7 PHONETICS

(Upon being asked by Director George Cukor to teach Rex Harrison, the star of the 1964 film "My Fair Lady", how to behave like a phonetician:)

"My immediate answer was, 'I don't have a singing butler and three maids who sing, but I will tell you what I can as an assistant professor.'"

Peter Ladefoged, quoted in his obituary, LA Times, 2004

The debate between the "whole language" and "phonics" methods of teaching reading to children seems at very glance like a purely modern educational debate. Like many modern debates, however, this one recapitulates an important historical dialectic, in this case in writing systems. The earliest independently-invented writing systems (Sumerian, Chinese, Mayan) were mainly logographic: one symbol represented a whole word. But from the earliest stages we can find, most such systems contain elements of syllabic or phonemic writing systems, in which symbols are used to represent the sounds that make up the words. Thus the Sumerian symbol pronounced *ba* and meaning "ration" could also function purely as the sound /ba/. Even modern Chinese, which remains primarily logographic, uses sound-based characters to spell out foreign words. 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 which were used to represent 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, is the Ur-theory that underlies all our modern theories of **phonology**. This idea of decomposing speech and words into smaller units also underlies the modern algorithms for **speech recognition** (transcrbining acoustic waveforms into strings of text words) and **speech synthesis** or **text-to-speech** (converting strings of text words into acoustic waveforms).

In this chapter we introduce **phonetics** from a computational perspective. Phonetics is the study of linguistic sounds, 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.

We begin with a key element of both speech recognition and text-to-speech systems: how words are pronounced in terms of individual speech units called **phones**. A speech recognition system needs to have a pronunciation for every word it can recognize, and a text-to-speech system needs to have a pronunciation for every word it can say. The first section of this chapter will introduce **phonetic alphabets** for describing these pronunciations. We then introduce the two main areas of phonetics, **articulatory phonetics**, the study of how speech sounds are produced by articulators in the mouth, and **acoustic phonetics**, the study of the acoustic analysis of speech sounds.

We also briefly touch on **phonology**, the area of linguistics that describes the systematic way that sounds are differently realized in different environments, and how this system of sounds is related to the rest of the grammar. In doing so we focus on the crucial fact of **variation** in modeling speech; phones are pronounced differently in different contexts.

7.1 Speech Sounds and Phonetic Transcription

PHONETICS

The study of the pronunciation of words is part of the field of **phonetics**, the study of the speech sounds used in the languages of the world. We model the pronunciation of a

PHONES

word as a string of symbols which represent **phones** or **segments**. A phone is a speech sound; phones are represented with phonetic symbols that bear some resemblance to a letter in an alphabetic language like English.

This section surveys the different phones of English, particularly American English, showing how they are produced and how they are represented symbolically. We will be using two different alphabets for describing phones. The **International Phonetic Alphabet** (**IPA**)

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 IPA is not just an alphabet but also a set of principles for transcription, which differ according to the needs of the transcription, so the same utterance can be transcribed in different ways all according to the principles of the IPA. The ARPAbet (Shoup, 1980) is another phonetic alphabet, but one that is specifically designed for American English and which uses ASCII symbols; it can be thought of as a convenient ASCII representation of an American-English subset of the IPA. ARPAbet symbols are often used in applications where non-ASCII fonts are inconvenient, such as in on-line pronunciation dictionaries. Because the ARPAbet is very common for computational representations of pronunciations, we will rely on it rather than the IPA in the remainder of this book. Fig. 7.1 and Fig. 7.2 show the ARPAbet symbols for transcribing consonants and vowels, respectively, together with their IPA equivalents.

[1] The phone [ux] is rare in general American English and not generally used in speech systems. It is used to represent the fronted [uw] which appeared in (at least) Western and Northern Cities dialects of American English starting in the late 1970s (Labov, 1994). This fronting was first called to public by imitations

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]
Less commonly used phones and allophones			
[q]	[ʔ]	uh-oh	[q ah q ow]
[dx]	[ɾ]	butter	[b ah dx axr]
[nx]	[ɾ̃]	winner	[w ih nx axr]
[el]	[l̩]	table	[t ey b el]

Figure 7.1 ARPAbet symbols for transcription of English consonants, with IPA equivalents. Note that some rarer symbols like the flap [dx], nasal flap [nx], glottal stop [q] and the syllabic consonants, are used mainly for narrow transcriptions.

Many of the IPA and ARPAbet symbols are equivalent to the Roman letters used in the orthography of English and many other languages. So for example the ARPAbet phone [p] represents the consonant sound at the beginning of *platypus*, *puma*, and *pachyderm*, 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

and recordings of 'Valley Girls' speech by Moon Zappa (Zappa and Zappa, 1982). Nevertheless, for most speakers [uw] is still much more common than [ux] in words like *dude*.

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]
Reduced and uncommon phones			
[ax]	[ə]	lotus	[l ow dx ax s]
[axr]	[ɚ]	heather	[h eh dh axr]
[ix]	[ɨ]	tulip	[t uw l ix p]
[ux]	[ʉ]	dude[1]	[d ux d]

Figure 7.2 ARPAbet symbols for transcription of English vowels, with IPA equivalents. Note again the list of rarer phones and reduced vowels (see Sec. 7.2.4); for example [ax] is the reduced vowel schwa, [ix] is the reduced vowel corresponding to [ih], and [axr] is the reduced vowel corresponding to [er].

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.

7.2 Articulatory Phonetics

ARTICULATORY PHONETICS

The list of ARPAbet phones is useless without an understanding of how each phone is produced. We thus turn to **articulatory phonetics**, the study of how phones are produced, as the various organs in the mouth, throat, and nose modify the airflow from the lungs.

7.2.1 The Vocal Organs

Sound is produced by the rapid movement of air. Most sounds in human spoken languages are produced 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,

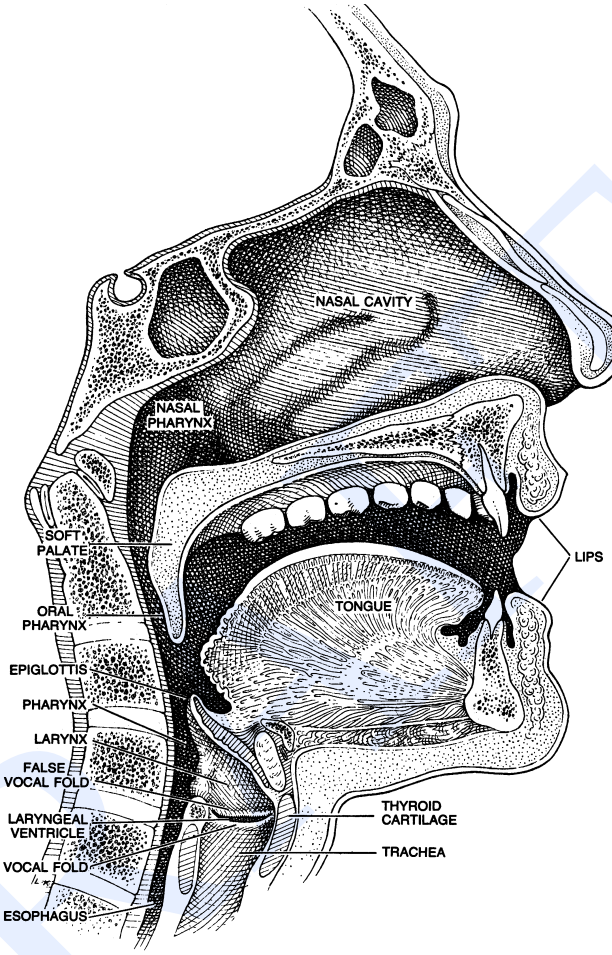

Figure 7.3 The vocal organs, shown in side view. Drawing by Laszlo Kubinyi from Sundberg (1977), ©Scientific American, used by permission.

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
UNVOICED
VOICELESS

The area above the trachea is called the **vocal tract**, and consists of the **oral tract** and the **nasal tract**. After the air leaves the trachea, it can exit the body through the

NASAL SOUNDS

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*, and *n*, and *ng*.

CONSONANTS
VOWELS

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 restricting or blocking the airflow in some way, and may 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 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.

7.2.2 Consonants: Place of Articulation

PLACE

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 **place of articulation** of a consonant. Places of articulation, shown in Fig. 7.4, are often used in automatic speech recognition as a useful way of grouping phones together into equivalence classes:

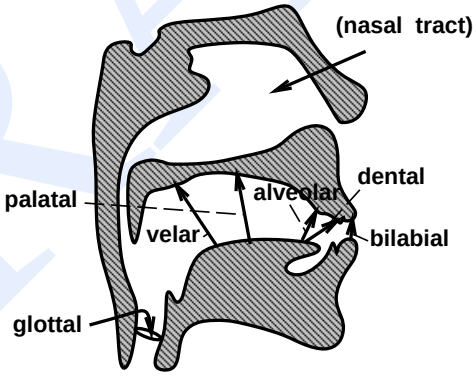

Figure 7.4 Major English places of articulation.

