Prosodic Typology and Transcription: A Unified Approach*. Oxford University Press.

Koehn, P., Abney, S. P., Hirschberg, J., and Collins, M. (2000). Improving intonational phrasing with syntactic information. In *ICASSP-00*, 1289–1290.

Koenig, W., Dunn, H. K., and Lacy, L. Y. (1946). The sound spectrograph. *JASA*, *18*, 19–49.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge Studies in Linguistics. Cambridge University Press.

Ladefoged, P. (1993). *A Course in Phonetics*. Harcourt Brace Jovanovich. (3rd ed.).

Ladefoged, P. (1996). *Elements of Acoustic Phonetics* (2nd Ed.). University of Chicago.

LDC (1995). COMLEX English Pronunciation Dictionary Version 0.2 (COMLEX 0.2). Linguistic Data Consortium.

Lehiste, I. (Ed.). (1967). *Readings in Acoustic Phonetics*. MIT Press.

Li, A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Yi, L., Song, Z., Ruhi, U., Venkataramani, V., and Chen, X. (2000). CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. In *ICSLP-00*, 485–488.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1952). The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, *65*, 497–516.

Lyons, R. G. (2004). *Understanding Digital Signal Processing*. Prentice Hall. (2nd. ed).

Miller, C. A. (1998). Pronunciation modeling in speech synthesis. Tech. rep. IRCS 98–09, University of Pennsylvania Institute for Research in Cognitive Science, Philadephia, PA.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *JASA*, *27*, 338–352.

NIST (1990). TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology Speech Disc 1-1.1. NIST Order No. PB91-505065.

O'Shaughnessy, D. (2000). *Speech Communications: Human and Machine*. IEEE Press, New York. 2nd. ed.

Ostendorf, M. and Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, *20*(1), 27–54.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *JASA*, *24*, 175–184.

Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, MIT.

Pitrelli, J. F., Beckman, M. E., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *ICSLP-94*, Vol. 1, 123–126.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W. D., Hume, E., and Fosler-Lussier, E. (2007). Buckeye corpus of conversational speech (2nd release).. Department of Psychology, Ohio State University (Distributor).

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. D. (2005). The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*, 90–95.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). The use of prosody in syntactic disambiguation. *JASA*, *90*(6).

Pullum, G. K. and Ladusaw, W. A. (1996). *Phonetic Symbol Guide* (2nd Ed.). University of Chicago.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice Hall.

Robins, R. H. (1967). *A Short History of Linguistics*. Indiana University Press, Bloomington.

Seneff, S. and Zue, V. W. (1988). Transcription and alignment of the TIMIT database. In *Proceedings of the Second Symposium on Advanced Man-Machine Interface through Spoken Language*.

Shockey, L. (2003). *Sound Patterns of Spoken English*. Blackwell.

Shoup, J. E. (1980). Phonological aspects of speech recognition. In Lea, W. A. (Ed.), *Trends in Speech Recognition*, 125–138. Prentice Hall.

Silverman, K., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P. J., Pierrehumbert, J. B., and Hirschberg, J. (1992). ToBI: A standard for labelling English prosody. In *ICSLP-92*, Vol. 2, 867–870.

Steedman, M. (2007). Information-structural semantics for English intonation. In Lee, C., Gordon, M., and Büring, D. (Eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, 245–264. Springer.

Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press.

Stevens, K. N. and House, A. S. (1955). Development of a quantitative description of vowel articulation. *JASA*, *27*, 484–493.

Stevens, K. N. and House, A. S. (1961). An acoustical theory of vowel production and some of its implications. *Journal of Speech and Hearing Research*, *4*, 303–320.

Stevens, K. N., Kasowski, S., and Fant, G. M. (1953). An electrical analog of the vocal tract. *JASA*, *25*(4), 734–742.

Stevens, S. S. and Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*(3), 329–353.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *JASA*, *8*, 185–190.

Sweet, H. (1877). *A Handbook of Phonetics*. Clarendon Press.

Venditti, J. J. (2005). The j_tobi model of japanese intonation. In Jun, S.-A. (Ed.), *Prosodic Typology and Transcription: A Unified Approach*. Oxford University Press.

Wells, J. C. (1982). *Accents of English*. Cambridge University Press.