L3: Organization of speech sounds

Phonemes, phones, and allophones

Taxonomies of phoneme classes

Articulatory phonetics

Acoustic phonetics

Speech perception

Prosody

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	ba**t**	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	botto**m**	en	n̩	butto**n**
ng	ŋ	si**ng**	eng	ŋ̩	Washi**ng**ton
nx	ɾ̃	wi**nn**er	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**h**ead
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**ou**ght	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	ab**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]

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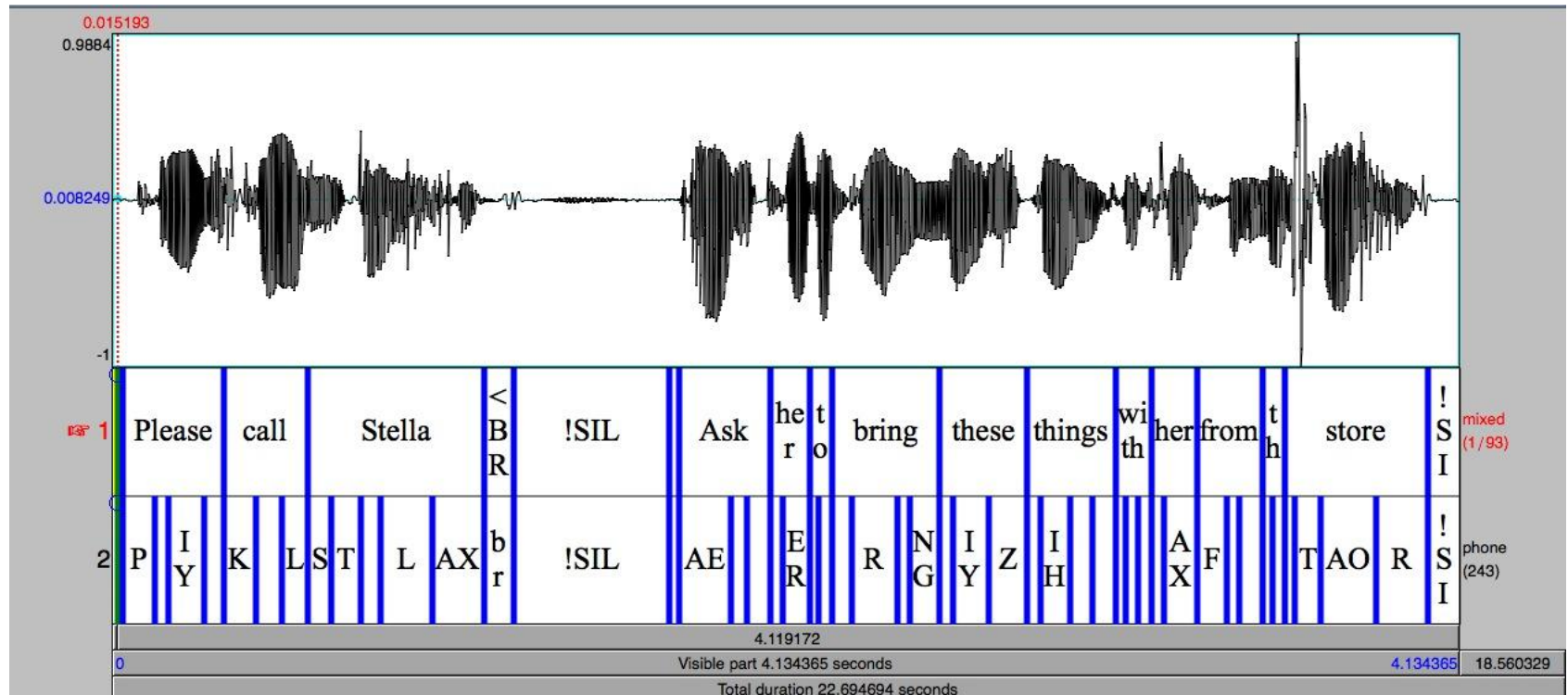
http://groups.linguistics.northwestern.edu/documentation/images/praat_aligned.jpg

Branches of phonetics

Phonology vs. phonetics

- <u>Phonology</u> is concerned with the distribution and patterns of speech sounds in a particular language, or in languages in general
- <u>Phonetics</u> 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

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

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)
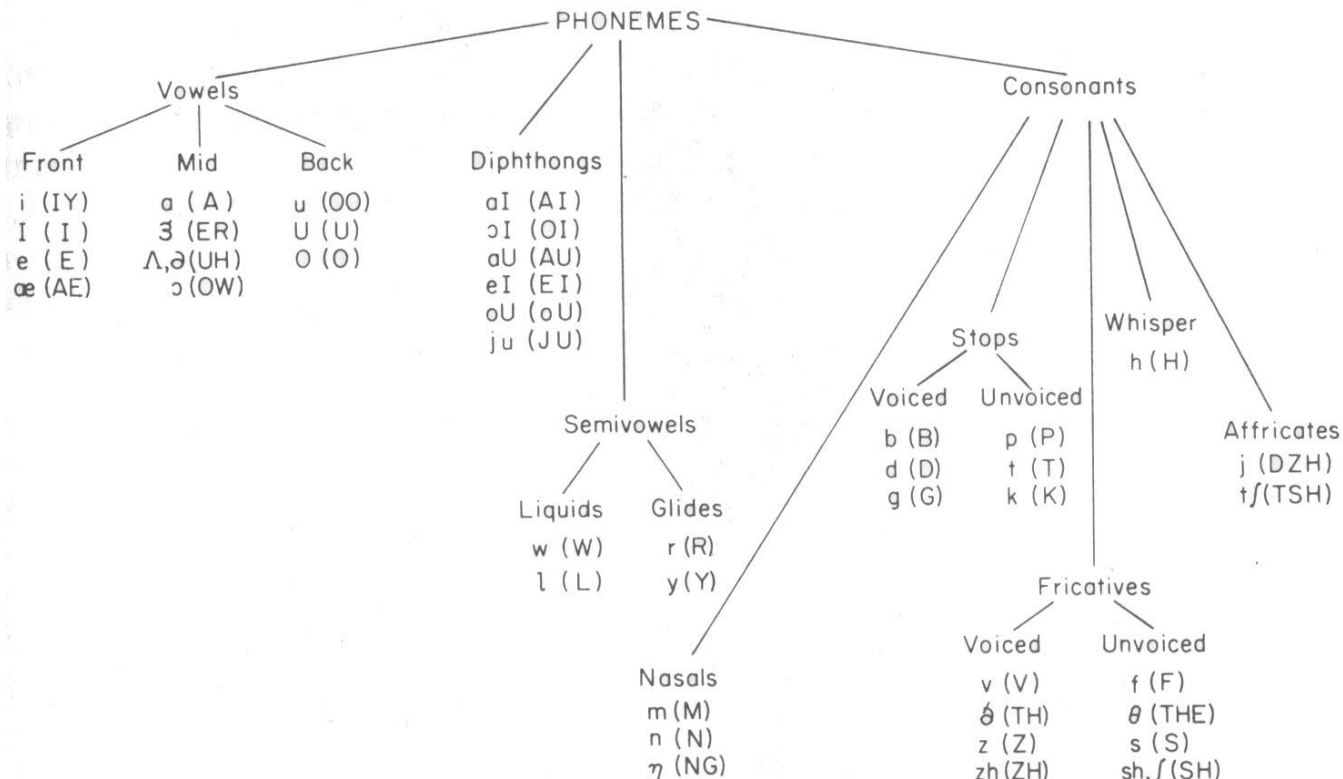
[Rabiner & Schafer, 1978]