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** 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* or 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 **coronal** is often used to refer to both dental and alveolar.

PALATAL **palatal:** The roof of the mouth (the **palate**) rises sharply from the back of the alveolar PALATE ridge. The **palato-alveolar** sounds [sh] (*shrimp*), [ch] (*china*), [zh] (*Asian*), and [jh] (*jar*) are made with the blade of the tongue against this 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 VELUM 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).

7.2.3 Consonants: Manner of Articulation

Consonants are also distinguished by *how* the restriction in airflow is made, for example whether there is a complete stoppage of air, or only a partial blockage, etc. This feature MANNER 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:

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**. Some computational systems use a more narrow (detailed) transcription style that has separate labels for the closure and release parts of a stop. In one version of the ARPAbet, for example, the closure of a [p], [t], or [k] is represented as [pcl], [tcl], or [kcl] (respectively), while the symbols [p], [t], and [k] are used to mean only the release portion of the stop. In another version the symbols [pd], [td], [kd], [bd], [dd], [gd] are used to mean unreleased stops (stops at the end of words or phrases often are missing the explosive release), while [p], [t], [k], etc are used to mean normal stops with a closure and a release. The IPA uses a special symbol to mark unreleased stops: [p̚], [t̚], or [k̚]. We will not be using these narrow transcription styles in this chapter; we will always use [p] to mean a full stop with both a closure and a release.

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 frica-

tives [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

APPROXIMANTS

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

FLAP

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 UK dialects would use a [t] instead of a tap in this word.

7.2.4 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**, which indicates whether this high point is toward the front or back of the oral tract, and the shape of the lips (**rounded** or not). Fig. 7.5 shows the position of the tongue for different vowels.

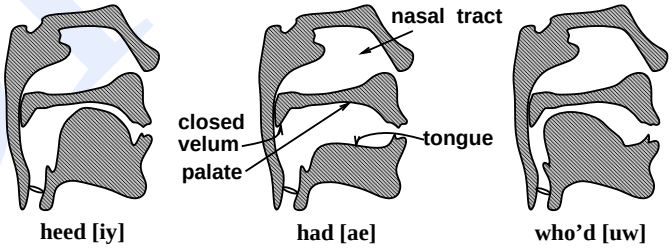

Figure 7.5 Positions of the tongue for three English vowels, high front [iy], low front [ae] and high back [uw]; tongue positions modeled after Ladefoged (1996).

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

FRONT

BACK
HIGH

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.

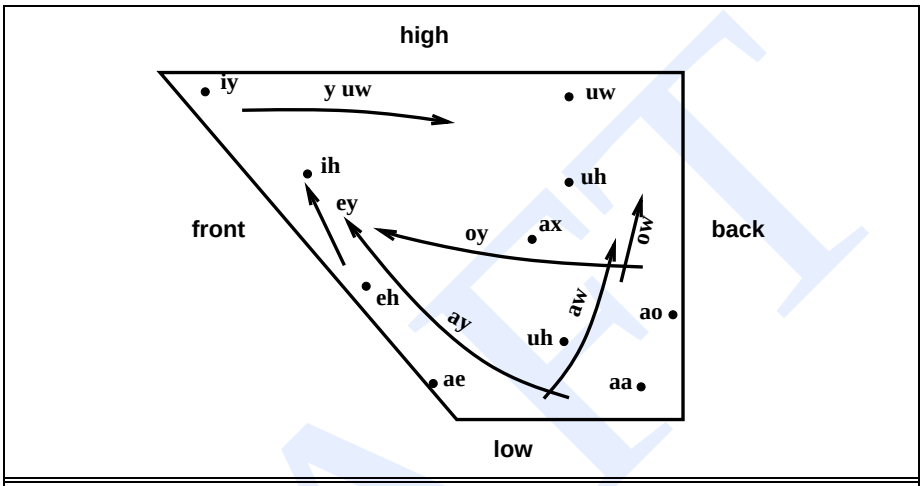

Figure 7.6 Qualities of English vowels (after Ladefoged (1993)).

Fig. 7.6 shows a schematic characterization of the vowel height of different vowels. It is schematic because the abstract property **height** only correlates roughly with actual 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 vector. A vowel in which the tongue
DIPHTHONG
position changes markedly during the production of the vowel is a **diphthong**. English is particularly rich in diphthongs.

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
ROUNDED
whistling). These **rounded** vowels include [uw], [ao], and [ow].

Syllables

SYLLABLE
Consonants and vowels combine to make a **syllable**. There is no completely agreed-upon definition of a syllable; roughly speaking 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], while the word *catnip* has two
NUCLEUS
syllables, [k ae t] and [n ih p], We call the vowel at the core of a syllable the **nucleus**.
ONSET
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
CODA
RIME
nucleus. Thus [d] is the onset of *dog*, while [g] is the coda. The **rime**. or **rhyme**. is the
RHYME
nucleus plus coda. Fig. 7.7 shows some sample syllable structures.

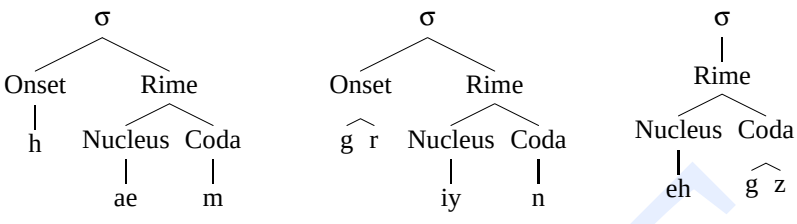

Figure 7.7 Syllable structure of *ham*, *green*, *eggs*. σ=syllable.

The task of automatically breaking up a word into syllables is called **syllabification**, and will be discussed in Sec. **??**.

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 listing constraints on fillers of syllable positions, or by creating a finite-state model of possible phone sequences. It is also possible to create a probabilistic phonotactics, by training *N*-gram grammars on phone sequences.

PHONOTACTICS

Lexical Stress and Schwa

In a natural sentence of American English, certain syllables are more **prominent** than others. These are called **accented** syllables, and the linguistic marker associated with this prominence is called a **pitch accent**. Words or syllables which are prominent are said to **bear** (be associated with) a pitch accent. Pitch accent is also sometimes referred to as **sentence stress**, although sentence stress can instead refer to only the most prominent accent in a sentence.

ACCENTED

PITCH ACCENT

BEAR

Accented syllables may be prominent by being louder, longer, by being associated with a pitch movement, or by any combination of the above. Since accent plays important roles in meaning, understanding exactly why a speaker chooses to accent a particular syllable is very complex, and we will return to this in detail in Sec. **??**. But one important factor in accent is often represented in pronunciation dictionaries. This factor is called **lexical stress**. The syllable that has lexical stress is the one that will be louder or longer if the word is accented. For example the word *parsley* is stressed in its first syllable, not its second. Thus if the word *parsley* receives a pitch accent in a sentence, it is the first syllable that will be stronger.

LEXICAL STRESS

In IPA we write the symbol [ˈ] before a syllable to indicate that it has lexical stress (e.g. [ˈpar.sli]). This difference in lexical stress can affect the meaning of a word. For example the word *content* can be a noun or an adjective. When pronounced in isolation the two senses are pronounced differently since they have different stressed syllables (the noun is pronounced [ˈkɑn.tɛnt] and the adjective [kən.ˈtɛnt]).

Vowels which 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

REDUCED VOWELS

SCHWA

the shape of the mouth is somewhat neutral; the tongue is neither particularly high nor particularly low. For example 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) proposed that American English had three reduced vowels: a reduced mid vowel [ə], a reduced front vowel [ɨ], and a reduced rounded vowel [ɵ]. The full ARPAbet includes two of these, the schwa [ax] and [ix] ([ɨ]), as well as [axr] which is an r-colored schwa (often called **schwar**), although [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).

Not all unstressed vowels are reduced; any vowel, and diphthongs in particular can retain their 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 in the word *carry* [k ae r iy].

Some computational ARPAbet lexicons mark reduced vowels like schwa explicitly. But in general predicting reduction requires knowledge of things outside the lexicon (the prosodic context, rate of speech, etc, as we will see the next section). Thus other ARPAbet versions mark stress but don't mark how stress affects reduction. 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]

SECONDARY STRESS

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. We return to this discussion when we introduce prosody in more detail in Sec. **??**.

