The Sounds of Speech

or

'Things that go bump in the mouth'

Overview

- **Speech production**
  - the vocal tract; how we use it to produce speech
- **Phonemes** - the minimal units of speech
  - describing and classifying speech sounds
- **Phonetics** - speech sounds in the real world
  - how speech sounds change in context
- **Beyond the phoneme**
  - the wider context, and effects of external factors

Speech Production

- **air pressure** from lungs builds up behind closed vocal cords
- **vocal cords** are repeatedly forced apart and pulled together again, causing the air in the vocal tract to **vibrate** quasi-periodically
- **rate of vibration** of the vocal folds determines the **fundamental frequency** of the speech waveform
- **fundamental frequency** (F0) contributes to the perceived **pitch** of the voice

Speech Production

- The vocal tract forms a **resonator** with a complex shape. Speech is produced by using the **articulators** to change the shape of the vocal tract, modifying its resonant characteristics
- Different **configurations of the vocal tract** enhance some of the harmonics of the fundamental, and suppress (damp) others
- Principal articulator is the tongue, but jaw, lips, soft palate and teeth are also involved

Phonemes

Speech sounds as idealisations...

- the minimal units of **discrimination** between words; defined in terms of their **distribution**
- **phonological constraints** determine possible combinations
  - language (even dialect) dependent
- described in terms of manner and place of production
- Useful in speech technology because all English words can be represented using around 44 phonemes
- **BUT** they are idealisations - not real objects

Phonemes and phones

Phoneme

- The smallest meaningful contrastive unit in the phonology of a language
- Each language uses small set of phonemes, much smaller than the number of sounds than can be produced by a human
- The number of phonemes varies per language, with most languages having 20-40 phonemes
  - General American has ~40 phonemes (24 consonants, 16 vowels)
  - The Rotokas language (Paupa New Guinea) has ~11 phonemes
  - The Taa language (Botswana) has ~112 phonemes

Phonetic notation

- International Phonetic Alphabet (IPA): consists of about 75 consonants, 25 vowels, and 50 diacritics (to modify base phones)
- TIMIT corpus: uses 61 phones, represented as ASCII characters for machine readability
  - TIMIT only covers English, whereas IPA covers most languages

Phones in the TIMIT Database

TIMIT	IPA	Example	TIMIT	IPA	Example
pcl	p˺	(p closure)	bcl	b˺	(b closure)
tcl	t˺	(t closure)	dcl	d˺	(d closure)
kcl	k˺	(k closure)	gcl	g˺	(g closure)
p	p	**p**ea	b	b	**b**ee
t	t	**t**ea	d	d	**d**ay
k	k	**k**ey	g	g	**g**ay
q	ʔ	bat	dx	ɾ	dir**t**y
ch	t͡ʃ	**ch**oke	jh	d͡ʒ	**j**oke
f	f	**f**ish	v	v	**v**ote
th	θ	**th**in	dh	ð	**th**en
s	s	**s**ound	z	z	**z**oo
sh	ʃ	**sh**out	zh	ʒ	a**z**ure
m	m	**m**oon	n	n	**n**oon
em	m̩	bottom	en	n̩	button
ng	ŋ	si**ng**	eng	ŋ̩	Washington
nx	ɾ̃	wi**n**ner	el	l̩	bott**le**
l	l	**l**ike	r	r	**r**ight
w	w	**w**ire	y	j	**y**es
hh	h	**h**ay	hv	ɦ	**a**head
er	ɝ	b**ir**d	axr	ɚ	butt**er**
iy	i	b**ee**t	ih	ɪ	b**i**t
ey	e	b**ai**t	eh	ɛ	b**e**t
ae	æ	b**a**t	aa	ɑ	f**a**ther
ao	ɔ	b**ough**t	ah	ʌ	b**u**t
ow	o	b**oa**t	uh	ʊ	b**oo**k
uw	u	b**oo**t	ux	ü	t**oo**t
aw	ɑʷ	**ou**t	ay	ɑʸ	b**i**te
oy	ɔʸ	b**oy**	ax-h	ə̥	s**u**spect
ax	ə	**a**bout	ix	ɨ	deb**i**t
epi		(epenthetic sil.)	pau		(pause)
h#		(silence)			

[Gold & Morgan, 2000]

The English Phonemes

IPA	SAMPA	Example	IPA	SAMPA	Example
i:	i:	heed	p	p	pea
I	I	hid	t	t	tea
e	e	head	k	k	key
Q	{	had	b	b	bee
A:	A:	hard	d	d	dog
ˇ	V	hut	g	g	good
Å	Q	hod	f	f	fat
ç:	O:	hoard	T	T	thin
U	U	hood	s	s	sip
u:	u:	who'd	S	S	ship
´	@	about	v	v	vat
Œ:	3:	heard	D	D	that
eI	eI	hay	z	z	zip
aI	aI	high	Z	Z	measure
çI	OI	boy	tS	tS	chin
´U	@U	hoe	dZ	dZ	gin
aU	aU	how	m	m	map
I´	I@	here	n	n	nap
e´	e@	there	N	N	hang
U´	U@	moor	l	l	led
w	w	wet	r	r	red
j	j	yet	h	h	hit

Advanced Research Project Agency-ARPAbet

- *Uses ASCII characters:*

  http://en.wikipedia.org/wiki/Arpabet

- *CMU - dictionary*

  http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Phone

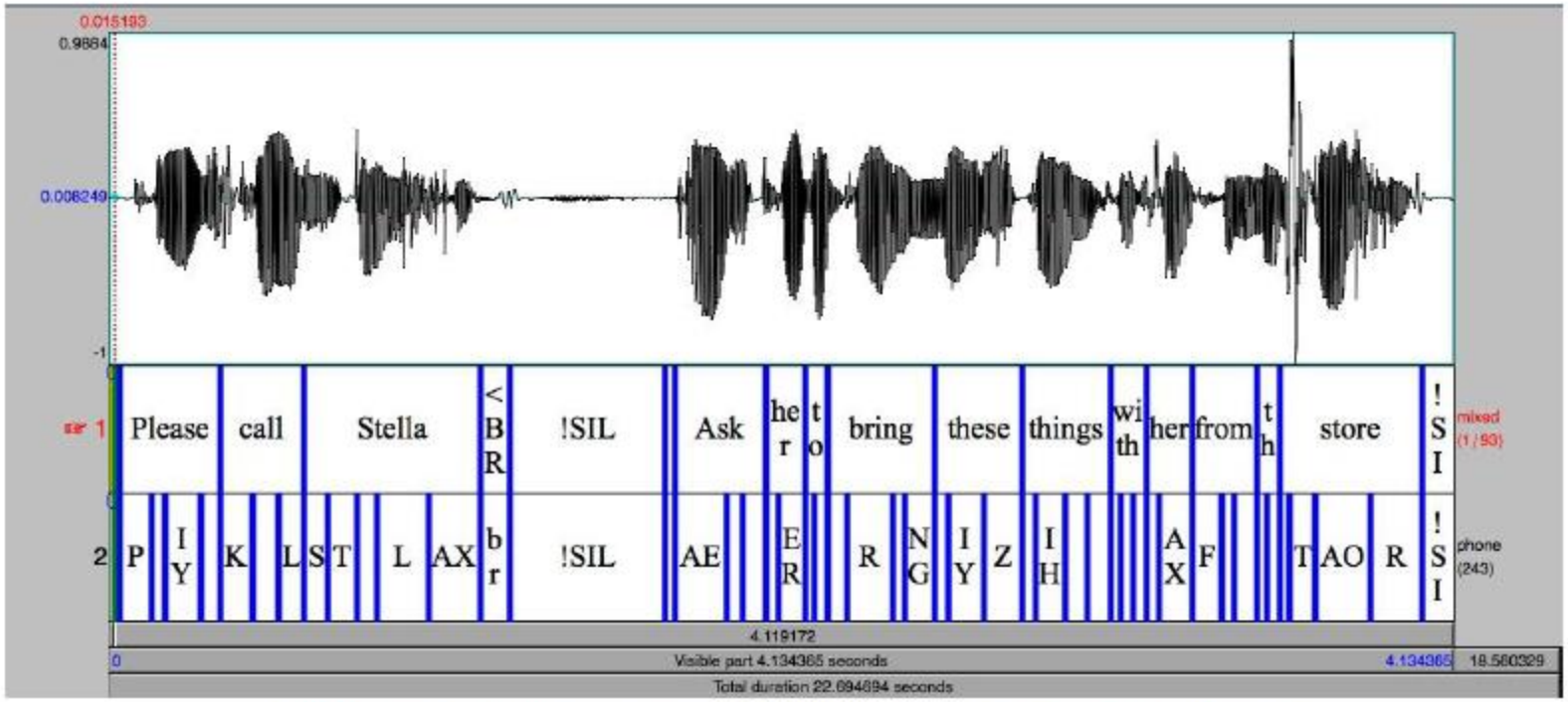
- The physical sound produced when a phoneme is articulated
- Since the vocal tract can have an infinite number of configurations, there is an infinite number of phones that correspond to a phoneme