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
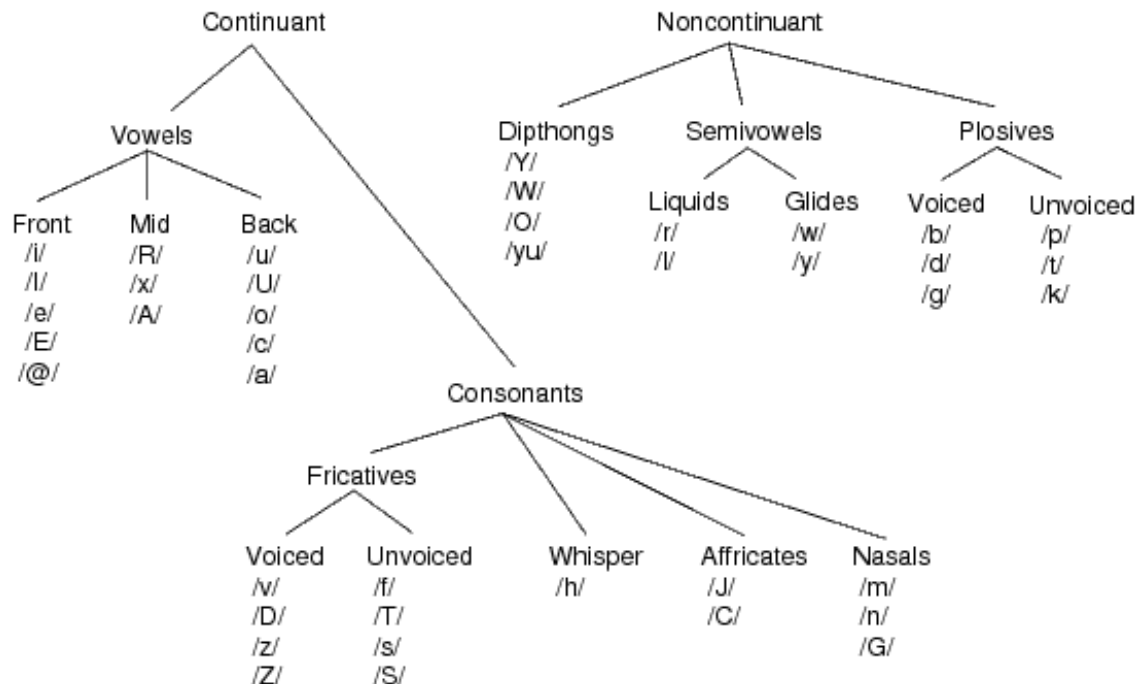
Articulatory phonetics (consonants)

In terms of articulators, consonants can be described by

– <u>Place of articulation</u>: Defines the place of contact between an active articulator (i.e. tongue) and a passive articulator (i.e. palate)

– <u>Manner of articulation</u>: Concerned with airflow, the path it takes and the degree to which it is impeded

– <u>Voicing</u>: 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					tʃ		
		Voiced					ʤ		
SONORANTS	Nasal	Voiced	m			n		ŋ	
	LIQUID Lateral	Voiced				l			
	LIQUID 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/#
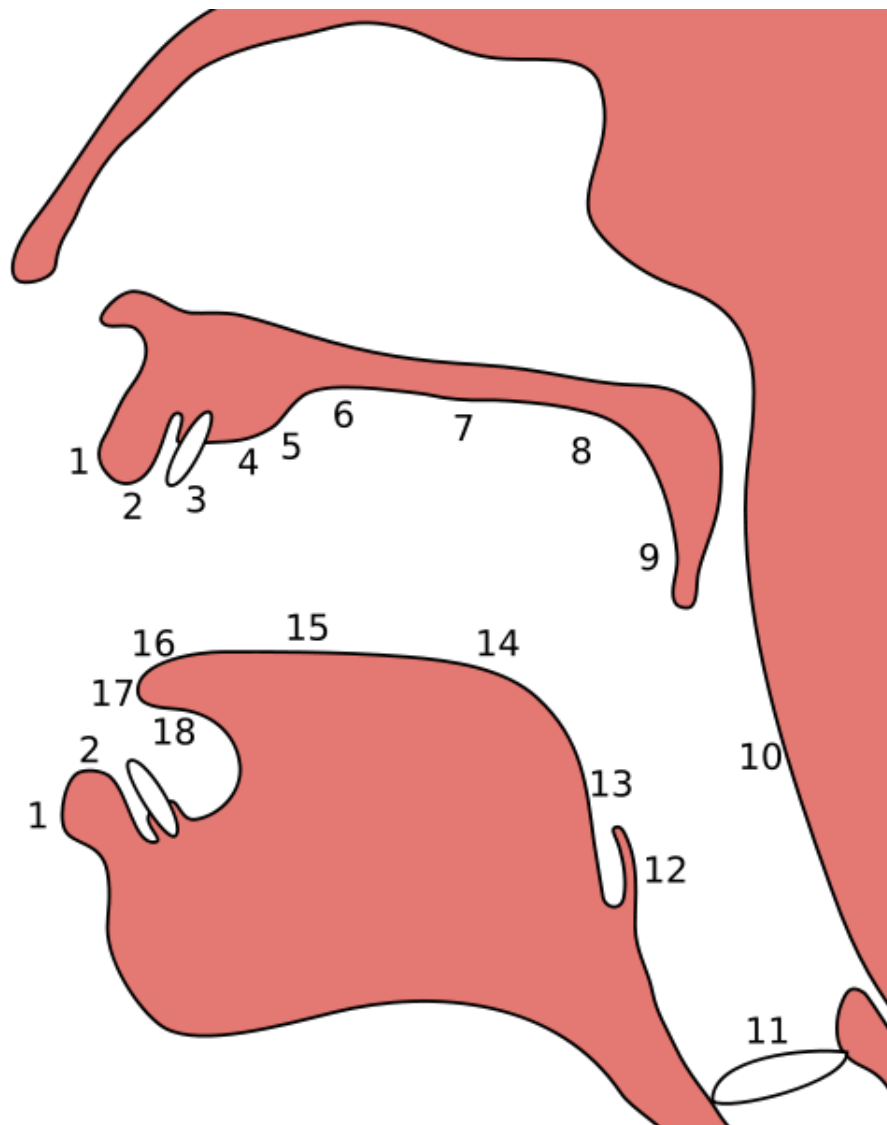
Places of articulation
1. Exo-labial
2. Endo-labial
3. Dental
4. Alveolar
5. Post-alveolar
6. Pre-palatal
7. Palatal
8. Velar
9. Uvular
10. Pharyngeal
11. Glottal
12. Epiglottal
13. Radical
14. Postero-dorsal
15. Antero-dorsal
16. Laminal
17. Apical
18. Sub-apical

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/#

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*ou*ght, [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

F2

Front　　Central　　Back

Close

Close-mid

Open-mid

Open

Where symbols appear in pairs, the one
to the right represents a rounded vowel.