7.3 Phonological Categories and Pronunciation Variation

'Scuse me, while I kiss the sky
Jimi Hendrix, *Purple Haze*
'Scuse me, while I kiss this guy
Common mis-hearing of same lyrics

If each word was pronounced with a fixed string of phones, each of which was pronounced the same in all contexts and by all speakers, the speech recognition and speech synthesis tasks would be really easy. Alas, the realization of words and phones varies massively depending on many factors. Fig. 7.8 shows a sample of the wide variation in pronunciation in the words *because* and *about* from the hand-transcribed Switchboard corpus of American English telephone conversations (Greenberg et al., 1996).

How can we model and predict this extensive variation? One useful tool is the assumption that what is mentally represented in the speaker's mind are abstract cate-

because				about			
ARPAbet	**%**	**ARPAbet**	**%**	**ARPAbet**	**%**	**ARPAbet**	**%**
b iy k ah z	27%	k s	2%	ax b aw	32%	b ae	3%
b ix k ah z	14%	k ix z	2%	ax b aw t	16%	b aw t	3%
k ah z	7%	k ih z	2%	b aw	9%	ax b aw dx	3%
k ax z	5%	b iy k ah zh	2%	ix b aw	8%	ax b ae	3%
b ix k ax z	4%	b iy k ah s	2%	ix b aw t	5%	b aa	3%
b ih k ah z	3%	b iy k ah	2%	ix b ae	4%	b ae dx	3%
b ax k ah z	3%	b iy k aa z	2%	ax b ae dx	3%	ix b aw dx	2%
k uh z	2%	ax z	2%	b aw dx	3%	ix b aa t	2%

Figure 7.8 The 16 most common pronunciations of *because* and *about* from the hand-transcribed Switchboard corpus of American English conversational telephone speech (Godfrey et al., 1992; Greenberg et al., 1996).

gories rather than phones in all their gory phonetic detail. For example consider the different pronunciations of [t] in the words *tunafish* and *starfish*. The [t] of *tunafish* is **aspirated**. Aspiration is a period of voicelessness after a stop closure and before the onset of voicing of the following vowel. Since the vocal cords are not vibrating, aspiration sounds like a puff of air after the [t] and before the vowel. By contrast, a [t] following an initial [s] is **unaspirated**; thus the [t] in *starfish* ([s t aa r f ih sh]) has no period of voicelessness after the [t] closure. This variation in the realization of [t] is predictable: whenever a [t] begins a word or unreduced syllable in English, it is aspirated. The same variation occurs for [k]; the [k] of *sky* is often mis-heard as [g] in Jimi Hendrix's lyrics because [k] and [g] are both unaspirated.[2]

UNASPIRATED

There are other contextual variants of [t]. For example, when [t] occurs between two vowels, particularly when the first is stressed, it is often pronounced as a **tap**. Recall that a tap is a voiced sound in which the top of the tongue is curled up and back and struck quickly against the alveolar ridge. Thus the word *buttercup* is usually pronounced [b ah dx axr k uh p] rather than [b ah t axr k uh p]. Another variant of [t] occurs before the dental consonant [th]. Here the [t] becomes dentalized (IPA [t̪]). That is, instead of the tongue forming a closure against the alveolar ridge, the tongue touches the back of the teeth.

In both linguistics and in speech processing, we use abstract classes to capture the similarity among all these [t]s. The simplest abstract class is called the **phoneme**, and its different surface realizations in different contexts are called **allophones**. We traditionally write phonemes inside slashes. So in the above examples, /t/ is a phoneme whose allophones include (in IPA) [tʰ], [ɾ], and [t̪]. Fig. 7.9 summarizes a number of allophones of /t/. In speech synthesis and recognition, we use phonesets like the ARPAbet to approximate this idea of abstract phoneme units, and represent pronunciation lexicons using ARPAbet phones. For this reason, the allophones listed in Fig. 7.1 tend to be used for narrow transcriptions for analysis purposes, and less often used in speech recognition or synthesis systems.

PHONEME

ALLOPHONES

[2] The ARPAbet does not have a way of marking aspiration; in the IPA aspiration is marked as [ʰ], so in IPA the word *tunafish* would be transcribed [tʰunəfɪʃ].

IPA	ARPABet	Description	Environment	Example
t^h	[t]	aspirated	in initial position	*toucan*
t		unaspirated	after [s] or in reduced syllables	*starfish*
ʔ	[q]	glottal stop	word-finally or after vowel before [n]	*kitten*
ʔt	[qt]	glottal stop t	sometimes word-finally	*cat*
ɾ	[dx]	tap	between vowels	*butter*
$t^{\urcorner}$	[tcl]	unreleased t	before consonants or word-finally	*fruitcake*
$\underset{\sqcap}{t}$		dental t	before dental consonants ([θ])	*eighth*
		deleted t	sometimes word-finally	*past*

Figure 7.9 Some allophones of /t/ in General American English.

Variation is even more common than Fig. 7.9 suggests. One factor influencing variation is that the more natural and colloquial speech becomes, and the faster the speaker talks, the more the sounds are shortened, reduced and generally run together. This phenomena is known as **reduction** or **hypoarticulation**. For example **assimilation** is the change in a segment to make it more like a neighboring segment. The dentalization of [t] to ([$\underset{\sqcap}{t}$]) before the dental consonant [θ] is an example of assimilation. A common type of assimilation cross-linguistically is **palatalization**, when the constriction for a segment moves closer to the palate than it normally would, because the following segment is palatal or alveolo-palatal. In the most common cases, /s/ becomes [sh], /z/ becomes [zh], /t/ becomes [ch] and /d/ becomes [jh], We saw one case of palatalization in Fig. 7.8 in the pronunciation of *because* as [b iy k ah zh], because the following word was *you've*. The lemma *you* (*you*, *your*, *you've*, and *you'd*) is extremely likely to cause palatalization in the Switchboard corpus.

REDUCTION
HYPOARTICULATION
ASSIMILATION
PALATALIZATION

DELETION

Deletion is quite common in English speech. We saw examples of deletion of final /t/ above, in the words *about* and *it*. Deletion of final /t/ and /d/ has been extensively studied. /d/ is more likely to be deleted than /t/, and both are more likely to be deleted before a consonant (Labov, 1972). Fig. 7.10 shows examples of palatalization and final t/d deletion from the Switchboard corpus.

Palatalization			Final t/d Deletion		
Phrase	**Lexical**	**Reduced**	**Phrase**	**Lexical**	**Reduced**
set your	s eh t y ow r	s eh ch er	find him	f ay n d h ih m	f ay n ix m
not yet	n aa t y eh t	n aa ch eh t	and we	ae n d w iy	eh n w iy
did you	d ih d y uw	d ih jh y ah	draft the	d r ae f t dh iy	d r ae f dh iy

Figure 7.10 Examples of palatalization and final t/d/ deletion from the Switchboard corpus. Some of the t/d examples may have glottalization instead of being completely deleted.

7.3.1 Phonetic Features

The phoneme gives us only a very gross way to model contextual effects. Many of the phonetic processes like assimilation and deletion are best modeled by more fine-grained articulatory facts about the neighboring context. Fig. 7.10 showed that /t/ and /d/ were deleted before [h], [dh], and [w]; rather than list all the possible following

phones which could influence deletion, we'd like to generalize that /t/ often deletes "before consonants". Similarly, flapping can be viewed as a kind of voicing assimilation, in which unvoiced /t/ becomes a voiced tap [dx] in between voiced vowels or glides. Rather than list every possible vowel or glide, we'd like to say that flapping happens 'near vowels or voiced segments'. Finally, vowels that precede nasal sounds [n], [m], and [ng], often acquire some of the nasal quality of the following vowel. In each of these cases, a phone is influenced by the articulation of the neighboring phones (nasal, consonantal, voiced). The reason these changes happen is that the movement of the speech articulators (tongue, lips, velum) during speech production is continuous and is subject to physical constraints like momentum. Thus an articulator may start moving during one phone to get into place in time for the next phone. When the realization of a phone is influenced by the articulatory movement of neighboring phones, we say it is influenced by **coarticulation**. **Coarticulation** is the movement of articulators to anticipate the next sound, or perseverating movement from the last sound.

