Lecture 04:
[Rabiner] Chapter 3. Fundamentals of Human Speech Production

DEEE725 음성신호처리실습

Speech Signal Processing Lab

Instructor: 장길진

Original slides from Lawrence Rabiner

Topics

- Sound production mechanisms of the human vocal tract
- Phonemes to represent distinctive sounds
- Conversion of text to sounds via letter-to-sound rules and dictionary lookup
- Sound representation by acoustic waveforms (time domain)
- Sound representation by spectrograms (frequency domain)
- Articulatory properties of speech sounds—place and manner of articulation
- Sound propagation in the human vocal tract
- Time-varying linear system approaches

The Speech Signal

- Speech is a ***sequence*** of ever changing sounds
- Sound properties are highly dependent on ***context*** (i.e., neighboring sounds which occur before and after the current sound)
- The state of the vocal cords, the positions, shapes and sizes of the various articulators—all change ***slowly*** over time, thereby producing the desired speech sounds
  - need to determine the physical properties of speech by observing and measuring the speech waveform (as well as signals derived from the speech waveform—e.g., the signal spectrum)

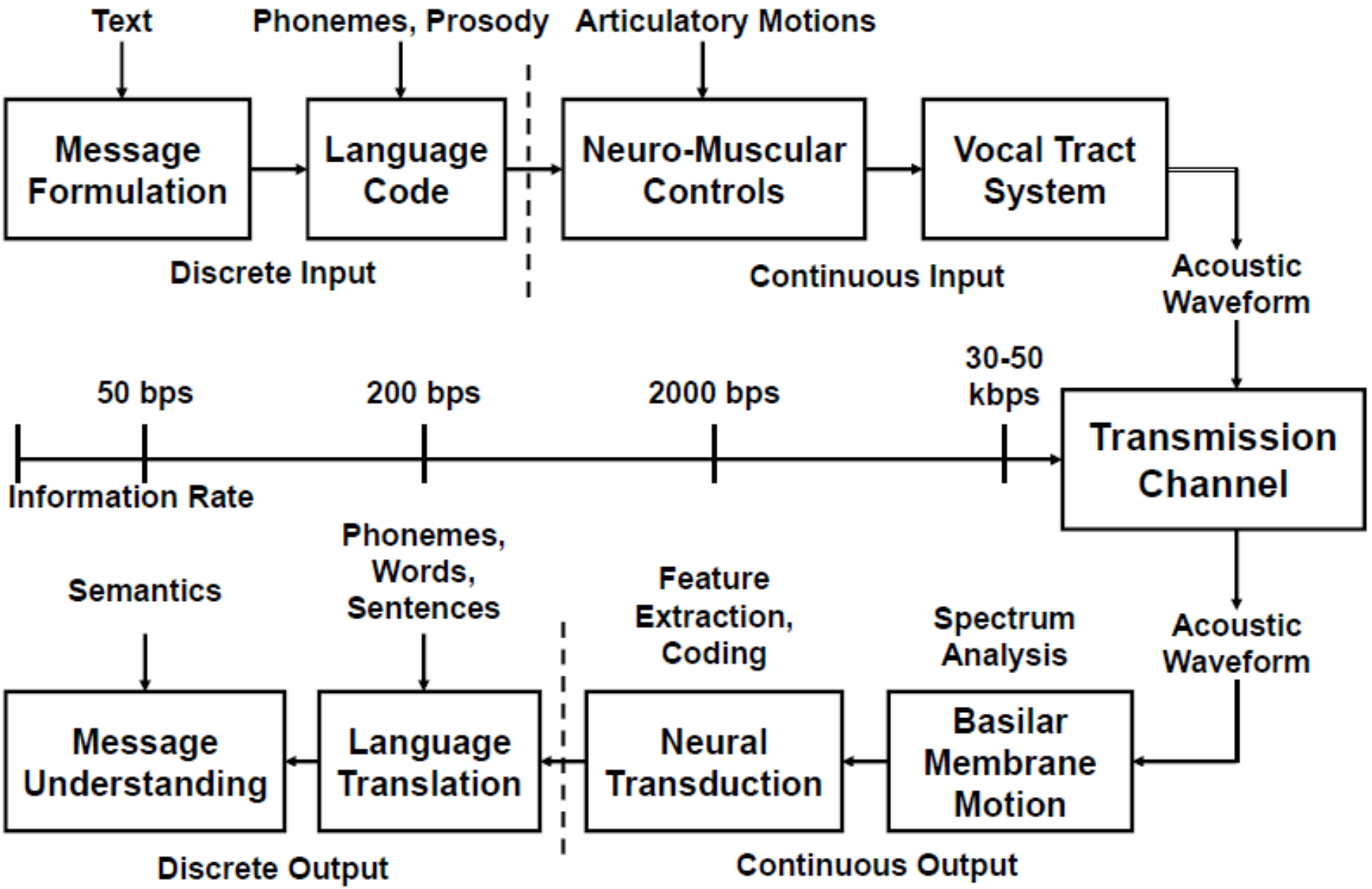
Basic Speech Processes

- **Sender**: idea → sentences → words → sounds → waveform
  - ***Idea***: *it's getting late, I should go to lunch, I should call Al and see if he wants to join me for lunch today*
  - ***Words***: "Hi Al, did you eat yet?"
  - ***Sounds***: /h/ /ay/-/ae/ /l/-/d/ /ih/ /d/-/y/ /u/-/iy/ /t/-/y/ /ɛ/ /t/
  - ***Coarticulated sounds***: /h- ay-l/-/d-ih-j-uh/-/iy-t-j-ɛ-t/ (hial-dijaeajet)
- **Receiver**: waveform → sounds → words → sentences → idea