Allophone

- A class of phones corresponding to a specific variant of a phoneme
  - Example: aspirated [p^h] and unaspirated [p] in the words *pit* and spit
  - Example: /t/ sounds in the words *tub*, *stub*, *but*, *butter*

Coarticulation

- The phenomenon whereby the articulatory configuration of a phoneme is affected by that of its neighboring phonemes
  - As a result, crisp boundaries between phonemes are hard to define
- Coarticulation is involved in the transformation of phonemes into allophones



http://groups.linguistics.northwestern.edu/documentation/images/praat_aligned.jpg

Branches of phonetics

Phonology vs. phonetics

- Phonology is concerned with the distribution and patterns of speech sounds in a particular language, or in languages in general
- Phonetics is concerned with the study of speech sounds and their production, classification, and transcription
  - In a nutshell, phonetics deals with the physical nature of speech sounds, and not with their relations to other speech sounds in particular languages

Three basic approaches to the study of phonetics

- Articulatory phonetics is concerned with the position, shape and movements of speech articulators
- Acoustic phonetics is concerned with the spectro-temporal properties of the speech sound waves
- Auditory phonetics is concerned with the perception, categorization, and recognition of speech sounds and the role of the auditory system

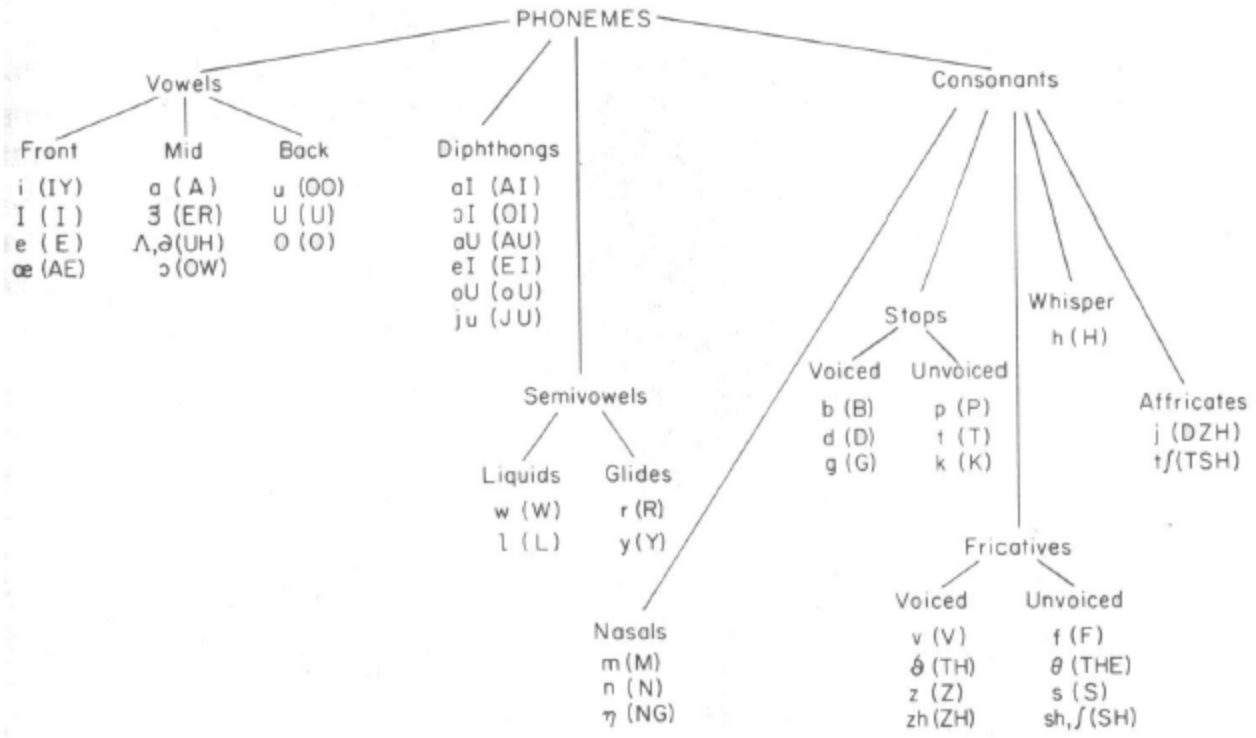
The Sounds of Language

- The International Phonetic Association (IPA) classifies speech sounds according to:
  - where the **air stream** comes from and whether it is going in or out
  - whether the **vocal folds** are vibrating
  - whether the **soft palate (velum)** is raised or lowered (nasals)
  - how the sound is made - **manner of articulation**
  - which bits of the vocal tract are involved - **place of articulation**
  - shape of the **lips** (rounded, spread)

General organization of sounds

Four general classes of sounds in American English

- Vowels, diphthongs, semivowels, and consonants
- Each can be further divided according to articulators (manner, place)

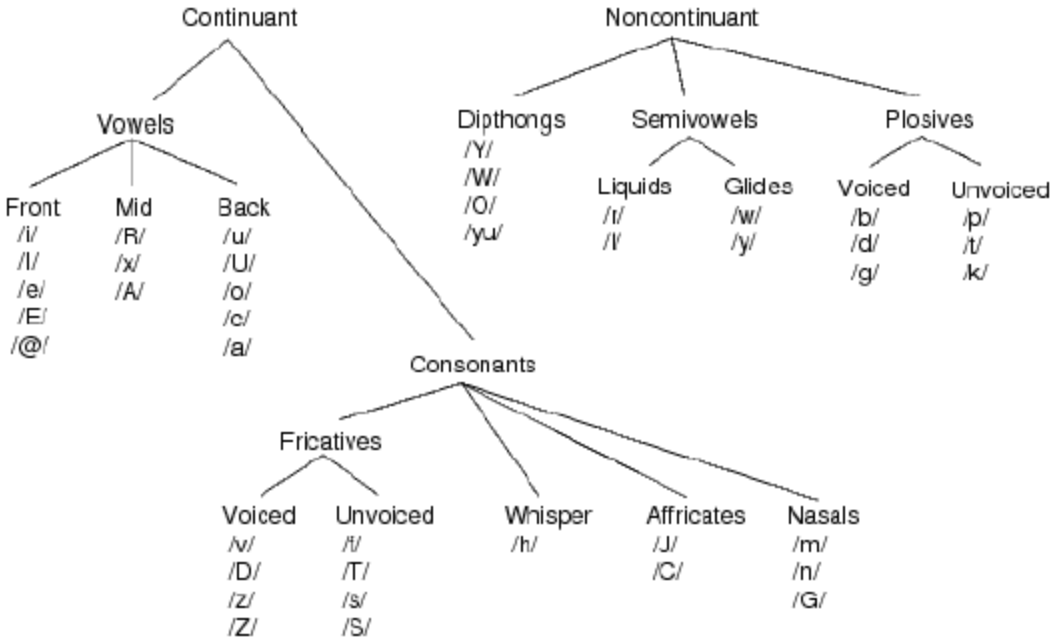


[Rabiner & Schafer, 1978]

