UNIT-4
Speech Fundamentals

Phonetics

Phonetics is the systematic study of the human ability to make and hear sounds which use the vocal organs of speech, especially for producing oral language. It is usually divided into the three branches of (1) articulatory, (2) acoustic and (3) auditory phonetics. It is also traditionally differentiated from (though overlaps with) the field of phonology, which is the formal study of the sound systems (phonologies) of languages, especially the universal properties displayed in ALL languages, such as the psycholinguistic aspects of phonological processing and acquisition.

Phonetic transcription and the IPA

In phonetics the most basic segments are called *phones*, which may be defined as units in speech which can be distinguished acoustically or articulatorily. This definition allows for different degrees of *wideness*.[1] In many contexts phones may be thought of as acoustic or articulatory *targets* which may or may not be fully reached in actual speech. Another, more commonly used segment is the *phoneme*

The International Phonetic Alphabet (IPA) is a system of phonetic notation which provides a standardized system of transcribing phonetic segments up to a certain degree of detail. It may be represented visually using charts, which may be found in full in Appendix A. We will leave a more detailed description of the IPA to the end of this chapter, but for now just be aware that text in square brackets [] is phonetic transcription in IPA

To understand the IPA's taxonomy of phones, it is important to consider articulatory, acoustic, and auditory phonetics.

Articulatory phonetics

Articulatory phonetics is concerned with how the sounds of language are physically produced by the vocal apparatus. The units articulatory phonetics deals with are known as *gestures*, which are abstract characterizations of articulatory events.

Speaking in terms of articulation, the sounds that we utter to make language can be split into two different types: consonants and vowels. For the purposes of articulatory phonetics, consonant sounds are typically characterized as sounds that have constricted or closed configurations of the vocal tract. Vowels, on the other hand, are characterized in articulatory terms as having relatively little constriction; that is, an open configuration of the vocal tract. Vowels carry much of the pitch of speech and can be held different durations, such as a half a beat, one beat, two beats, three beats, etc. of speech rhythm. Consonants, on the other hand, do not carry the prosodic pitch (especially if devoiced and not nasalized) and do not display the potential for the durations that vowels can have. Linguists may also speak of 'semi-vowels' or 'semi-consonants' (often used as synonymous terms). For example, a sound such as [w] phonetically seems more like a vowel (with relative lack of constriction or closure

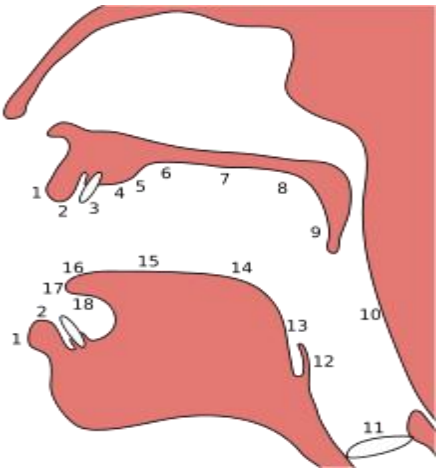
of the vocal tract) but, phonologically speaking, behaves as a consonant in that it always appears before a vowel sound at the beginning (onset) of a syllable.

Consonants

Phoneticians generally characterize consonants as being distinguished by settings of the independent variables *place of articulation* (POA) and *manner of articulation* (MOA). In layman's terminology, POA is "where" the consonant is produced, while MOA is "how" the consonant is produced.

The following are descriptions of the different POAs:

A diagram of the vocal tract showing the different places of articulation



- Bilabial segments are produced with the lips held together, for instance the [p] sound of the English *pin*, or the [b] sound in *bin*.
- Labiodental segments are produced by holding the upper teeth to the lower lip, like in the [f] sound of English *fin*.
- Dental consonants have the tongue making contact with the upper teeth (area 3 in the diagram). An example from English is the [θ] sound in the word *thin*.
- Alveolar consonants have the tongue touching the area of the mouth known as the alveolar ridge (area 4 in the diagram). Examples include the [t] in *tin* and [s] in *sin*.
- Postalveolar consonants are similar to alveolars but more retracted (in area 5 in the diagram), like the [ʃ] of *shin*.
- Palatal consonants are articulated at the hard palate (the middle part of the roof of the mouth, area 7 in the diagram). In English the palatal [j] sound appears in the word *young*.
- Velar consonants are articulated at the soft palate (the back part of the roof of the mouth, known also as the *velum*, area 8 in the diagram). English [k] is velar, like in the word *kin*.
- Glottal consonants are articulated far back in the throat, at the *glottis* (area 11 in the diagram, effectively the vocal folds). English [h] may be regarded as glottal.[2]