F1

http://www.singwise.com/images/CardinalVowelChart.gif
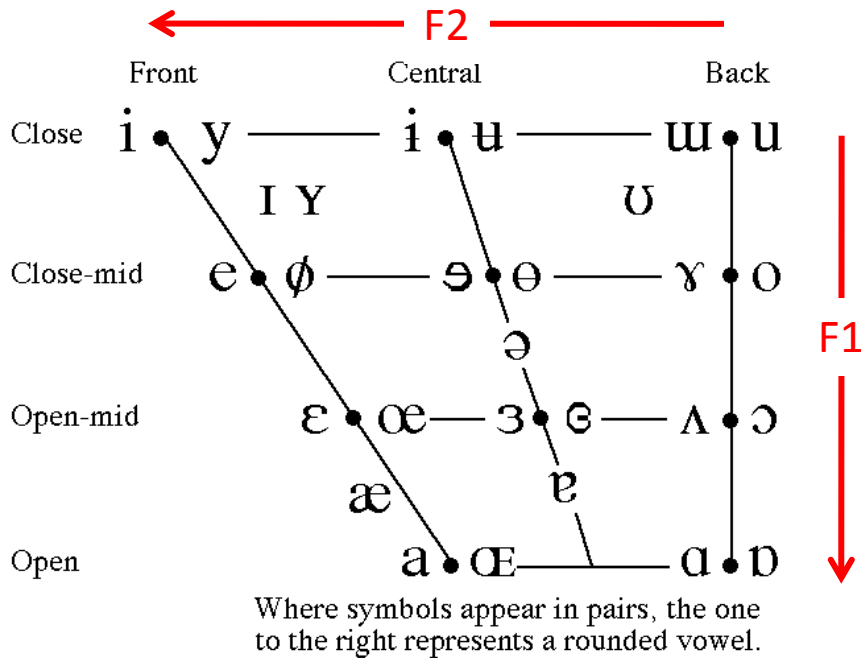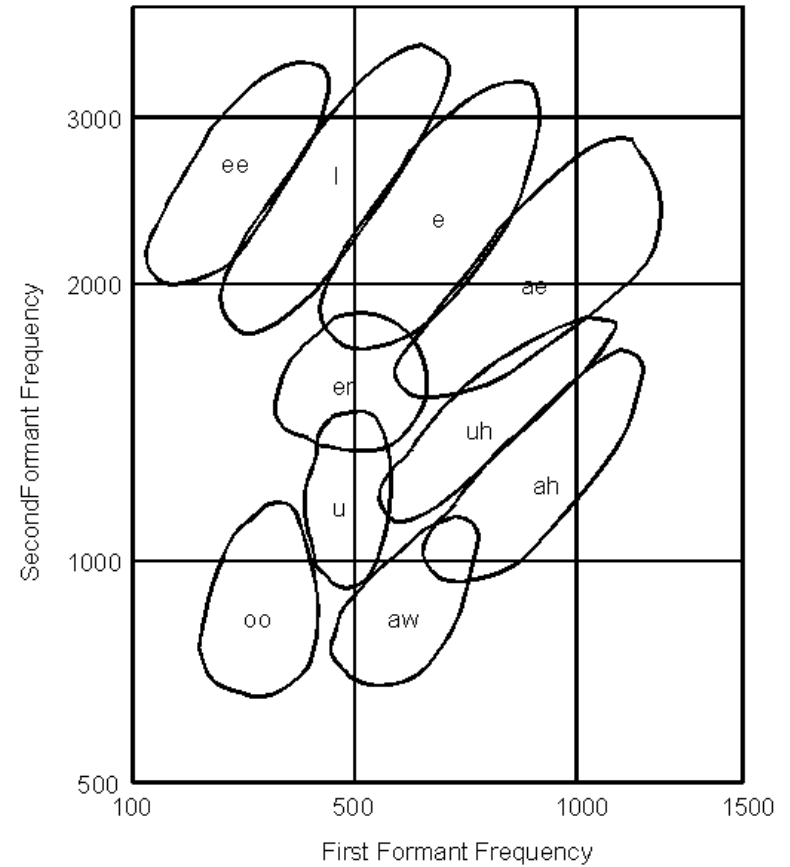
http://www.geofex.com/Article_Folders/sing-wah/sing-wah.htm
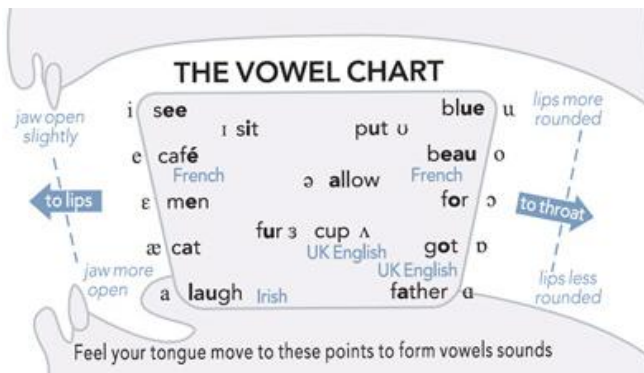
http://www.thedialectcoach.com/images/content/vowel_chart.jpg

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 …

Spectral envelope
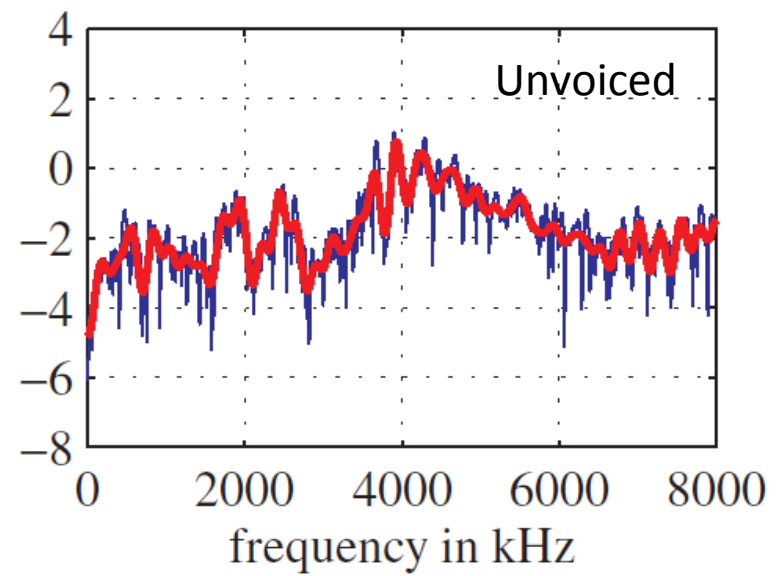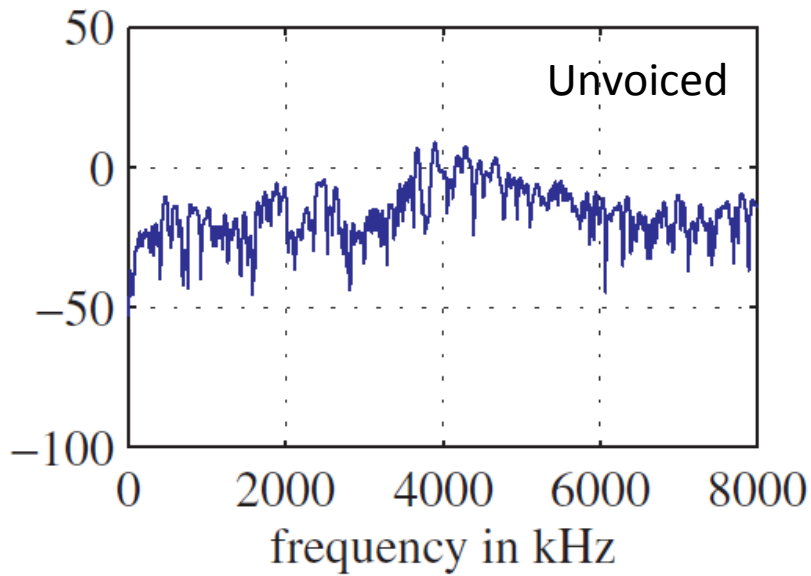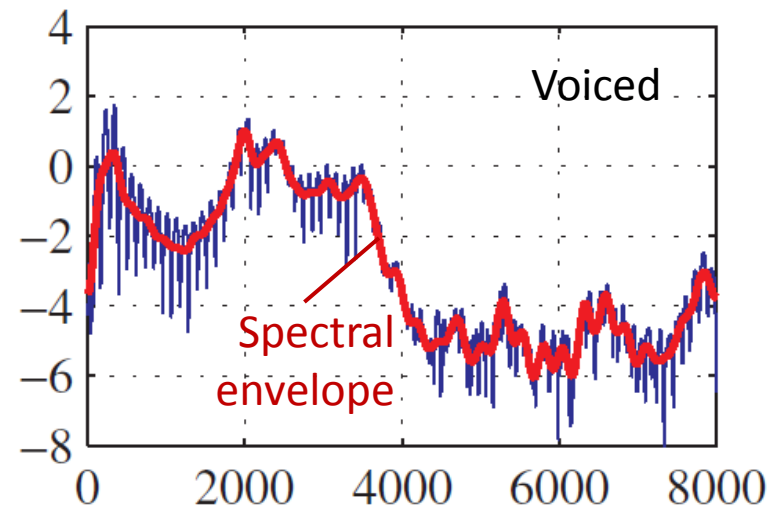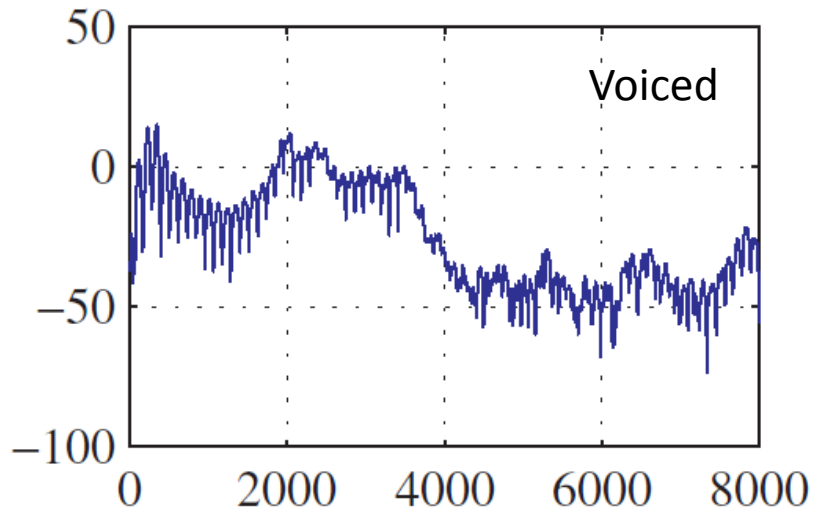
The spectrogram