http://www.uiowa.edu/~acadtech/phonetics/english/frameset-ad3.html

Alternatively, phoneme classes can be divided into

- Continuants: produced by a fixed vocal tract configuration
  - Includes vowels, fricatives, and nasals
- Non-continuants: vocal tract configuration changes over time
  - Diphthongs, semivowels, stops and affricatives



http://cnx.org/content/m18086/latest/phoneme.png

Articulatory phonetics (consonants)

In terms of articulators, consonants can be described by

- Place of articulation: Defines the place of contact between an active articulator (i.e. tongue) and a passive articulator (i.e. palate)
- Manner of articulation: Concerned with airflow, the path it takes and the degree to which it is impeded
- Voicing: Determined by the behavior of the vocal folds (vibrating vs. open)

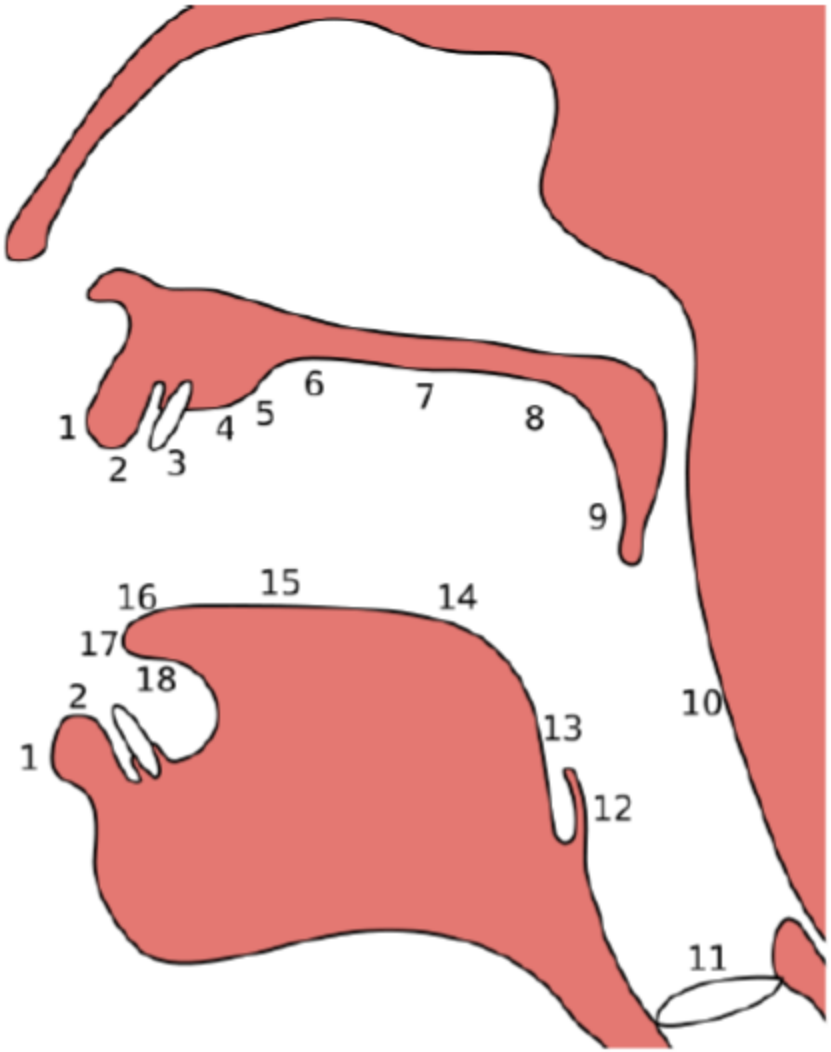
	MANNER		VOICING	PLACE						
				Bilabial	Labiodental	Interdental	Alveolar	Palatal	Velar	Glottal
OBSTRUENTS	Stop		Voiceless	p			t		k	ʔ
			Voiced	b			d		g	
	Fricative		Voiceless		f	θ	s	ʃ		h
			Voiced		v	ð	z	ʒ		
	Affricate		Voiceless					ʧ		
			Voiced					ʤ		
SONORANTS	Nasal		Voiced	m			n		ŋ	
	LIQUID	Lateral	Voiced				l			
		Rhotic	Voiced					r		
	Glide		Voiced	w				j	w	

http://www.speech-language-therapy.com/pvm.pdf

Place of articulation

- Bilabial: constriction at the lips: [b], [m]
- Labiodental: Lower lips against upper teeth: [f], [v]
- Interdental: constriction between the teeth: [th] *th*ing, [dh] *th*at
- Alveolar: constriction is at the alveolar ridge: [t], [n], [z]
- Palatal-alveolar: constriction slightly behind alveolar ridge: [sh] *sh*erry, [zh] measure
- Palatal: constriction in the hard palate: [jh] *j*oke
- Velar: constriction closer to the soft palate: [k], [ng] si*ng*
- Labiovelar: constriction both at lips and velum: [w]
- Glottal: when closure occurs as far back as the glottis: glottal stop [], as in the negative utterance uh-uh
- Uvular: constriction in the uvula; none in English, French /r/ in *r*ouge

http://www.uiowa.edu/~acadtech/phonetics/#



Places of articulation

1. Exo-labial
2. Endo-labial
3. Dental
4. Alveolar
5. Post-alveolar
6. Pre-palatal
7. Palatal
8. Velar
9. Uvular
10. Pharyngeal
11. Glottal
12. Epiglottal
13. Radical
14. Postero-dorsal
15. Antero-dorsal
16. Laminal
17. Apical
18. Sub-apical

http://en.wikipedia.org/wiki/Place_of_articulation

Manner of articulation

- Stops: produced by complete stoppage of the airstream: [p]
- Fricatives: tongue comes very close to a full closure: [f], [sh] *sh*erry
- Affricatives: combination b/w stops and fricatives: *ch*erry
- Nasals: closed oral passage (as in stops), open nasal cavity: [n], [ng] si*ng*
- Approximants : halfway between consonants and vowels
  - Liquids: [l], [r]
  - Glides: [y], [w]

Voicing

- When the vocal folds vibrate, it is voiced; otherwise it is voiceless
- Examples of voiceless/voiced: *S*ue vs. *z*oo, *p*at vs. *b*at

http://www.uiowa.edu/~acadtech/phonetics/#

Vowels

- The largest phoneme group and most interesting one
  - Carry little information in written speech, but most ASR systems rely heavily on them for performance
- Vowels are voiced (except when whispered) and have the greatest intensity and duration in the range of 50 to 400ms
- Vowels are distinguished mainly by their first three formants
  - However, there is a significant individual variability, so other cues can be employed for discrimination (upper formants, bandwidths)

Articulatory phonetics (vowels)

Vowels can be described in a similar way

- Manner of articulation, just considered to be "vowel"
- Place of articulation is generally described with three major parameters: frontness, height, and roundness

Frontness (or backness)