- Doubly articulated consonants have two points of articulation, such as the English labio-velar [w] of *wit*.

Other POAs are also possible.

MOA involves a number of different variables which may vary independently:

- Voicing: Try pronouncing the hissing sound [s] of the English word *sip*. Elongate the sound until you can produce it continuously for five seconds. Then do the same for the [z] sound in *zip*. Hold your hand to your throat, observing the different in tactile sensation between the two. you should notice that [z] creates vibrations, while [s] does not. This rapid vibration is in fact caused by the *vocal folds*, and it is referred to as *voicing*. Many different sounds can contrast solely based on a voicing difference: English [b, p] in *bin*, *pin*, [d, t] in *din*, *tin*, et cetera.

- Nasality: Some sounds are produced with airflow through the nasal cavity. These are known as *nasals*. Nasal consonants in English include the [n] of *not*, the [m] of *mit*, and the [ŋ] of *sing*. Nasals may also contrast for voicing in some languages, but this is rare — in most languages, nasals are voiced.

- Obstruency: Consonants involving a total obstruction of airflow are known as *stops* or *plosives*. Examples include English [p, b, t, d, k, g]. *Fricatives* are consonants with a steady stricture causing friction, for example [f, v, s, z, ʃ, ʒ]. *Affricates* begin with a stop-like closure followed by frication, like the [tʃ, dʒ] of English *chip*, *jeans*.

- Sonorancy: Non-obstruents are classed as sonorants. This includes the already-mentioned nasals. Another important type of sonorant found in English is the approximant, in which articulatory organs produce a narrowing of the vocal tract, but leave enough space for air to flow without much audible turbulence. Examples include English [w, j, l, ɹ].

Knowing this information is enough to construct a simplified IPA chart of the consonants of English. As is conventional, MOA is organized in rows, and POA columns. Voicing pairs occur in the same cells; the ones in bold are voiced while the rest are voiceless.

Simplified IPA consonant chart (for English)

	Bilabial (or Labio-velar)	Labiodental	Dental	Alveolar	Postalveolar	Palatal	Velar	Glottal
Plosive	pb			td			kg	
Fricative		fv	θ[3] ð[4]	sz	fʒ			
Affricate					tʃ dʒ			
Nasal	m			n			ŋ	
Approximant	w			ɹ l			j	

Vowels

Vowels are very different from consonants, but our method of decomposing soundinto sets of features works equally well. Vowels can essentially be viewed as being combinations of three variables:

- Height: This measured how close your tongue is to the roof of your mouth. For example, try pronouncing [æ] (as in "cat") and [i] (as in "feet"). Your mouth should be much more open for the former than the latter. Thus [æ] is called either *open* or *low*, and [i] either *closed* or *high*.

- Backness: This is what is sounds like. Try, for example, alternating between pronouncing the vowels [æ] (as in "cat") and [ɑ] (as in "cot"), and get a feel for the position of your tongue in your mouth. It should move forward for [æ] and back for [ɑ], which is why the former is called a *front* vowel and the latter a *back* vowel.

- Rounding: Pronouncing the vowels [i] and [u], and look at your lips in a mirror. They should look puckered up for [u] and spread out for /i/.[5] In general, this "puckering" is referred to in phonetics as *rounding*.

Back vowels[6] tend to be rounded, and front vowels unrounded, for reasons which will be covered later in this chapter. However, this tendency is not universal. For instance, the vowel in the French word *bœuf* is what would result from the vowel of the English word *bet* being pronounced with rounding. Some East and Southeast Asian languages possess unrounded back vowels, which are difficult to describe without a sound sample.[7]

The *cardinal vowels* are a set of idealized vowels used by phoneticians as a base of reference.