- Can be thought of as a moving spectrum over time
- Typically represented as a 3D graphic where
  - Horizontal dimension represents time,
  - Vertical dimension represents frequency, and
  - Color represents magnitude (typically in log scale)

Two general types of spectrograms

- Wideband
  - Computed over a short window of time (e.g., 5 ms)
  - High temporal resolution, poor frequency resolution
  - Vertical striations represent individual pitch periods (for voiced phones)
- Narrowband
  - Computed over a relatively large window of time (e.g., 25 ms)
  - High frequency resolution, poor temporal resolution

"*Should we chase.*" Voiced Unvoiced

[Rabiner & Schafer, 2007]

[Taylor, 2009]

Time domain signal   Log magnitude spectrum

64 samples

128 samples

256 samples   Pitch period
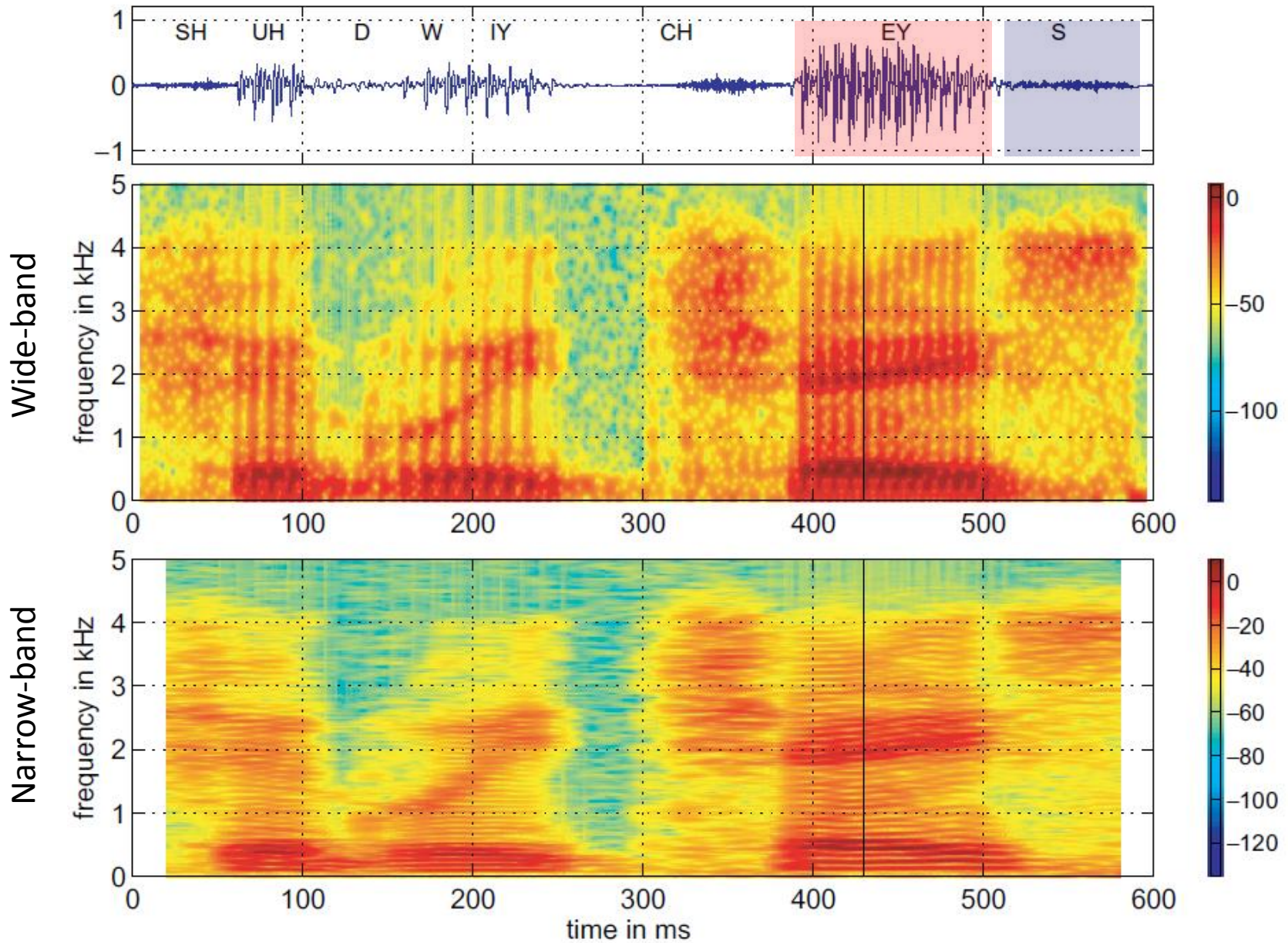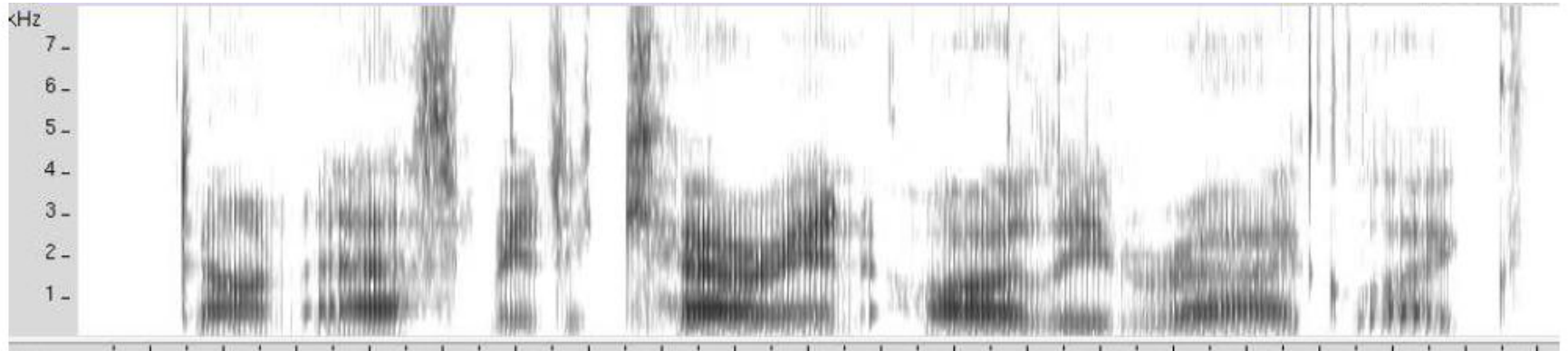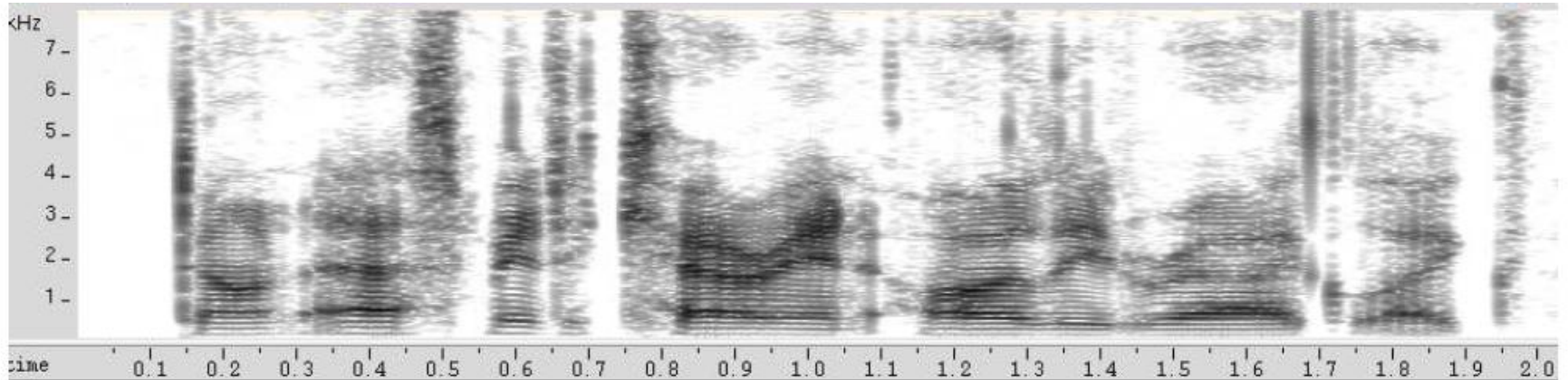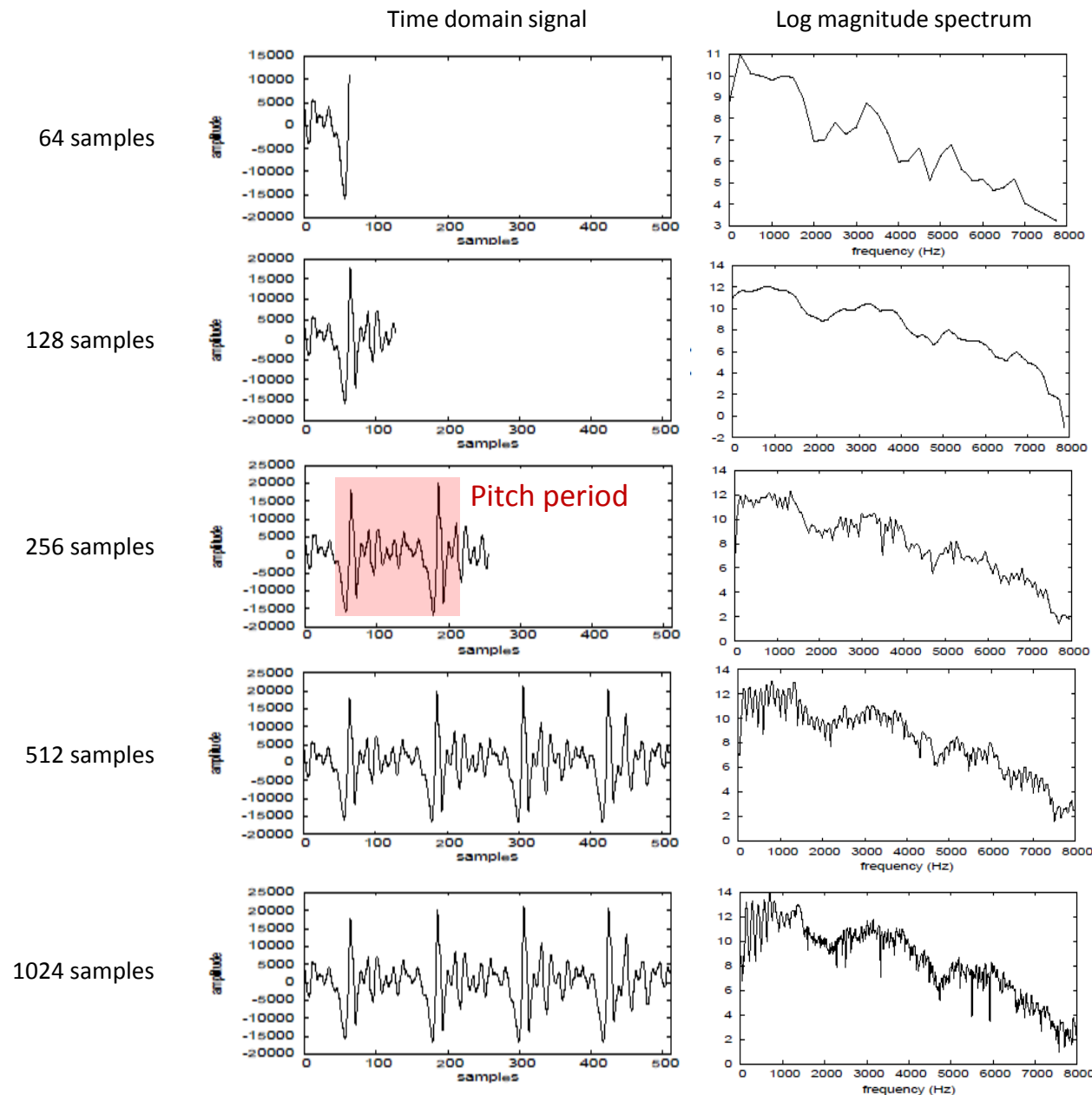
512 samples   Harmonics are clearest for 512 samples.