- Provides a general indication of the greatest place of constriction, and correlates with F2
- Three positions in English
  - Front: [iy] b*ea*t, [ih] b*i*t, [eh] b*e*t, [ae] b*a*t
  - Central: "schwa" [ax] *a*bout
  - Back: [uw] b*oo*t, [ao] b*ough*t, [ah] b*u*t, [aa] f*a*ther

http://www.utexas.edu/courses/linguistics/resources/phonetics/vowelmap/index.html

Height

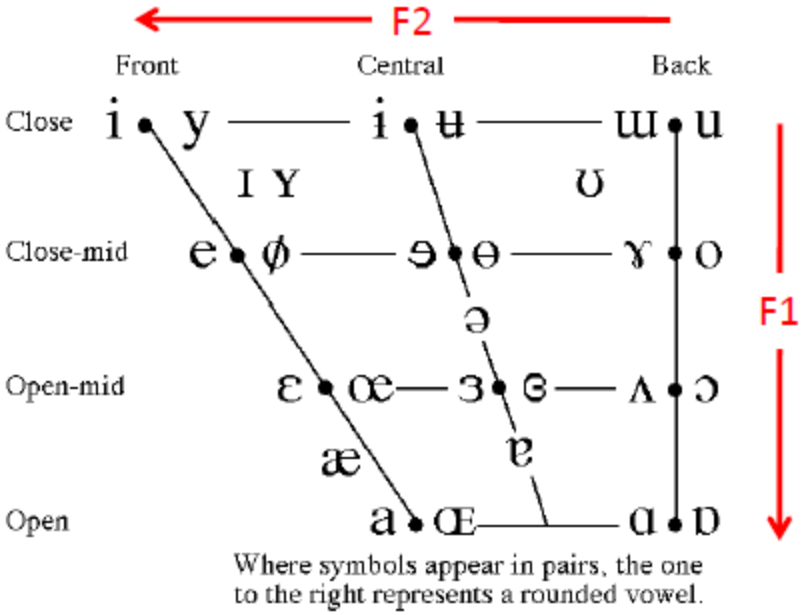
- Refers to how far lower jaw is from upper jaw when making the vowel
  - High vowels have lower and upper jaw close: [iy], [uw]
  - Low vowels have a more open oral cavity: [ae], [aa]
- Correlates with F1 (high vowel: low F1; low vowel: high F1)

Roundness

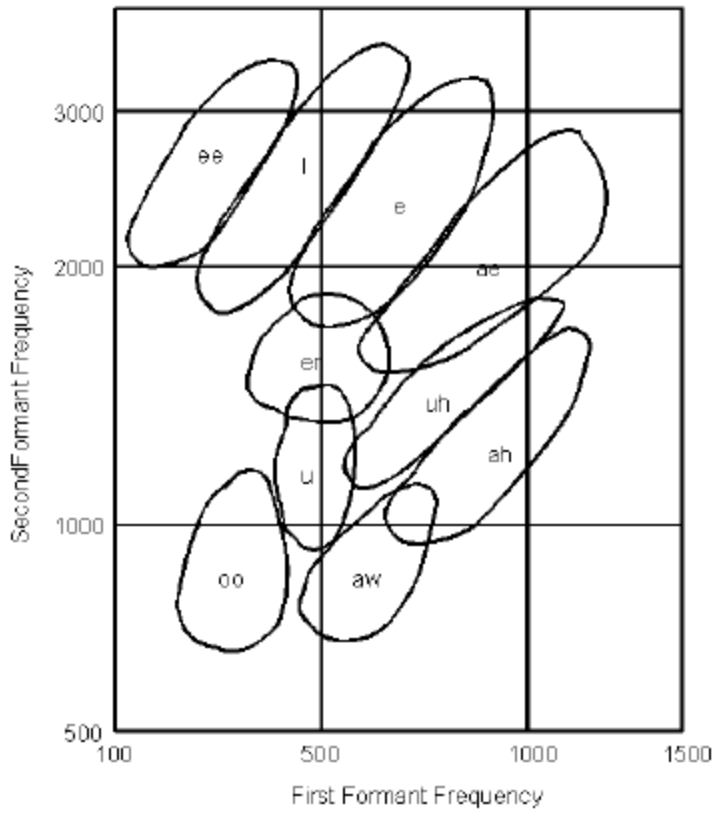
- Refers to whether the lips have been rounded as opposed to spread
- In English, front vowels are unrounded whereas back vowels are rounded: b*i*t vs. b*oo*t

/ARPABET, IPA/

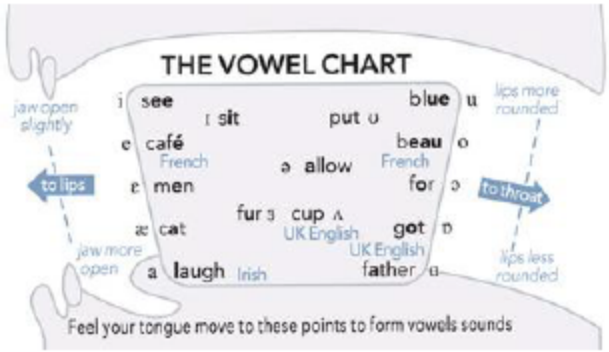
/iy, i/ f*ee*l, el*i*te, /ih,ɪ/ f*i*ll, /ae, æ/ g*a*s, /aa, ɑ/ f*a*ther,
/ah, ʌ/ c*u*t, /ao, ɑ/ d*o*g, /ax, ə/ c*o*mply, /eh, ɛ/ p*e*t,
/er, ɝ/ t*ur*n, /uh, ʊ/ g*oo*d, /uw, u/ t*oo*l



http://www.singwise.com/images/CardinalVowelChart.gif



http://www.geofex.com/Article_Folders/sing-wah/sing-wah.htm



http://www.thedialectcoach.com/images/content/vowel_chart.jpg

Acoustic phonetics

Acoustic phonetics is concerned with

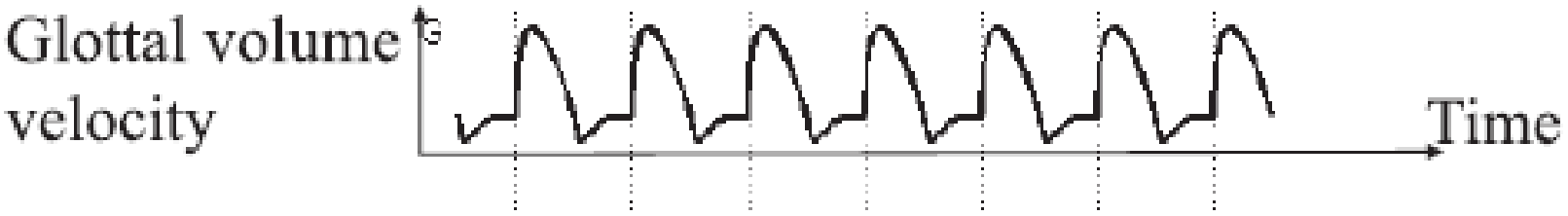
- Time domain waveform of the speech signal, and
- Its time-varying spectral characteristics

Visualizations of speech waveforms

- Time-domain waveforms are rarely studied directly
  - This is because phase differences can significantly affect its shape but are rarely relevant for speech perception
- Instead, frequency-domain signals are commonly used
- The spectrum (log-magnitude) of a voiced phone shows two types of information
  - A comb-like structure, which represent the harmonics of F0 (the source),
  - A broader envelope, which represents the resonances (formants) of the vocal tract filter
- Various techniques exist to separate the two sources of information
  - Linear prediction, homomorphic (cepstral) analysis ...