COARTICULATION

We can capture generalizations about the different phones that cause coarticulation by using **distinctive features**. Features are (generally) binary variables which express some generalizations about groups of phonemes. For example the feature [voice] is true of the voiced sounds (vowels, [n], [v], [b], etc); we say they are [+voice] while unvoiced sounds are [-voice]. These articulatory features can draw on the articulatory ideas of **place** and **manner** that we described earlier. Common **place** features include [+labial] ([p, b, m]), [+coronal] ([ch d dh jh l n r s sh t th z zh]), and [+dorsal]. Manner features include [+consonantal] (or alternatively [+vocalic]), [+continuant], [+sonorant]. For vowels, features include [+high], [+low], [+back], [+round] and so on. Distinctive features are used to represent each phoneme as a matrix of feature values. Many different sets of distinctive features exist; probably any of these are perfectly adequate for most computational purposes. Fig. 7.11 shows the values for some phones from one partial set of features.

DISTINCTIVE FEATURES

	syl	son	cons	strident	nasal	high	back	round	tense	voice	labial	coronal	dorsal
b	-	-	+	-	-	-	-	+	+	+	+	-	-
p	-	-	+	-	-	-	-	-	+	-	+	-	-
iy	+	+	-	-	-	+	-	-	-	+	-	-	-

Figure 7.11 Some partial feature matrices for phones, values simplified from Chomsky and Halle (1968). *Syl* is short for syllabic; *son* for sonorant, and *cons* for consonantal.

One main use of these distinctive features is in capturing natural articulatory classes of phones. In both synthesis and recognition, as we will see, we often need to build models of how a phone behaves in a certain context. But we rarely have enough data to model the interaction of every possible left and right context phone on the behavior of a phone. For this reason we can use the relevant feature ([voice], [nasal], etc) as a useful model of the context; the feature functions as a kind of backoff model of the phone. Another use in speech recognition is to build articulatory feature detectors and use them to help in the task of phone detection; for example Kirchhoff et al. (2002) built neural-net detectors for the following set of multi-valued articulatory features and used them to improve the detection of phones in German speech recognition:

Feature	Values	Feature	Value
voicing	+voice, -voice, silence	**manner**	stop, vowel, lateral, nasal, fricative, silence
cplace	labial, coronal, palatal, velar	**vplace**	glottal, high, mid, low, silence
front-back	front, back, nil, silence	**rounding**	+round, -round, nil, silence

7.3.2 Predicting Phonetic Variation

For speech synthesis as well as recognition, we need to be able to represent the relation between the abstract category and its surface appearance, and predict the surface appearance from the abstract category and the context of the utterance. In early work in phonology, the relationship between a phoneme and its allophones was captured by writing a **phonological rule**. Here is the phonological rule for flapping in the traditional notation of Chomsky and Halle (1968):

$$\left/ \left\{ \begin{array}{c} t \\ d \end{array} \right\} \right/ \rightarrow \text{[dx]} \;/\; \acute{\text{V}} \;\underline{\quad}\; \text{V} \tag{7.1}$$

In this notation, the surface allophone appears to the right of the arrow, and the phonetic environment is indicated by the symbols surrounding the underbar (___). Simple rules like these are used in both speech recognition and synthesis when we want to generate many pronunciations for a word; in speech recognition this is often used as a first step toward picking the most likely single pronunciation for a word (see Sec. **??**).

In general, however, there are two reasons why these simple 'Chomsky-Halle'-type rules don't do well at telling us **when** a given surface variant is likely to be used. First, variation is a stochastic process; flapping sometimes occurs, and sometimes doesn't, even in the same environment. Second, many factors that are not related to the phonetic environment are important to this prediction task. Thus linguistic research and speech recognition/synthesis both rely on statistical tools to predict the surface form of a word by showing which factors cause, e.g., a particular /t/ to flap in a particular context.

7.3.3 Factors Influencing Phonetic Variation

RATE OF SPEECH

One important factor that influences phonetic variation is the **rate of speech**, generally measured in syllables per second. Rate of speech varies both across and within speakers. Many kinds of phonetic reduction processes are much more common in fast speech, including flapping, is vowel reduction, and final /t/ and /d/ deletion (Wolfram, 1969). Measuring syllables per second (or words per second) can be done with a transcription (by counting the number of words or syllables in the transcription of a region and dividing by the number of seconds), or by using signal-processing metrics (Morgan and Fosler-Lussier, 1989). Another factor affecting variation is word frequency or predictability. Final /t/ and /d/ deletion is particularly likely to happen in words which are very frequent like *and* and *just* (Labov, 1975; Neu, 1980). Deletion is also more likely when the two words surrounding the segment are a collocation (Bybee, 2000; Gregory et al., 1999; Zwicky, 1972). The phone [t] is more likely to be palatalized in frequent words and phrases. Words with higher conditional probability given the previous word are more likely to have reduced vowels, deleted consonants, and flapping (Bell et al., 2003; Gregory et al., 1999).

Other phonetic, phonological, and morphological factors affect variation as well. For example /t/ is much more likely to flap than /d/; and there are complicated interactions with syllable, foot, and word boundaries (Rhodes, 1992). As we will discuss in Ch. 8, speech is broken up into units called **intonation phrases** or **breath groups**. Words at the beginning or end of intonation phrases are longer and less likely to be reduced. As for morphology, it turns out that deletion is less likely if the word-final /t/ or /d/ is the English past tense ending (Guy, 1980). For example in Switchboard, deletion is more likely in the word *around* (73% /d/-deletion) than in the word *turned* (30% /d/-deletion) even though the two words have similar frequencies.

Variation is also affected by the speaker's state of mind. For example the word *the* can be pronounced with a full vowel [dh iy] or reduced vowel [dh ax]. It is more likely to be pronounced with the full vowel [iy] when the speaker is disfluent and having "planning problems"; in general speakers are more likely to use a full vowel than a reduced one if they don't know what they are going to say next (Fox Tree and Clark, 1997; Bell et al., 2003; Keating et al., 1994).

SOCIOLINGUISTIC
DIALECT

Sociolinguistic factors like gender, class, and **dialect** also affect pronunciation variation. North American English is often divided into eight dialect regions (Northern, Southern, New England, New York/Mid-Atlantic, North Midlands, South Midlands, Western, Canadian). Southern dialect speakers use a monophthong or near-monophthong [aa] or [ae] instead of a diphthong in some words with the vowel [ay]. In these dialects *rice* is pronounced [r aa s]. **African-American Vernacular English** (AAVE) shares many vowels with Southern American English, and also has individual words with specific pronunciations such as [b ih d n ih s] for *business* and [ae k s] for *ask*. For older speakers or those not from the American West or Midwest, the words *caught* and *cot* have different vowels ([k ao t] and [k aa t] respectively). Young American speakers or those from the West pronounce the two words *cot* and *caught* the same; the vowels [ao] and [aa] are usually not distinguished in these dialects except before [r]. For speakers of most non-American and some American dialects of English (for example Australian English), the words *Mary* ([m ey r iy]), *marry* ([m ae r iy]) and *merry* ([m eh r iy]) are all pronounced differently. Most American speakers pronounce all three of these words identically as ([m eh r iy]).

AFRICAN-AMERICAN VERNACULAR ENGLISH

REGISTER
STYLE

Other sociolinguistic differences are due to **register** or **style**; a speaker might pronounce the same word differently depending on who they were talking to or what the social situation is. One of the most well-studied examples of style-variation is the suffix *-ing* (as in *something*), which can be pronounced [ih ng] or [ih n] (this is often written *somethin'*). Most speakers use both forms; as Labov (1966) shows, they use [ih ng] when they are being more formal, and [ih n] when more casual. Wald and Shopen (1981) found that men are more likely to use the non-standard form [ih n] than women, that both men and women are more likely to use more of the standard form [ih ng] when the addressee is a women, and that men (but not women) tend to switch to [ih n] when they are talking with friends.

Many of these results on predicting variation rely on logistic regression on phonetically-transcribed corpora, a technique with a long history in the analysis of phonetic variation (Cedergren and Sankoff, 1974), particularly using the VARBRUL and GOLDVARB software (Rand and Sankoff, 1990).

Finally, the detailed acoustic realization of a particular phone is very strongly in-

fluenced by **coarticulation** with its neighboring phones. We will return to these fine-grained phonetic details in the following chapters (Sec. **??** and Sec. **??**) after we introduce acoustic phonetics.