1024 samples   For longer windows, F0 is not constant and there is no single harmonic spacing

[Taylor, 2009]

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)

[O'Shaughnessy, 2000]
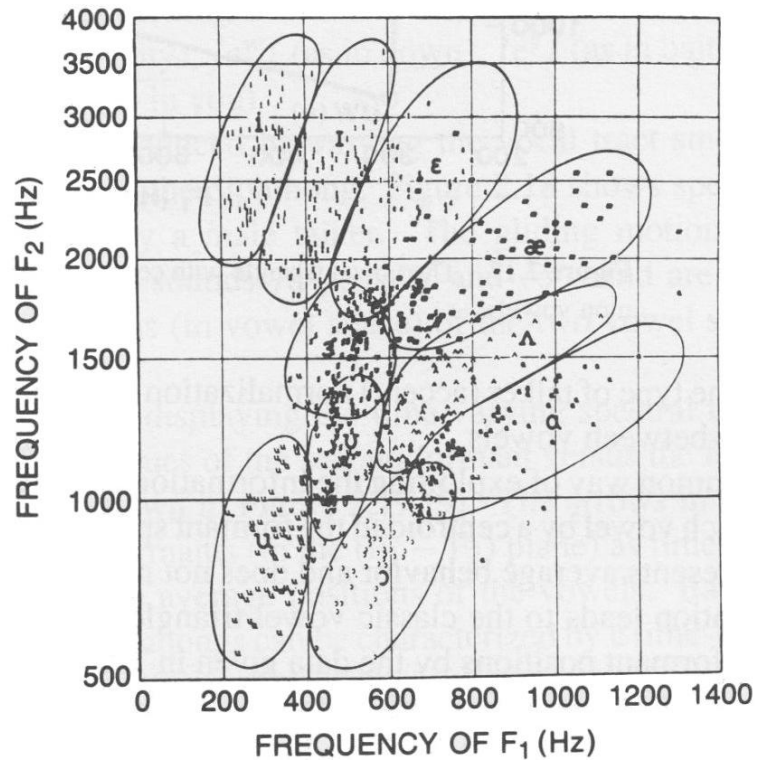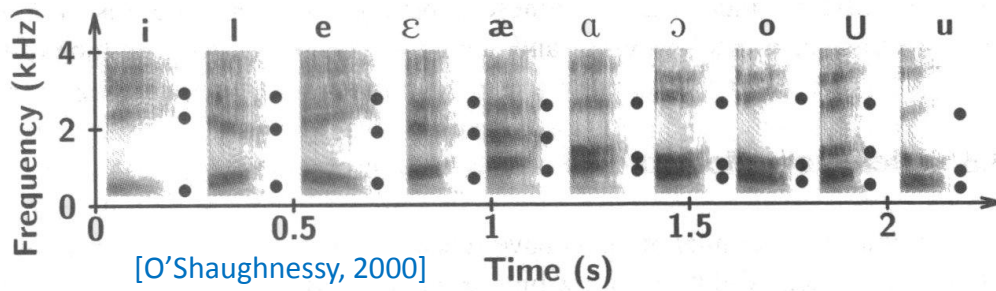
Figure 2.16 Measured frequencies of first and second formants for a wide range of talkers for several vowels (after Peterson & Barney [7]).