"Voiced" Speech

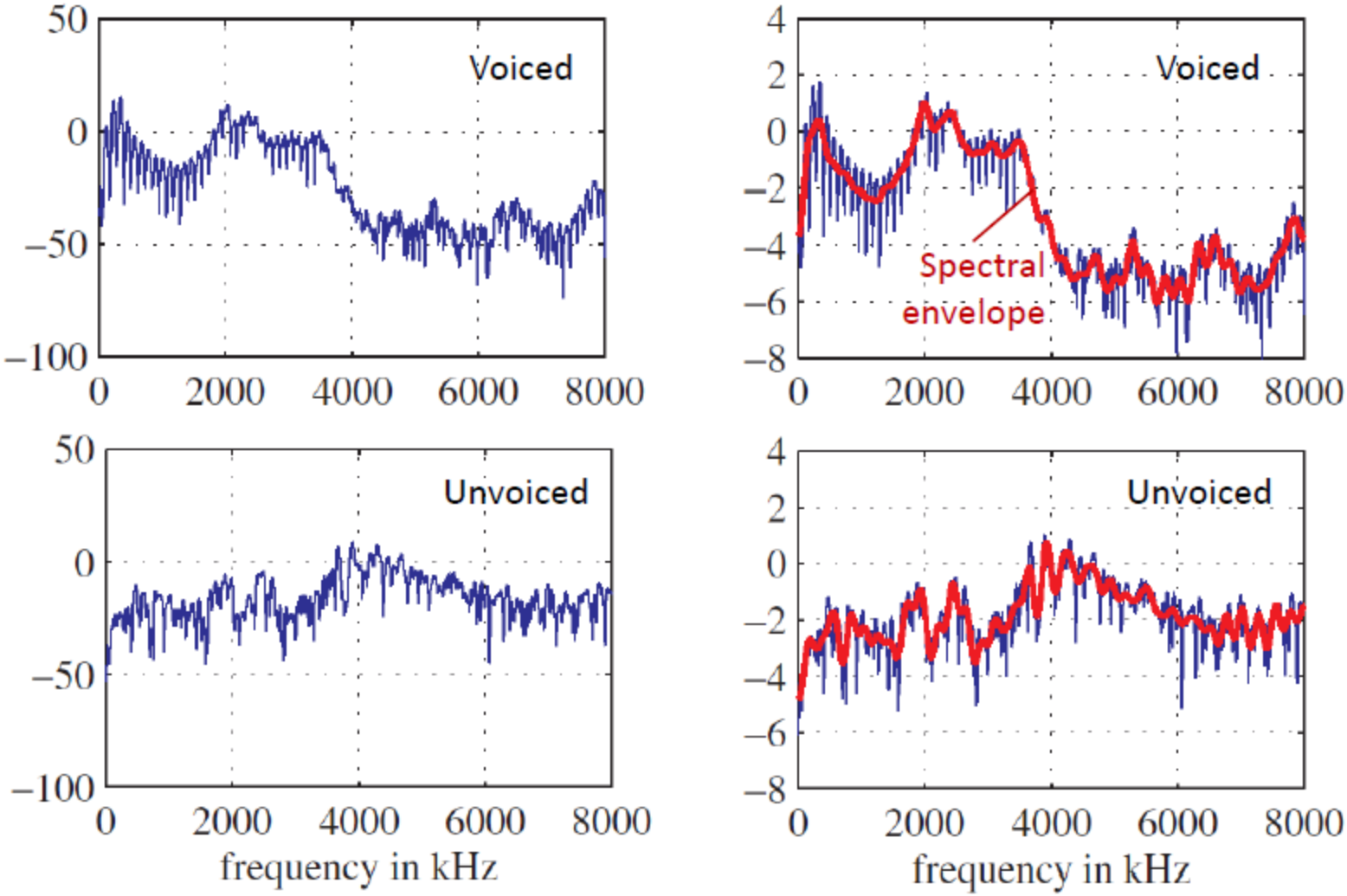
- Voiced speech occurs when air flows through the vocal chords into the vocal tract in discrete "puffs" rather than as a continuous flow



- The vocal chords vibrate at a particular frequency, which is called the fundamental frequency of the sound
  - 50 : 200 Hz for male speakers
  - 150:300 Hz for female speakers
  - 200:400 Hz child speakers

“Unvoiced” Speech

- For unvoiced speech, the vocal chords are held open and air flows continuously through them
- The vocal tract, however, is narrowed resulting in a turbulent flow of air along the tract
- Examples include the unvoiced fricatives /f/ & /s/
- Characterised by high frequency components


Voiced
Unvoiced
Spectral
envelope
frequency in kHz
frequency in kHz

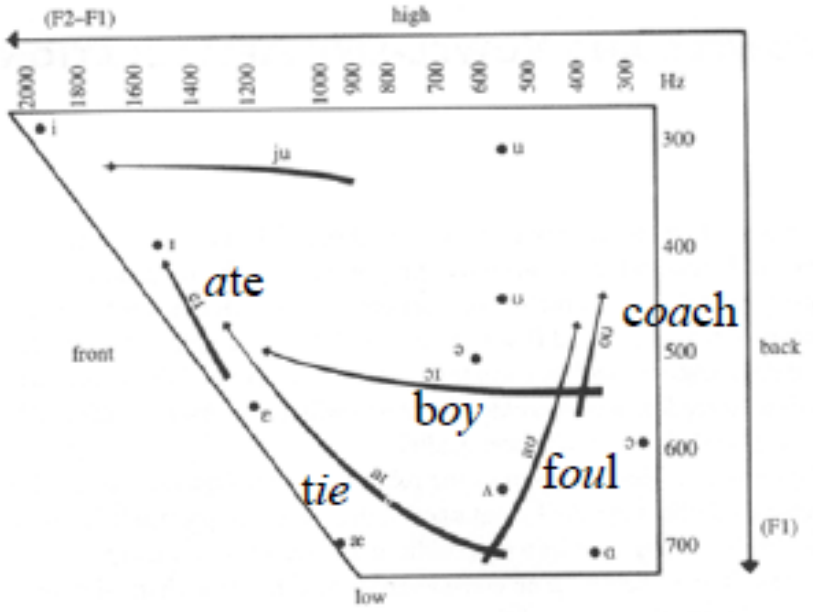
Diphthongs (vowels)

/ARPABET, IPA/

/ay, aɪ/ t*ie*, /ey, eɪ/ *a*te, /oy, ɔɪ/ c*oi*n, /aw, aʊ/, f*ou*l, /ow, oʊ/ c*oa*ch, /ow, oʊ/ t*o*ne

- Articulators start to form one vowel & move into another:

diphthong	from	to
/ay/ t*ie*	/aa/ f*a*ther	/iy/ *e*ve
/ey/ *a*te	/eh/ t*e*n	/iy/ *e*ve
/oy/ c*oi*n	/ao/ d*o*g	/iy/ *e*ve
/aw/ f*ou*l	/aa/ f*a*ther	/uw/ t*oo*l
/ow/ c*oa*ch		



Ladefoged, 2001, p. 200

Manner of Articulation: Plosives

- produced by the abrupt **release** of a **constriction** somewhere along the length of the vocal tract
- English has plosives at the following places
  - bilabial (lips): **p b** (pin, bin);
  - alveolar (alveolar or teeth ridge): **t d** (tin, din);
  - velar (soft palate or velum): **k g** (kin, good)
- characterised by a short period of no (or little) energy, followed by sudden 'explosion' and aspiration when constriction is released
- characteristics vary depending on context