7.4 Acoustic Phonetics and Signals

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

7.4.1 Waves

Acoustic analysis is based on the sine and cosine functions. Fig. 7.12 shows a plot of a sine wave, in particular the function:

$$y = A * sin(2\pi f t) \tag{7.2}$$

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

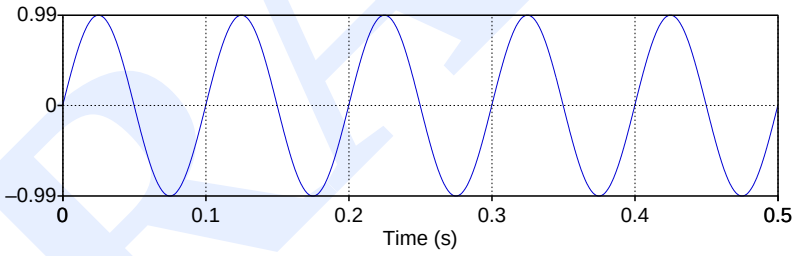

Figure 7.12 A sine wave with a frequency of 10 Hz and an amplitude of 1.

FREQUENCY
AMPLITUDE
CYCLES PER SECOND
HERTZ

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, i.e. the number of **cycles**. We usually measure frequency in **cycles per second**. The signal in Fig. 7.12 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. 7.12 would be described as 10 Hz. The **amplitude** A of a sine wave is the maximum value on the Y axis.

PERIOD

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

$$T = \frac{1}{f} \tag{7.3}$$

In Fig. 7.12 we can see that each cycle lasts a tenth of a second, hence $T = .1$ seconds.

7.4.2 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 to imagine a vertical plate which is 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. Fig. 7.13 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".

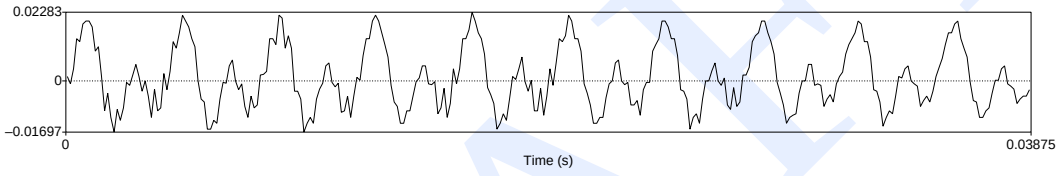

Figure 7.13 A waveform of the vowel [iy] from an utterance to be shown in Fig. 7.17. The *y*-axis shows the level of air pressure above and below normal atmospheric pressure. The *x*-axis shows time. Notice that the wave repeats regularly.

SAMPLING

SAMPLING RATE

NYQUIST FREQUENCY

TELEPHONE-BANDWIDTH

WIDEBAND

Let's explore how the digital representation of the sound wave shown in Fig. 7.13 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**. A signal is sampled by measuring its amplitude at a particular time; the **sampling rate** is the number of samples taken per second. In order to accurately measure a wave, it is necessary to 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 less than two samples will cause the frequency of the wave to be completely missed. Thus the maximum 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.

Even an 8,000 Hz sampling rate requires 8000 amplitude measurements for each second of speech, and so it is important to store the amplitude measurement 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
QUANTIZATION called **quantization** because there is a minimum granularity (the quantum size) and all values which are closer together than this quantum size are represented identically.

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, while microphone data is often sampled at 16 kHz and stored as 16-bit samples. Another parameter of these formats is the
CHANNELS 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 whether each sample is stored linearly or whether it is 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
PCM 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):

$$F(x) = \frac{sgn(s)\log(1+\mu|s|)}{\log(1+\mu)} \tag{7.4}$$

There are a number of standard file formats for storing the resulting digitized wavefile, such as Microsoft's WAV, Apple AIFF and Sun 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. Fig. 7.14 shows a simple .wav file with a single data chunk together with its format chunk:

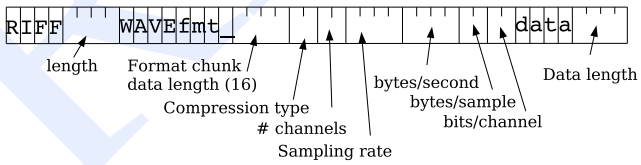

Figure 7.14 Microsoft wavefile header format, assuming simple file with one chunk. Following this 44-byte header would be the data chunk.

7.4.3 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. 7.13 that although not exactly a sine, that the wave is nonetheless periodic, and that there are 10 repetitions of the wave in the 38.75 milliseconds (.03875 seconds) we have captured in the figure. Thus the frequency of this segment of the wave is 10/.03875 or 258 Hz.

Where does this periodic 258Hz wave come from? It comes from the speed of vibration of the vocal folds; since the waveform in Fig. 7.13 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 longs. Thus when the vocal folds are vibrating, we expect to see regular peaks in amplitude of the kind we see in Fig. 7.13, 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 $\mathbf{F}_0$. We can plot F_0 over time in a **pitch track**. Fig. 7.15 shows the pitch track of a short question, "Three o'clock?" represented below the waveform. Note the rise in F_0 at the end of the question.

FUNDAMENTAL FREQUENCY
F0
PITCH TRACK

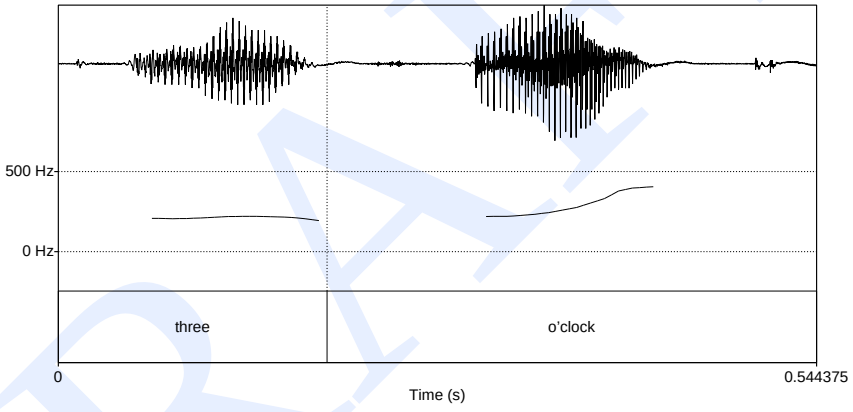

Figure 7.15 Pitch track of the question "Three o'clock?", shown below the wavefile. Note the rise in F0 at the end of the question. Note the lack of pitch trace during the very quiet part (the "o'" of "o'clock"; automatic pitch tracking is based on counting the pulses in the voiced regions, and doesn't work if there is no voicing (or insufficient sound at all).

The vertical axis in Fig. 7.13 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, while 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.

$$\text{RMS amplitude}_{i=1}^{N} = \sqrt{\sum_{i=1}^{N} \frac{x_i^2}{N}} \tag{7.5}$$

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

$$\text{Power} = \frac{1}{N}\sum_{i=1}^{n} x[i]^2 \tag{7.6}$$

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 = $2 \times 10^{-5} Pa$ then intensity is defined as follows:

$$\text{Intensity} = 10\log_{10}\frac{1}{NP_0}\sum_{i=1}^{n} x_i^2 \tag{7.7}$$

Fig. 7.16 shows an intensity plot for the sentence "Is it a long movie?" from the CallHome corpus, again shown below the waveform plot.

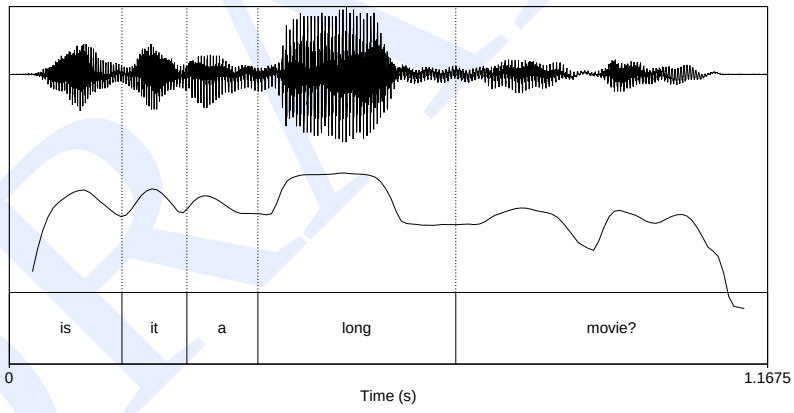

Figure 7.16 Intensity plot for the sentence "Is it a long movie?". Note the intensity peaks at each vowel, and the especially high peak for the word *long*.

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 100Hz and 1000Hz, 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 Volkmann, 1940). A mel is a unit of pitch defined so 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:

MEL

$$m = 1127 \ln(1 + \frac{f}{700}) \tag{7.8}$$

We will return to the mel scale in Ch. 9 when we introduce the MFCC representation of speech used 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.

Various algorithms exist for automatically extracting F_0. 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 will return to in Ch. 9. There are various publicly available pitch extraction toolkits; for example an augmented autocorrelation pitch tracker is provided with `Praat` (Boersma and Weenink, 2005).