[Rabiner & Juang, 1993]

Complete the following sections of text

```
Th_y n_t_d s_gn_f_c_nt _mpr_v_m_nts _n th_
c_mp_ny's _m_g_, s_p_rv_s__n, th__r w_rk_ng
c_nd_t__ns, b_n_f_ts _nd _pp_rt_n_t__s
```

```
A__i_u_e_ _o_a__ _ay __a_e_ e__e__ia__y __e
_a_e, _i__ __e __o_e_ o_ o__u_a_io_a_ e___oyee_
__i____y _e__ea_i__
```

They noted significant improvements in the company's image, supervision, their working conditions, benefits and opportunities

Attitudes toward pay stayed essentially the same, with the scores of occupational employees slightly decreasing

Diphthongs

– Diphthongs consist of a dynamic vowel sound in which the tongue and lips move between two vowel positions

– Three diphthongs are universally accepted

  • These can be described as a vowel followed by a glide, though the articulators rarely reach the full constriction of a glide

  • Examples: [ay] b*i*te, [oy] b*oy*, [aw] b*ou*t

–
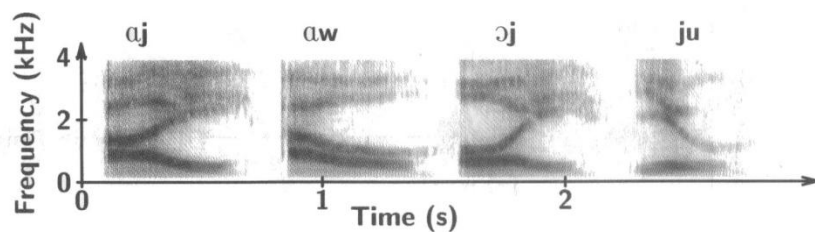
Figure 3.16 Spectrograms of four English diphthongs.

[O'Shaughnessy, 2000]

Figure 3.15 F1 and F2 time movements for diphthongs. The dashed circles indicate typical F1–F2 positions for vowels, and the solid contours enclose formant trajectories during diphthongs. (After Holbrook and Fairbanks [78].)
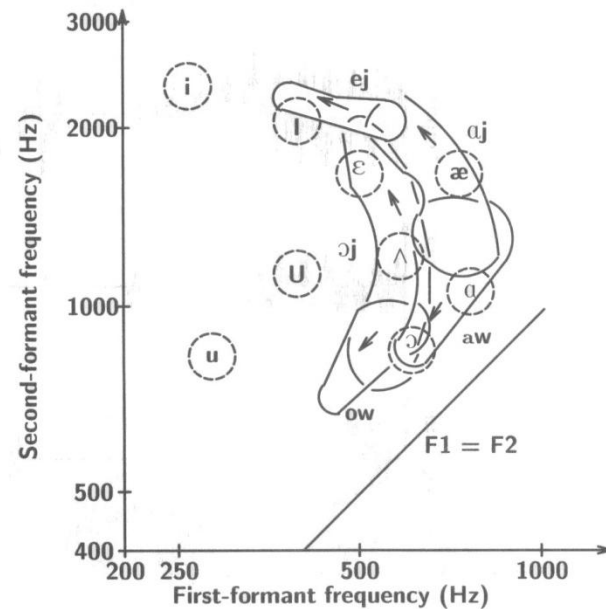
Glides and liquids

- Very similar to vowels: periodic, intense waveforms: most energy in low formant; hence, they are also known as semivowels
  - However, they are weaker than vowels since they require a greater constriction of the vocal tract
  - They are also transitional sounds, so their characteristics are strongly influenced by context
- Liquids ([l], [r]) tend to color preceding vowels with their sound
- Glides ([y], [w]) often lengthen vowels to create diphthongs

Nasals

– Produced with glottal excitation (i.e., voicing)

– Vocal tract totally constricted at some point in the oral cavity

  • Thus, the mouth serves as a resonant cavity that traps energy at certain acoustic frequencies, which appear as anti-resonances (zeros of the transfer function)

– Velum is lowered, so air flows through the nasal cavity

  • Since nasal cavity has large area, formant bandwidths are generally broader than for other sonorants

– Nasal consonants are distinguished by their place of constriction

  • [m] at the lips, [n] at the back of the teeth, [ng] forward of the velum

Effect of vowel nasalisation

- Spectral envelope of a normal [a] and a nasalised [a]

- The first three formants are labeled in the normal vowel

- Reductions in spectrum level in the nasalised vowel are due to the addition of zeroes (labeled as "z")
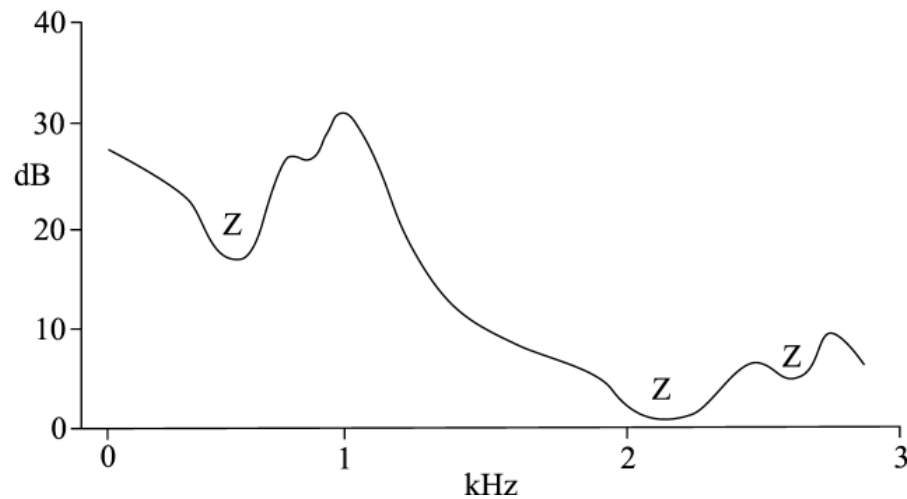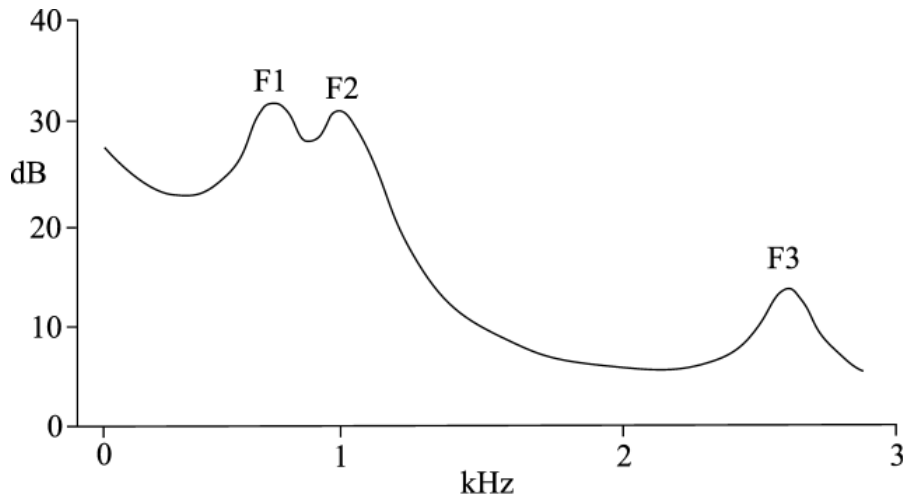
http://clas.mq.edu.au/physiology/nasal/nasality_review.html