Manner of Articulation: other stops

- **Trill** involving repeated **vibration** of one articulator against another: e.g. r (right)
- **Tap or Flap** like a trill, but with a single touch: e.g. ɾ (arrow)
- **Nasal**
  - vocal tract is **constricted**
  - the soft palate (velum) is lowered so air escapes through the **nose**
  - characterised by vowel-like structure, but with weaker energy
  - identity cued by transitions from surrounding sounds
  - English has only voiced nasals:
    - bilabial nasal m (map);
    - alveolar nasal n (nap);
    - velar nasal N (sing);

Manner of Articulation: Fricatives

- airstream is forced through a constriction, causing **turbulence**
- characterised by non-periodic sound (**friction**) – random energy across wide frequency range
- English has both voiced and voiceless fricatives
  - labio-dental fricatives: **f v** (fine, vine);
  - inter-dental fricatives: **T D** (thin, these);
  - alveolar fricatives: **s z** (sun, zoo);
  - palatal-alveolar **S Z** (shine, pleasure);
  - glottal fricative **h** (hat)

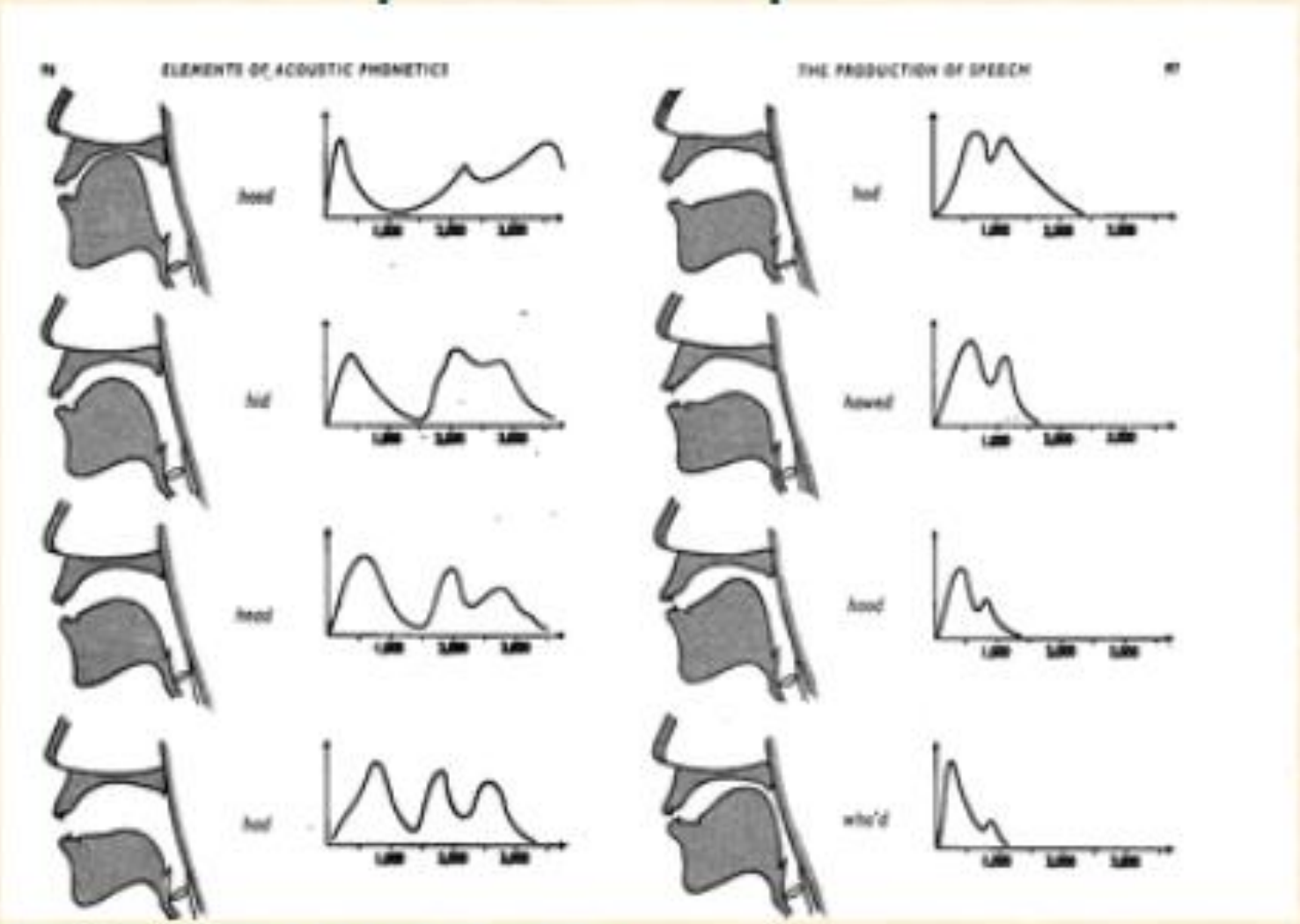
Manner of Articulation: Approximants

- similar to fricatives, but less obstruction, so no friction
  - have weaker vowel-like structure, but function as consonants
- **liquids** English has only one
  - lateral **l** (light)
  - some variants of **r** are approximant
- **glides** are transient sounds
  - palatal **j** (young)
  - labial **w** (wind)

Manner of Articulation: Vowels

- **Vowel** - made without any constriction in the vocal tract
  - the shape of the vocal tract enhances some harmonics of the fundamental, while suppressing (damping) others
- regions of enhanced harmonics are called **formants**
- formants are related to position of tongue and lips
  - first two formants are important for vowel discrimination

Mouth Shape and Spectra



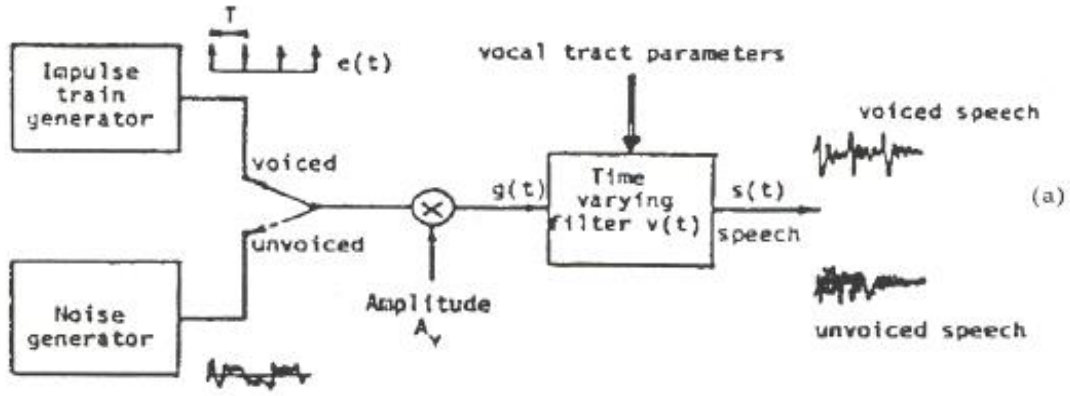
Taken from 'Elements of Acoustic Phonetics', P. Ladefoged, The University of Chicago Press, 1962.