The IPA orders the vowels in a similar way to the consonants, separating the three main distinguishing variables into different dimensions. The *vowel trapezoid* may be thought of as a rough diagram of the mouth, with the left being the front, the right the back, and the vertical direction representing height in the mouth. Each vowel is positioned thusly based on height and backness. Rounding isn't indicated by location, but when pairs of vowels sharing the same height and backness occur next to each other, the left member is always unrounded, and the right a rounded vowel. Otherwise, just use the general heuristic that rounded vowels are usually back. The following is a simplified version of the IPA vowel chart:

[æ]	cat	[ɑ]	cot	[ɐ]	Cut
[ɛ]	kelp	[ə]	carrot	[ɔ]	Caugh t
[e]	cake	[o]	coat	[ɪ]	Kin
[ʊ]	cushion	[i]	keep	[u]	Cool

Note, however, that in phonetics we can describe any segment in arbitrarily fine detail. As such, when we say that, say, the vowel in "cat" *is* [æ], we are sacrificing precision.

Some of these vowels have no English equivalent, but may be familiar from foreign languages. [a] represents the sound in Spanish "h*a*blo", the front rounded [œ] vowel is that in French "b*œu*f", [y] is that of German "h*ü*ten", and [ɯ] (perhaps the most exotic to most English speakers) is found as the first vowel of European Portuguese "p*e*gar".

Other types of phonetics

Acoustic phonetics

Acoustic phonetics deals with the physical medium of speech -- that is, how speech manipulates sound waves.

Sound is composed of waves of high- and low-pressure areas which propagate through air. The most basic way to view sound is as a *wave function*. This plots the pressure measured by the sound-recording device against time, corresponding closely to the physical nature of sound. Loudness may be found by looking at the *amplitude* of the sound at a given time.

However, this approach is fairly limited. Humans, in fact, don't process sound using this raw data. The ear analyzes sound by decomposing it into its constituent frequencies, a mathematical algorithm known as the *Fourier transform*.

As a sound is produced in the oral tract, the column of air in the tract serves as a *harmonic oscillator*, oscillating at numerous frequencies simultaneously. Some of the frequencies of oscillation are at higher amplitudes than others, a property called *resonance*. The *resonant frequencies* (frequencies with relatively high resonance) of

the vocal tract are known in phonetics as *formants*[8]. The formants in a speech sound are numbered by their frequency: f1 (pronounced *eff-one*) has the lowest frequency, followed by f2, f3, etc. The analysis of formants turns out to be key to acoustic phonetics, as any change in the shape of the vocal cavity changes which resonances are dominant.

There are two basic ways to analyze the formants of a speech signal. Firstly, at any given time the sound contains a mixture of different frequencies of sound. The relative amplitudes (strengths) of different frequencies at a particular time may be shown as a *frequency spectrum*. As you can see on the right, frequency is plotted against amplitude, and formants show up as peaks.

Spectrogram of American English vowels [i, u, ɑ] showing the formants *f*1 and *f*2

Another way to view formants is by using a *spectrogram*. This plots time against frequency, with amplitude represented by darkness. Formants show up as dark bands, and their movement may be tracked through time.

Given the development of modern technology, acoustic analysis is now accessible to anyone with a computer and a microphone

Phonetic Resources

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

In addition to resources like dictionaries and corpora, there are many useful pho- netic 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 avail- able on Microsoft, Macintosh, and UNIX environments.

Advanced: Articulatory and Gestural Phonology

Articulatory Phonology

The intuition behind articulatory phonology is that the gestural score is likely to bemuch 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**.

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)

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

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

Gestural phonology

Gestural phonology is phonological model in which each sound is broken down into the individual articulatory actions that are used to produce a sound. For example, the first gesture in the word "Matt" is to pull one's lips together to create the initial [m]. Under this model, sounds are described and stored as mental representations in terms of their composite gestures.