Fricatives

- Characterized by a major place of constriction
- Come in two flavors: voiced and unvoiced
  - Place of constriction is identical for corresponding pairs of phones
- Unvoiced
  - Vocal tract is excited by a stead air flow, which becomes turbulent at a place of constriction
    - Sorted by place of constriction: [f] *f*ish, [th] *th*in, [s] *s*ound, and [sh] *sh*out
  - Spectrum looks "noisy"
- Voiced
  - Glottis is vibrating, but the airflow becomes turbulent near the constriction, so these sounds display two distinct components
    - Sorted by place of constriction : [v] *v*ote, [dh] *th*en, [z] *z*oo, [zh] *a*zure
  - Spectrum contains both "noisy" and harmonic components

Unvoiced fricatives

Voiced fricatives
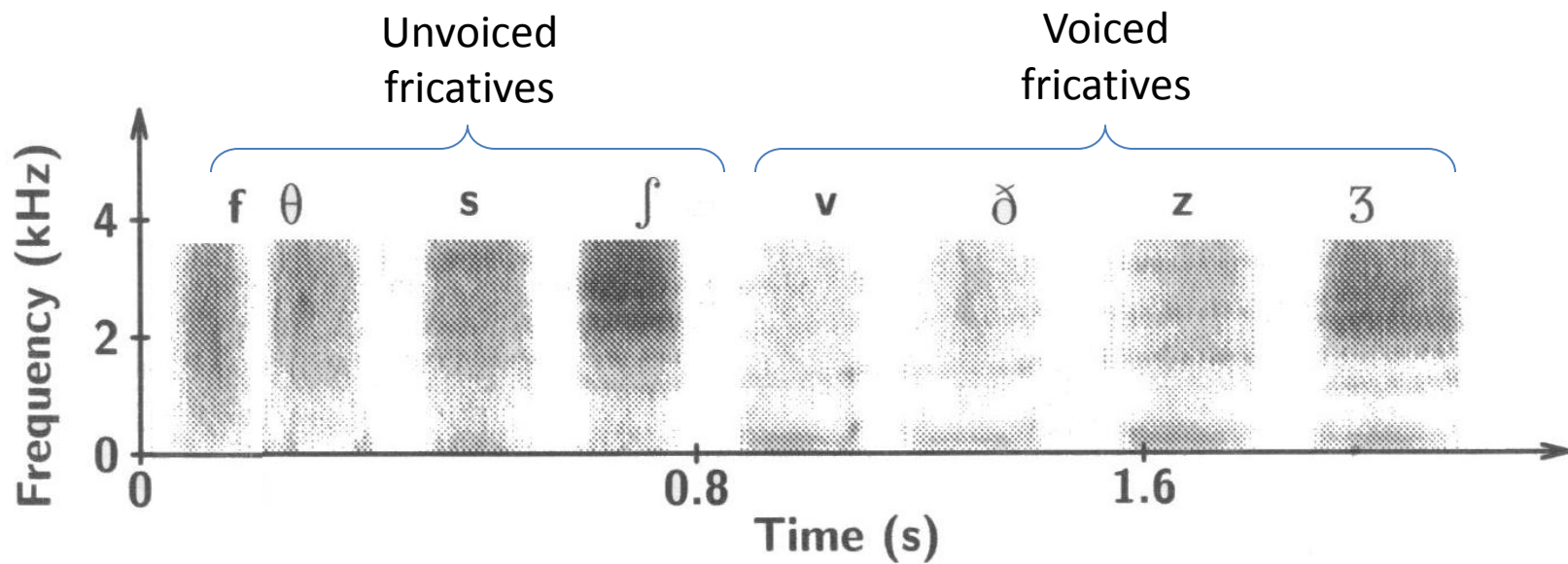
[O'Shaughnessy, 2000]

Stops (plosives)

- Unlike other sounds, which can be described by steady-state spectra, stops are transient phones

  - They are produced by building up pressure behind a constriction and then suddenly releasing it

- As with fricatives, stops come in two flavors: voiced and unvoiced

  - Place of constriction is identical for corresponding pairs of phones: bilabial, alveolar, velar

- Voiced

  - Vocal folds vibrate while pressure builds up, which can be heard as a low-frequency energy radiating through the walls of the throat (voice bar)

  - Sorted: [b] *b*ee, [d] *d*ay, [g] *g*uy

- Unvoiced

  - Vocal folds do not vibrate as pressure builds up

  - Sorted: [p] *p*ea, [t] *t*ea, [k] *k*ey
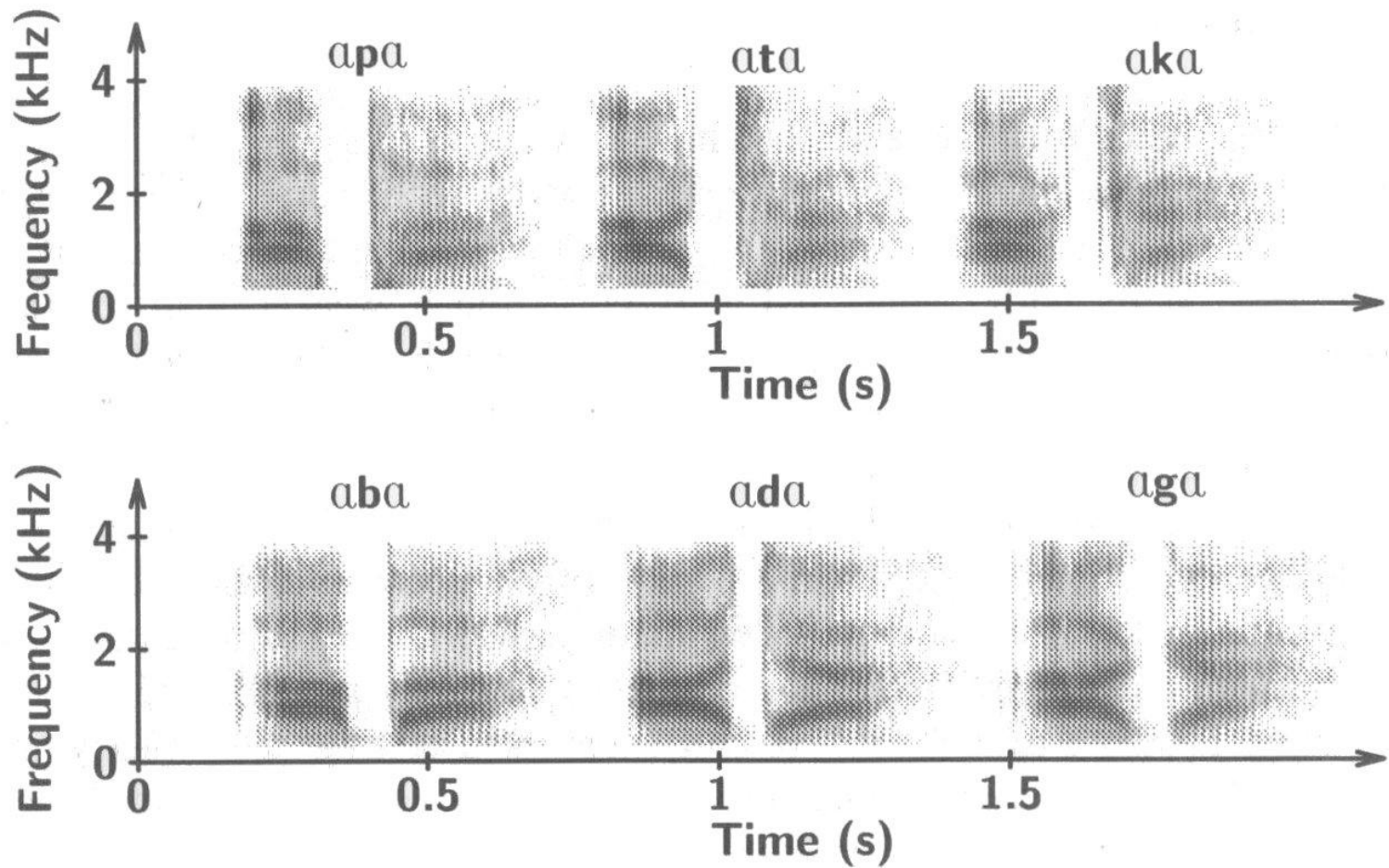
Figure 3.21 Spectrograms of the six English stops in vocalic context: /αpα, αtα, αkα, αbα, αdα, αgα/.