Resonant Frequencies of Vocal Tract

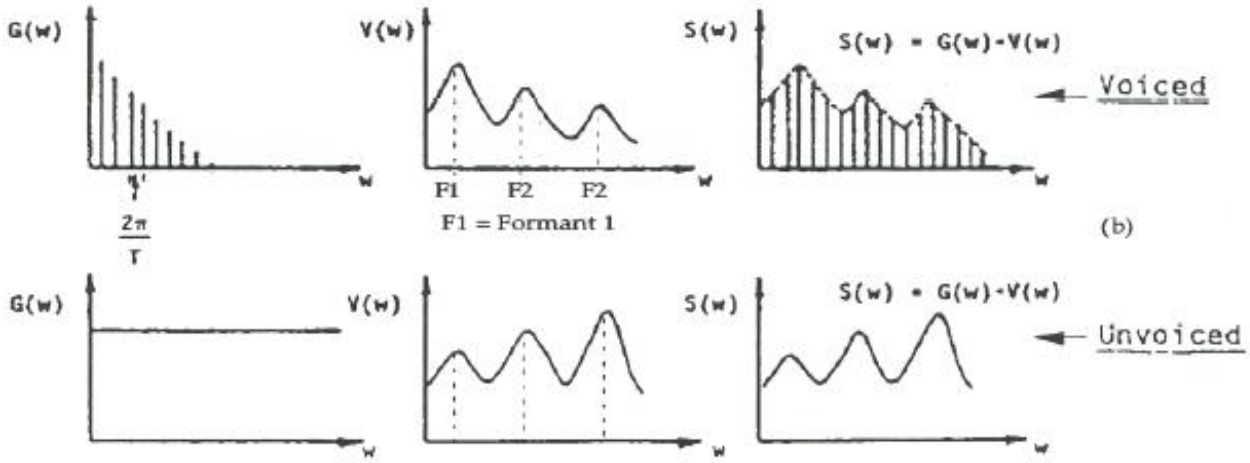
- Vocal Tract is a non-uniform acoustic tube that is terminated at one end by the vocal chords and at the other end by the lips
- The Cross-sectional area of the vocal tract determined by the positions of the tongue, lips, jaw and velum.depends on lips, tongue, jaw and velum
- The spectrum of vocal tract response consists of a number of resonant frequencies of the vocal tract.
- These frequencies are called Formants
- Three to four formants present below 4kHz of speech

Formant Frequencies

- Speech normally exhibits one formant frequency in every 1 kHz
- For VOICED speech, the magnitude of the lower formant frequencies is successively larger than the magnitude of the higher formant frequencies
- For UNVOICED speech, the magnitude of the higher formant frequencies is successively larger than the magnitude of the lower formant frequencies