PITCH EXTRACTION

7.4.4 Interpreting 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. 7.16). Length in time manifests itself directly on the x-axis, while 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. 7.13, each major peak corresponding to an opening of the vocal folds. Fig. 7.17 shows the waveform of the short phrase 'she just had a baby'. We have labeled this waveform with word and phone labels. Notice that each of the six vowels in Fig. 7.17, [iy], [ax], [ae], [ax], [ey], [iy], all have regular amplitude peaks indicating voicing.

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. 7.17.

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 very noisy, irregular waveform. This can be seen somewhat in Fig. 7.17; it's even clearer in Fig. 7.18, where we've magnified just the first word *she*.

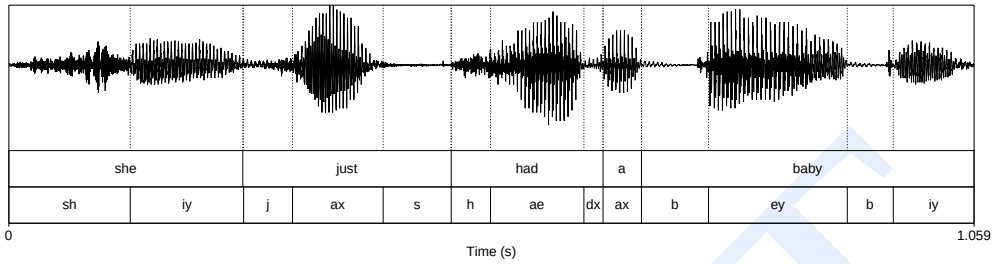

Figure 7.17 A waveform of the sentence "She just had a baby" from the Switchboard corpus (conversation 4325). The speaker is female, was 20 years old in 1991, which is approximately when the recording was made, and speaks the South Midlands dialect of American English.

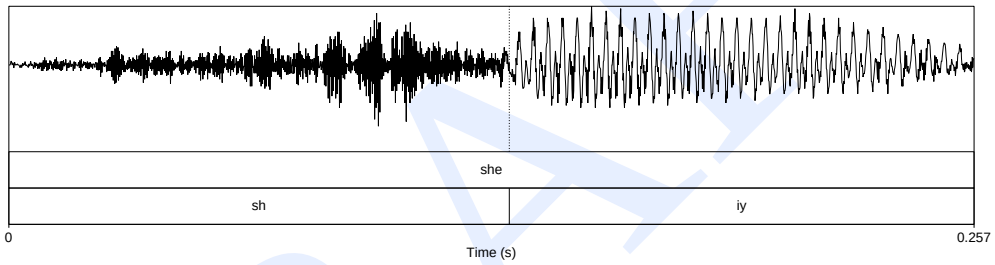

Figure 7.18 A more detailed view of the first word "she" extracted from the wavefile in Fig. 7.17. Notice the difference between the random noise of the fricative [sh] and the regular voicing of the vowel [iy].

7.4.5 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. 7.19. This waveform was created (in `Praat`) by summing two sine waveforms, one of frequency 10 Hz and one of frequency 100 Hz.

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. Fig. 7.20 shows the spectrum of Fig. 7.19. Frequency in Hz is on the x-axis and amplitude on the y-axis. Note that there are 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 soundwave at a particular time point.

Let's look now at the frequency components of a speech waveform. Fig. 7.21 shows

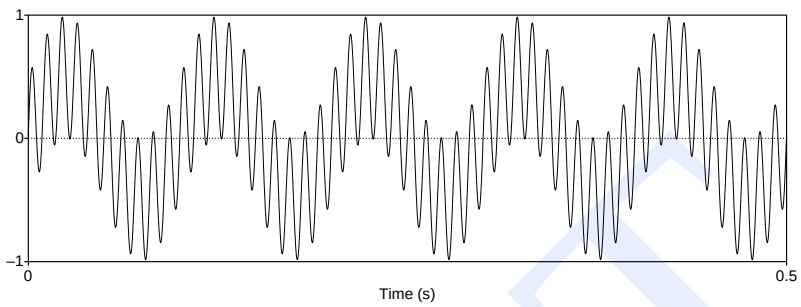

Figure 7.19 A waveform created by summing two sine waveforms, one of frequency 10 Hz (note the 5 repetitions in the half-second window) and one of frequency 100 Hz, both with amplitude 1.

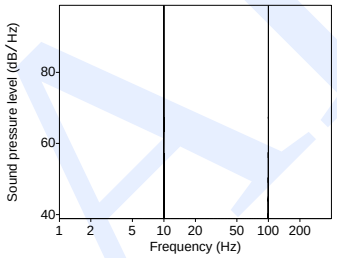

Figure 7.20 The spectrum of the waveform in Fig. 7.19.

part of the waveform for the vowel [ae] of the word *had*, cut out from the sentence shown in Fig. 7.17.

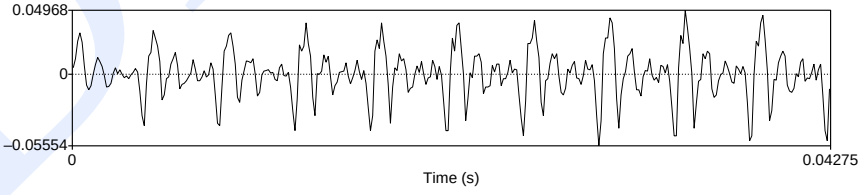

Figure 7.21 The waveform of part of the vowel [ae] from the word *had* cut out from the waveform shown in Fig. 7.17.

Note that there is a complex wave which repeats about ten times in the figure; but there is also a smaller repeated wave which 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.

Fig. 7.22 shows a smoothed spectrum for the waveform in Fig. 7.21, computed via a Discrete Fourier Transform (DFT).

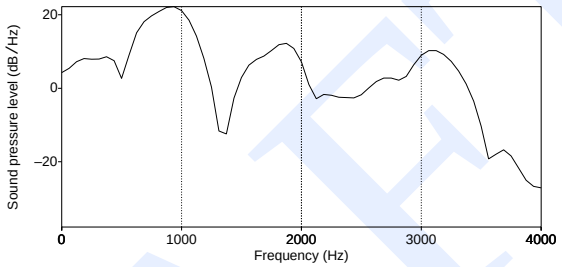

Figure 7.22 A spectrum for the vowel [ae] from the word *had* in the waveform of *She just had a baby* in Fig. 7.17.

The *x*-axis of a spectrum shows frequency while 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. 7.22 shows that there are 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. 7.21!

Why is a spectrum useful? It turns out that these spectral peaks that are easily visible in a spectrum are very 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 looking at the spectrum 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
INNER EAR

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 corresponding to the amplitude of

SPECTROGRAM

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).

Fig. 7.23 shows spectrograms of 3 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. 7.21.

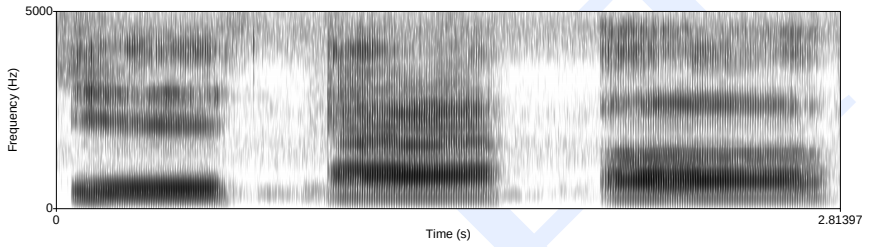

Figure 7.23 Spectrograms for three American English vowels, [ih], [ae], and [uh], spoken by the first author.

FORMANT

Each dark bar (or spectral peak) is called a **formant**. As we will 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 different position for the 3 vowels; it's low for [ih] (centered at about 470Hz) and somewhat higher for [ae] and [ah] (somewhere around 800Hz) 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. Fig. 7.24 shows the spectrogram of 'she just had a baby' whose waveform was shown in Fig. 7.17. F1 and F2 (and also F3) are pretty clear for the [ax] of *just*, the [ae] of *had*, and the [ey] of *baby*.

What specific clues can spectral representations give for phone identification? First, since different vowels have their formants at characteristic places, the spectrum can be used to 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. 7.17. The spectrum for this vowel is shown in Fig. 7.25. The first formant of [iy] is 540 Hz; much lower than the first formant for [ae], while 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. 7.24 just around 0.5 seconds.