[O'Shaughnessy, 2000]

Speech perception

Vowels
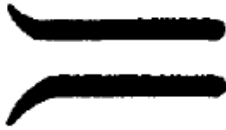
- Vowel perception is relatively simple: formant frequencies are the main factors in vowel identification
- However, formant frequencies scale with vocal tract length
  - There is evidence that listeners "normalize" formant location by making formant spacing essential features of vowel identification
- Vowel nasalization is cued primarily by
  - Increase in the bandwidth of F1, and
  - The introduction of zeros

Consonants

- More complex, and depends on a number of factors
  - Formant transition into the following vowel
  - Formant location
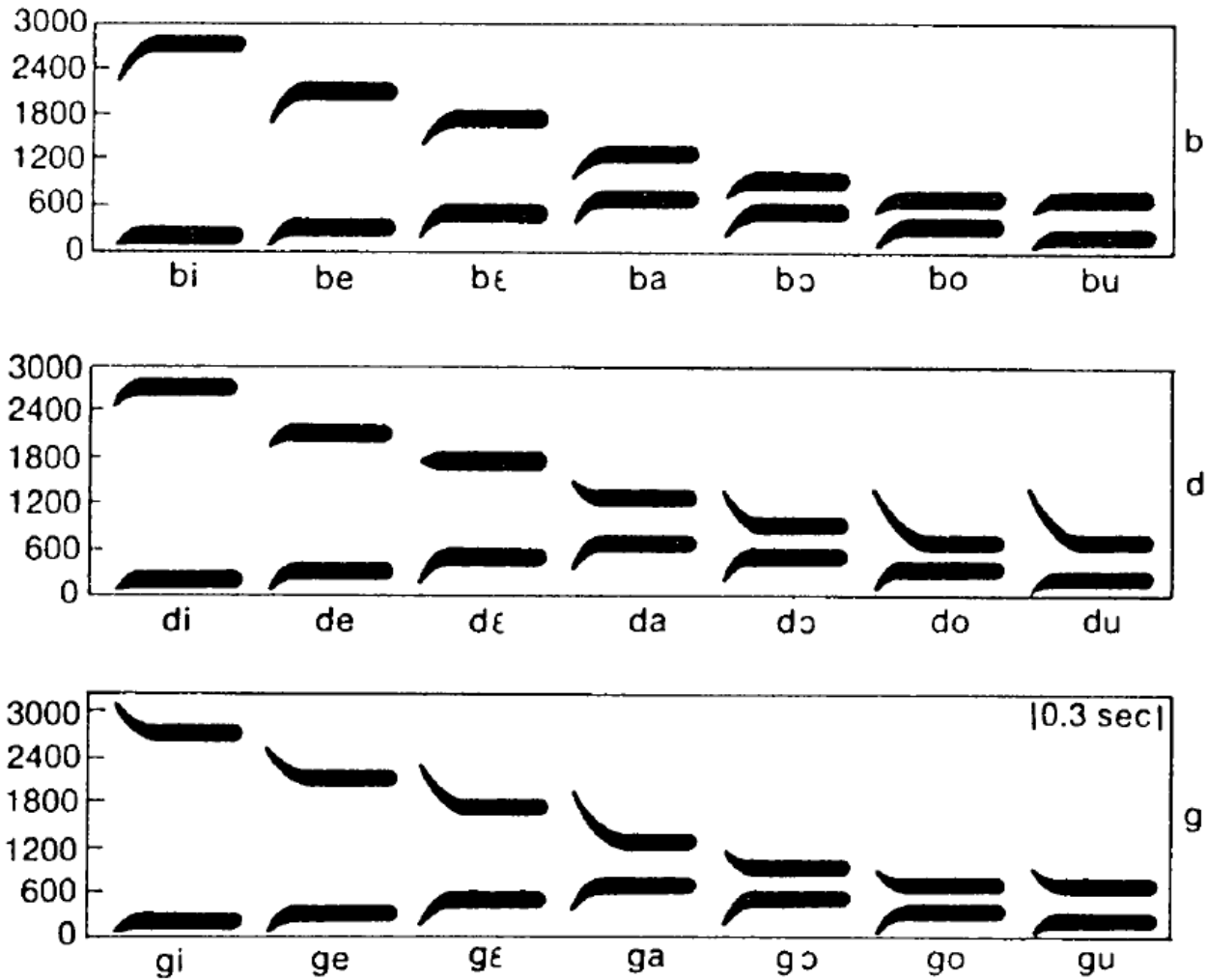 - Voice onset time
  - Voicing

- Direction of formant transition
  - See examples in the next few slides
- Rate of formant transition
  - It is possible to transform the perception of a plosive into a semi-vowel by decreasing the formant transition rate
  - Example: contrast between [b] and [w]
    - On the word 'be', the transition between [b] and [i] is 10 ms
    - As the transition increases beyond 30 ms, 'be' is transformed into 'we'
- Formant locus
  - Vocal tract configuration in front of the closure (for stops)
- Voice onset time (VOT)
  - Length of time between release of a closure and the start of voicing
  - Critical for the perception of stop consonants
  - Example: contrast between [t] and [d]
    - On the word 'do', segment the sound [d] and increase the delay wrt [o]
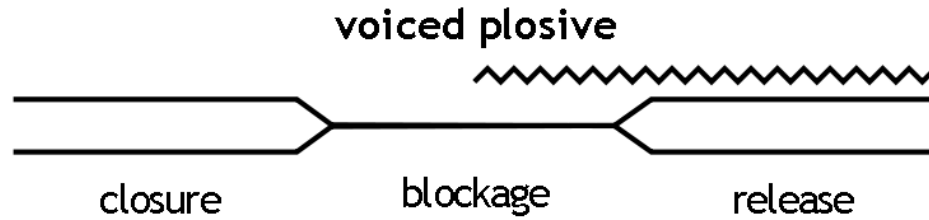   - When VOT exceeds about 25 ms, the word is perceived as 'to'

Schematic formant transition patterns varying in place and manner

Schematic formant transition patterns for voiced stop-vowel syllables

voiced plosive

closure   blockage   release

Negative VOT
[b,d,g]

voiceless unaspirated plosive

closure   blockage   release

Zero VOT
[p,t,k]

voiceless aspirated plosive

closure   blockage   release
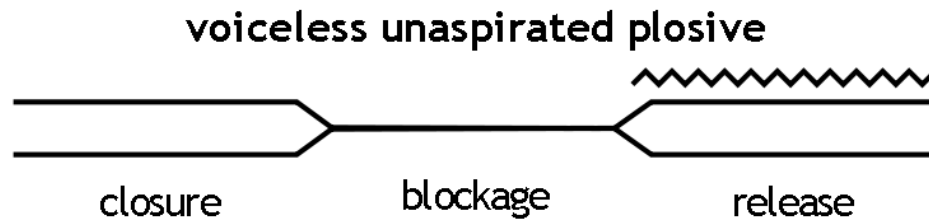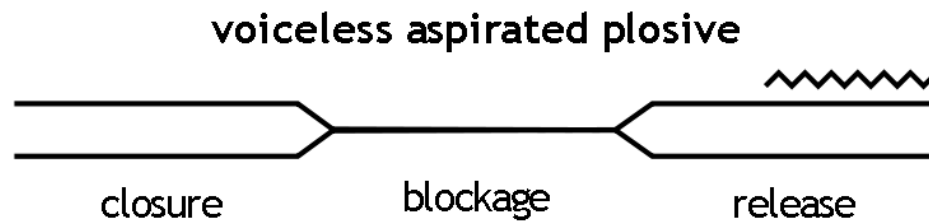
Positive VOT
[p^h,t^h,k^h]

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

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**