Time model for speech production



Speech spectrum for voiced and unvoiced

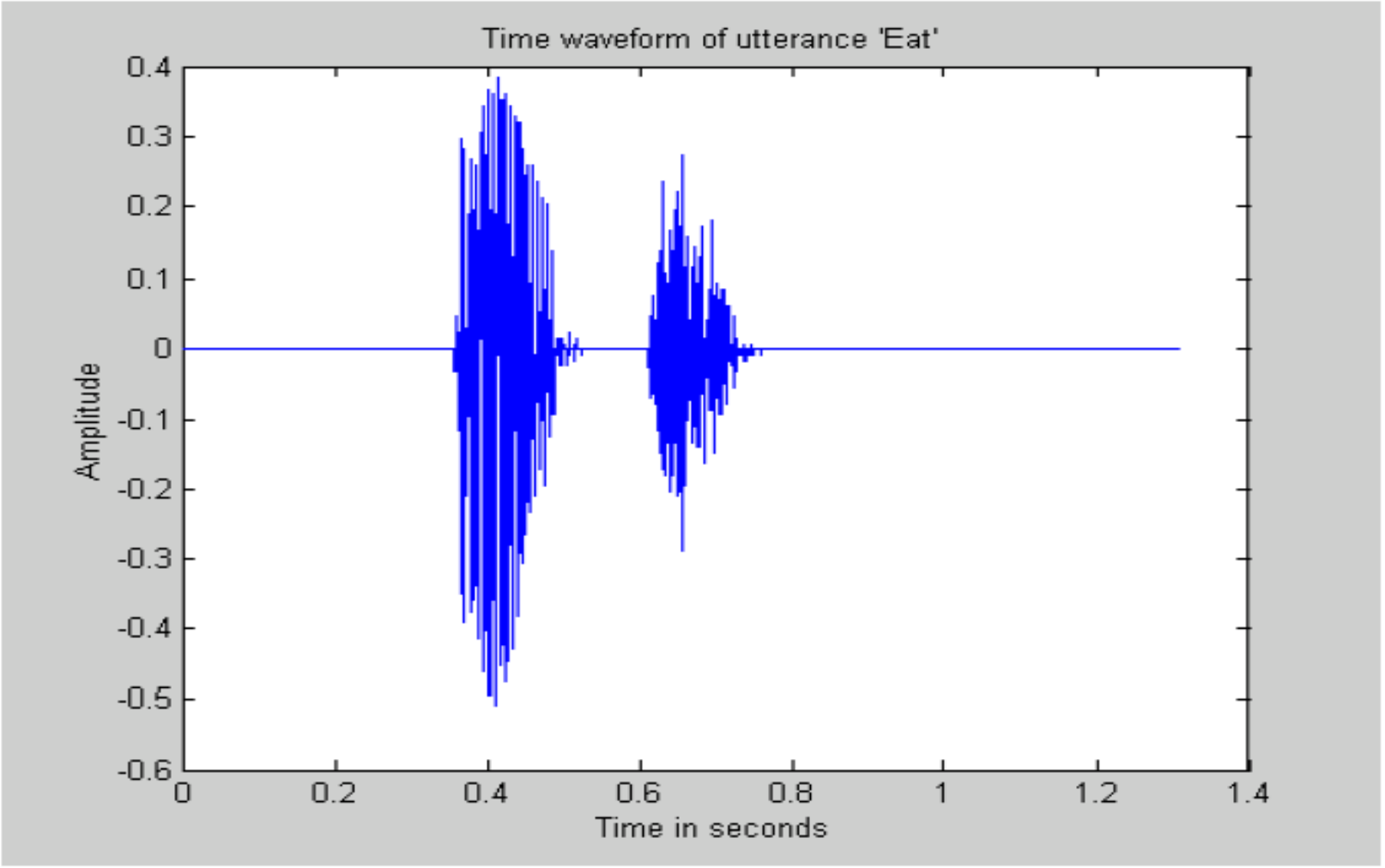
Speech

Figure 1.3: Time model for speech production (b) Speech spectrum

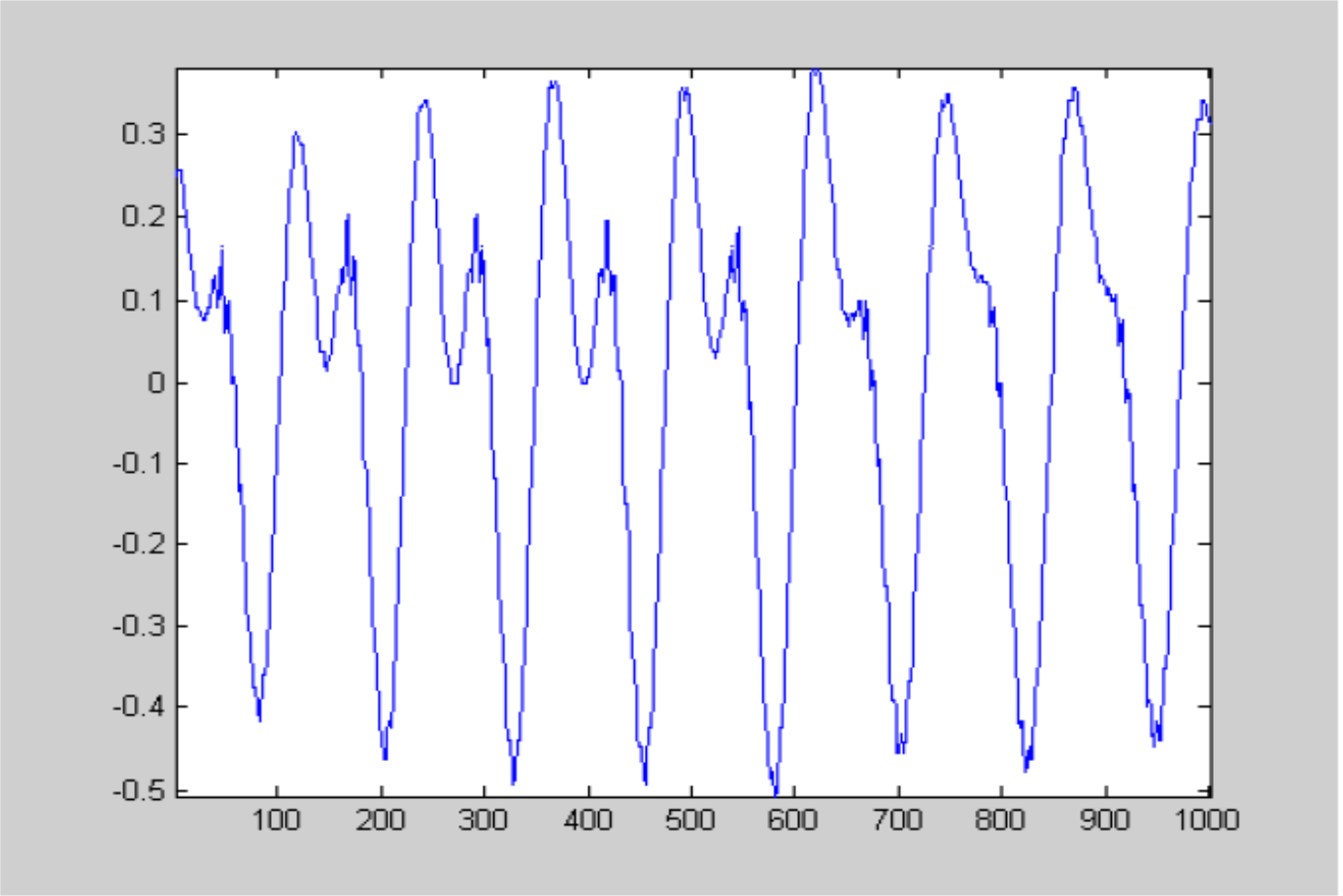
Basic Assumptions of Speech Processing

- The basic assumption of almost all speech processing systems is that the source of excitation and the vocal tract system are independent.
- Therefore, it is a reasonable approximation to model the source of excitation and the vocal tract system separately as shown (Figure 1.3)
- The vocal tract changes shape rather slowly in continuous speech and it is reasonable to assume that the vocal tract has a fixed characteristics over a time interval of the order of 10 ms.
- Thus once every 10 ms, on average, the vocal tract configuration is varied producing new vocal tract parameters (resonant frequencies)

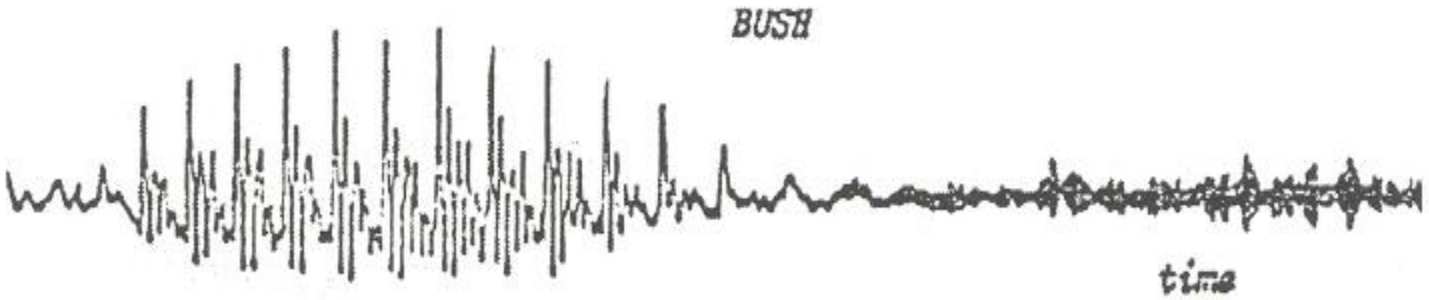
Acoustic Waveforms



Frame of waveform



Example: The acoustic waveform of the word 'Bush'



Amplitude/Time plot of 'Bush'

The speech signal is a slowly time varying signal in the sense that when examined over sufficiently short period of time, its characteristics are fairly stationary.

Phonetics: Variability

Speech is inherently **variable**:

- **Inter-speaker** variation: physical, age, gender, accent
- **Intra-speaker** variation: health, mood, external factors
- **Inherent** variation: rate of speaking, loudness
- **Phonological** (Contextual) variation
- **Style of speaking**: formal, casual, read, spontaneous
- **'Random'** variation

Phonetics: Context

- **Co-articulation** affects how each sound is realized in context
  - /{/ realized differently in ***cab*** and ***cat*** and ***can***
  - /r/ sound in ***train*** different to that in ***arrow***
  - /p/ may be different in each of ***pin, spin,*** and ***apt***
  - /l/ in ***leap*** different to that in ***milk***

  - ***can be --> cam be*** **(assimilation)**
- Variants of a phoneme which are caused by contextual influences and are not contrastive are called **allophones**

Sounds in Fluent Speech

- **Reduction -** 'target' positions may not be reached
  - vowels tend to be **neutralised (centralised)**
  - consonants may be not precisely articulated
- **Elision -** sounds get missed out
  - unstressed vowels disappear
  - consonant clusters may be simplified
- **Epenthesis** - sounds may be inserted
- Depends on speaking style (and rate)

Beyond the Phoneme ...

- **homophones**
  - to, too, two
  - glasses (aids to vision or drinking vessels?)
- **ambiguity of segmentation**
  - grey tape vs. great ape
  - this new display will recognise speech vs. this nudist play will wreck a nice beach
- **intonation changes meaning of utterance**
  - He's gone. vs. He's gone?
  - What's that? vs. What's that!
  - In some languages intonation changes the meaning of a word

Higher-level ambiguity

- **Lexical:** words can have more than one meaning
  - E.g. kind, plant, set
- **Syntactic**: words can have different functions, so it is not always possible to determine a unique interpretation:
  - Have the students who missed the exam **taken it today?** (Q)
  - Have the students who missed the exam **take it today** (C)
  - Is the liquid spreading across the floor?
  - Is the liquid spreading across the floor blood?
  - Visiting relatives can be a nuisance
  - He fired his secretary with enthusiasm

External Factors

- **Noise**
  - Lombard Effect
- **Vibration**
  - vibrations in the chest, oral and nasal cavity may cause interference in speech signal
- **Fatigue**
  - speaking rate may decrease, loss of control may result in slurring
- **Fear**
  - speaking rate may increase, pitch may rise due to muscle tightening
- **Cognitive loading**
  - interaction with other tasks, stress
- **Alcohol or drugs**

Prosody

Long-term variations (over more than one phoneme) in

- Pitch (intonation)
- Amplitude (loudness), and
- Timing (articulation rate or rhythm)

Roles of prosody

- Helps highlight the spoken message
  - Alternation of stressed and unstressed syllables identifies the words that the speaker considers more important
- Helps segment the spoken message
  - Provides cues to syntactic boundaries (e.g., main vs. subordinate clauses) and syntactic structure (e.g., declaratives statements vs. questions)
  - Serves as a "continuity guide" to track speakers in noisy environments
- Provides cues to the state of the speaker
  - F0 and amplitude patterns vary with emotions
- **Interestingly, however, prosody is typically ignored in ASR**

English Sounds – Examples with animation

http://www.uiowa.edu/~acadtech/phonetics/english/frameset.html