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 characteristic of the speakers vocal tract rather than by individual vowels. Formants also can be used to identify the nasal

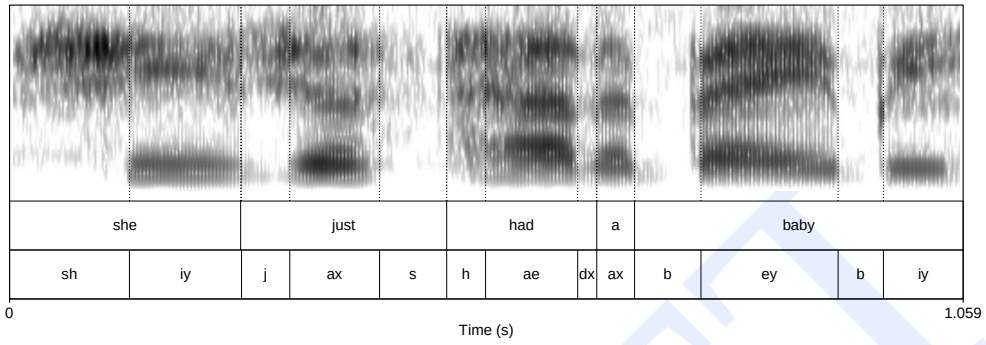

Figure 7.24 A spectrogram of the sentence "She just had a baby" whose waveform was shown in Fig. 7.17. We can think of a spectrogram is as a collection of spectra (time-slices) like Fig. 7.22 placed end to end. Note

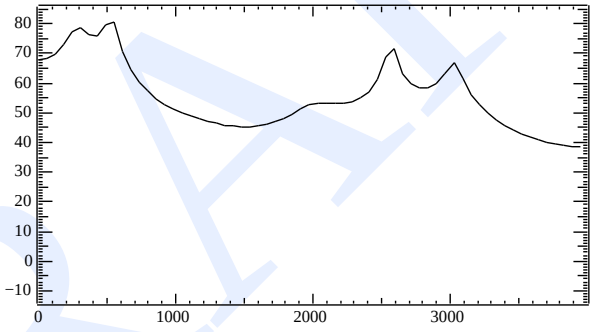

Figure 7.25 A smoothed (LPC) spectrum for the vowel [iy] at the start of *She just had a baby*. Note that the first formant (540 Hz) is much lower than the first formant for [ae] shown in Fig. 7.22, while the second formant (2581 Hz) is much higher than the second formant for [ae].

phones [n], [m], and [ng], and the liquids [l] and [r].

7.4.6 The Source-Filter Model

SOURCE-FILTER

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**).

HARMONICS

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, i.e. have much less amplitude than the wave at the fundamental frequency.

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.

Now 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 oral tract and the corresponding spectrum. Fig. 7.26 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.

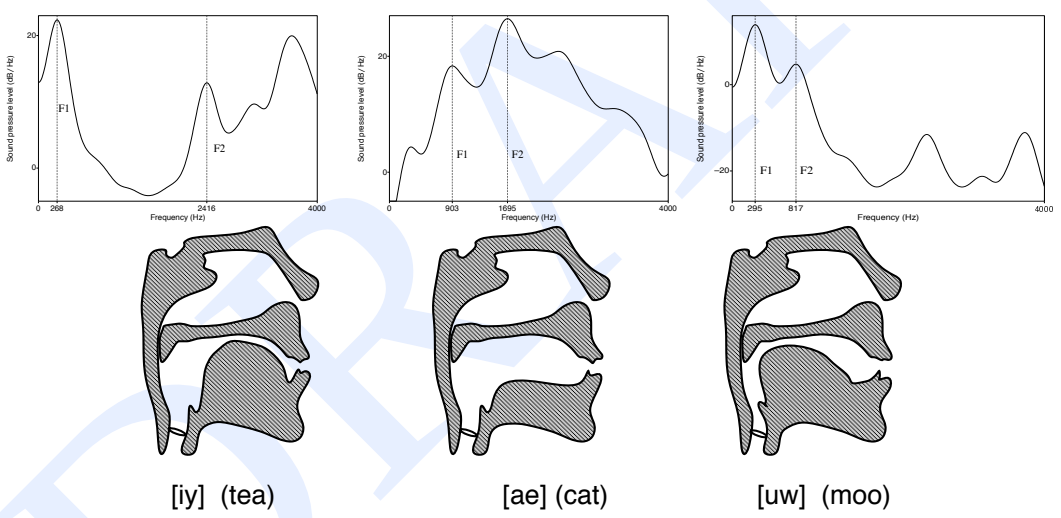

Figure 7.26 Visualizing the vocal tract position as a filter: the tongue positions for three English vowels and the resulting smoothed spectra showing F1 and F2. Tongue positions modeled after Ladefoged (1996).

7.5 PHONETIC RESOURCES

PRONUNCIATION DICTIONARIES

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.

The CELEX dictionary (Baayen et al., 1995) is the most richly annotated of the dictionaries. It includes all the words in the Oxford Advanced Learner's Dictionary (1974) (41,000 lemmata) and the Longman Dictionary of Contemporary English (1978) (53,000 lemmata), in total it has pronunciations for 160,595 wordforms. Its (British rather than American) pronunciations are transcribed using 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 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 an 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.

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, where 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 Ch. 8.

Another useful resource is a **phonetically annotated corpus**, in which a collection of waveforms is hand-labeled with the corresponding string of phones. Two important phonetic corpora in English are the TIMIT corpus and the Switchboard Transcription Project 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, where 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 in the transcript is associated with a start and end time in the waveform; we showed a graphical example of a time-aligned transcription in Fig. 7.17.

TIME-ALIGNED TRANSCRIPTION

The phoneset for TIMIT, and for the Switchboard Transcription Project corpus be-

low, 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. 7.1 and Fig. 7.2; 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. 7.27.

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 7.27 Phonetic transcription from the TIMIT corpus. Note 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. Fig. 7.28 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 7.28 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 resyllabification (the [v] of *of* attaches as the onset of *in*). Time is given in number of seconds from beginning of sentence to start of each syllable.

Phonetically transcribed corpora are also available for other languages; the Kiel corpus of German is commonly used, as are various 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 free Praat package (Boersma and Weenink, 2005), which includes spectrum and spectrogram analysis, pitch extraction and formant analysis, and an embedded scripting language for automation. It is available on Microsoft, Macintosh, and UNIX environments.

7.6 Advanced: Articulatory and Gestural Phonology

We saw in Sec. 7.3.1 that we could use **distinctive features** to capture generalizations across phone class. These generalizations were mainly articulatory (although some, like [strident] and the vowel height features, are primarily acoustic).

ARTICULATORY PHONOLOGY
ARTICULATORY GESTURE

This idea that articulation underlies phonetic production is used in a more sophisticated way in **articulatory phonology**, in which the **articulatory gesture** is the underlying phonological abstraction (Browman and Goldstein, 1986, 1992). Articulatory gestures are defined as parameterized **dynamical systems**. Since speech production requires the coordinated actions of tongue, lips, glottis, etc, articulatory phonology represents a speech utterances as a sequence of potentially overlapping articulatory gestures. Fig. 7.29 shows the sequent of gestures (or **gestural score**) required for the production of the word *pawn* [p aa n]. The lips first close, then the glottis opens, then the tongue body moves back toward the pharynx wall for the vowel [aa], the velum drops for the nasal sounds, and finally the tongue tip closes against the alveolar ridge. The lines in the diagram indicate gestures which are phased with respect to each other. With such a gestural representation, the nasality in the [aa] vowel is explained by the timing of the gestures; the velum drops before the tongue tip has quite closed.

GESTURAL SCORE

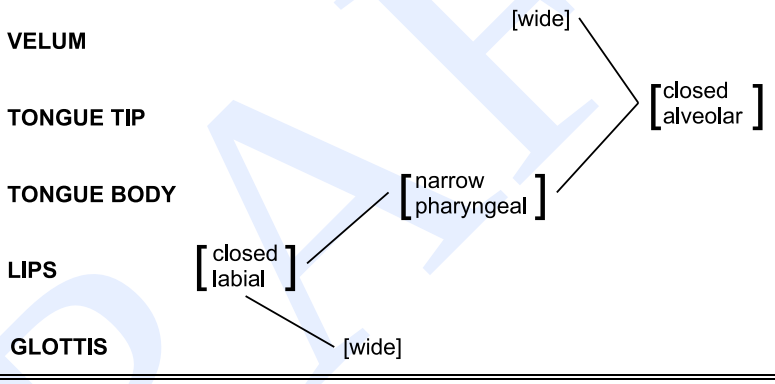

Figure 7.29 The gestural score for the word *pawn* as pronounced [p aa n], after Browman and Goldstein (1989) and Browman and Goldstein (1995).

The intuition behind articulatory phonology is that the gestural score is likely to be much better as a set of hidden states at capturing the continuous nature of speech than a discrete sequence of phones. In addition, using articulatory gestures as a basic unit can help in modeling the fine-grained effects of coarticulation of neighboring gestures that we will explore further when we introduce **diphones** (Sec. **??**) and **triphones** (Sec. **??**).

Computational implementations of articulatory phonology have recently appeared in speech recognition, using articulatory gestures rather than phones as the underlying representation or hidden variable. Since multiple articulators (tongue, lips, etc) can move simultaneously, using gestures as the hidden variable implies a multi-tier hidden representation. Fig. 7.30 shows the articulatory feature set used in the work of Livescu and Glass (2004) and Livescu (2005); Fig. 7.31 shows examples of how phones are mapped onto this feature set.

Feature	Description	value = meaning
LIP-LOC	position of lips	LAB = labial (neutral position); PRO = protruded (rounded); DEN = dental
LIP-OPEN	degree of opening of lips	CL = closed; CR = critical (labial/labio-dental fricative); NA = narrow (e.g., [w], [uw]); WI = wide (all other sounds)
TT-LOC	location of tongue tip	DEN = inter-dental ([th], [dh]); ALV = alveolar ([t], [n]); P-A = palato-alveolar ([sh]); RET = retroflex ([r])
TT-OPEN	degree of opening of tongue tip	CL = closed (stop); CR = critical (fricative); NA = narrow ([r], alveolar glide); M-N = medium-narrow;MID = medium;WI = wide
TB-LOC	location of tongue body	PAL = palatal (e.g. [sh], [y]); VEL = velar (e.g., [k], [ng]); UVU = uvular (neutral position); PHA = pharyngeal (e.g. [aa])
TB-OPEN	degree of opening of tongue body	CL = closed (stop); CR = critical (e.g. fricated [g] in "legal"); NA = narrow (e.g. [y]); M-N = medium-narrow; MID = medium; WI = wide
VEL	state of the velum	CL = closed (non-nasal); OP = open (nasal)
GLOT	state of the glottis	CL = closed (glottal stop); CR = critical (voiced); OP = open (voiceless)

Figure 7.30 Articulatory-phonology-based feature set from Livescu (2005).

phone	LIP-LOC	LIP-OPEN	TT-LOC	TT-OPEN	TB-LOC	TB-OPEN	VEL	GLOT
aa	LAB	W	ALV	W	PHA	M-N	CL(.9),OP(.1)	CR
ae	LAB	W	ALV	W	VEL	W	CL(.9),OP(.1)	CR
b	LAB	CR	ALV	M	UVU	W	CL	CR
f	DEN	CR	ALV	M	VEL	M	CL	OP
n	LAB	W	ALV	CL	UVU	M	OP	CR
s	LAB	W	ALV	CR	UVU	M	CL	OP
uw	PRO	N	P-A	W	VEL	N	CL(.9),OP(.1)	CR

Figure 7.31 Livescu (2005): sample of mapping from phones to underyling target articulatory feature values. Note that some values are probabilistic.

7.7 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 speech 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 twelfth century exploited the concept of the phoneme, 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 3-volume source on dialects of English.

Many of the classic insights in acoustic phonetics had been developed by the late 1950's or early 1960's; 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, 1970; 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 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.

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. There are a wide variety of books that 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). An excellent textbook on digital signal processing is Lyons (2004).

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

7.1 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]

7.2 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]

7.3 Ira Gershwin's lyric for *Let's Call the Whole Thing Off* talks about two pronunciations of the word "either" (in addition to the tomato and potato example given at the beginning of the chapter. Transcribe Ira Gershwin's two pronunciations of "either" in the ARPAbet.

7.4 Transcribe the following words in the ARPAbet:

a. dark
b. suit

c. greasy
d. wash
e. water

7.5 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 wavfile, and to look at the wavefile and the spectrogram.

7.6 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], Philadelphia, PA.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M. L., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, *113*(2), 1001–1024.

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.

Browman, C. P. and Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155–180.

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*, 201–250.

Browman, C. P. and Goldstein, L. (1995). Dynamics and articulatory phonology. In Port, R. and v. Gelder, T. (Eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 175–193. MIT Press.

Bybee, J. L. (2000). The phonology of the lexicon: evidence from lexical diffusion. In Barlow, M. and Kemmer, S. (Eds.), *Usage-based Models of Language*, pp. 65–85. CSLI, Stanford.

Cedergren, H. J. and Sankoff, D. (1974). Variable rules: performance as a statistical reflection of competence. *Language*, *50*(2), 333–355.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper and Row.

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

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.

Fox Tree, J. E. and Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*, 151–167.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*, 765–768.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE ICASSP-92*, San Francisco, pp. 517–520. IEEE.

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*, Philadelphia, PA, pp. S24–27.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., and Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *CLS-99*, pp. 151–166. University of Chicago, Chicago.

Guy, G. R. (1980). Variation in the group and the individual: The case of final stop deletion. In Labov, W. (Ed.), *Locating Language in Time and Space*, pp. 1–36. Academic.

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

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

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

Keating, P. A., Byrd, D., Flemming, E., and Todaka, Y. (1994). Phonetic analysis of word and segment variation using the TIMIT corpus of American English. *Speech Communication*, *14*, 131–142.

Kirchhoff, K., Fink, G. A., and Sagerer, G. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, *37*, 303–319.

Koenig, W., Dunn, H. K., Y., L., and Lacy (1946). The sound spectrograph. *Journal of the Acoustical Society of America*, *18*, 19–49.

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, D.C.

Labov, W. (1972). The internal evolution of linguistic rules. In Stockwell, R. P. and Macaulay, R. K. S. (Eds.), *Linguistic Change and Generative Theory*, pp. 101–171. Indiana University Press, Bloomington.

Labov, W. (1975). *The quantitative study of linguistic structure*. Pennsylvania Working Papers on Linguistic Change and Variation v.1 no. 3. U.S. Regional Survey, Philadelphia, PA.

Labov, W. (1994). *Principles of Linguistic Change: Internal Factors*. Blackwell, Oxford.

Ladefoged, P. (1993). *A Course in Phonetics*. Harcourt Brace Jovanovich. Third Edition.

Ladefoged, P. (1996). *Elements of Acoustic Phonetics*. University of Chicago, Chicago, IL. Second Edition.

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*, Beijing, China, pp. 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.

Livescu, K. (2005). *Feature-Based Pronuncaition Modeling for Automatic Speech Recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

Livescu, K. and Glass, J. (2004). Feature-based pronunciation modeling with trainable asynchrony probabilities. In *ICSLP-04*, Jeju, South Korea.

Lyons, R. G. (2004). *Understanding Digital Signal Processing*. Prentice Hall, Upper Saddle River, NJ. 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. *Journal of the Acoustical Society of America*, *27*, 338–352.

Morgan, N. and Fosler-Lussier, E. (1989). Combining multiple estimators of speaking rate. In *IEEE ICASSP-89*.

Neu, H. (1980). Ranking of constraints on /t,d/ deletion in American English: A statistical analysis. In Labov, W. (Ed.), *Locating Language in Time and Space*, pp. 37–54. Academic Press.

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.

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

Pullum, G. K. and Ladusaw, W. A. (1996). *Phonetic Symbol Guide*. University of Chicago, Chicago, IL. Second Edition.

Rand, D. and Sankoff, D. (1990). Goldvarb: A variable rule application for the macintosh. Available at http://www.crm.umontreal.ca/ sankoff/GoldVarb_Eng.html.

Rhodes, R. A. (1992). Flapping in American English. In Dressler, W. U., Prinzhorn, M., and Rennison, J. (Eds.), *Proceedings of the 7th International Phonology Meeting*, pp. 217–232. Rosenberg and Sellier, Turin.

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*, Oahu, Hawaii.

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

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

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

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

Stevens, K. N. and House, A. S. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, *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, 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. *Journal of the Acoustical Society of America*, *8*, 185–190.

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

Wald, B. and Shopen, T. (1981). A researcher's guide to the sociolinguistic variable (ING). In Shopen, T. and Williams, J. M. (Eds.), *Style and Variables in English*, pp. 219–249. Winthrop Publishers.

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

Wolfram, W. A. (1969). *A Sociolinguistic Description of Detroit Negro Speech*. Center for Applied Linguistics, Washington, D.C.

Zappa, F. and Zappa, M. U. (1982). Valley girl. From Frank Zappa album *Ship Arriving Too Late To Save A Drowning Witch*.

Zwicky, A. (1972). On Casual Speech. In *CLS-72*, pp. 607–615. University of Chicago.