Basic Speech Processes

- Remarkably, humans can decode these sounds and determine the meaning that was intended—at least at the idea/concept level (perhaps not completely at the word or sound level)
- Often machines can also do the same task in different levels
  - speech coding: waveform → (model) → waveform
  - speech synthesis: words → waveform
  - speech recognition: waveform → words/sentences
  - speech understanding: waveform → idea

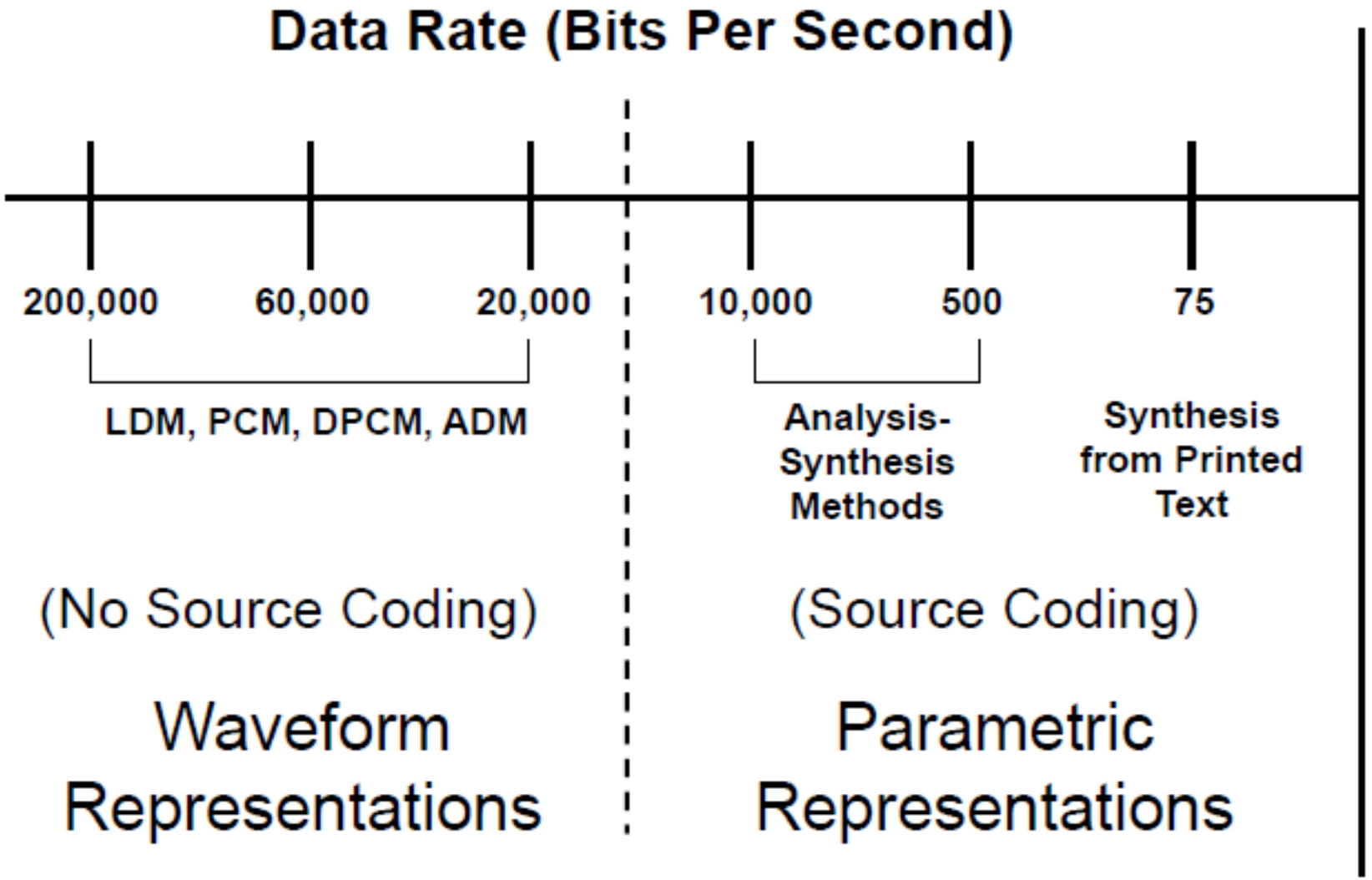
The Speech Chain



Information Rate Analysis

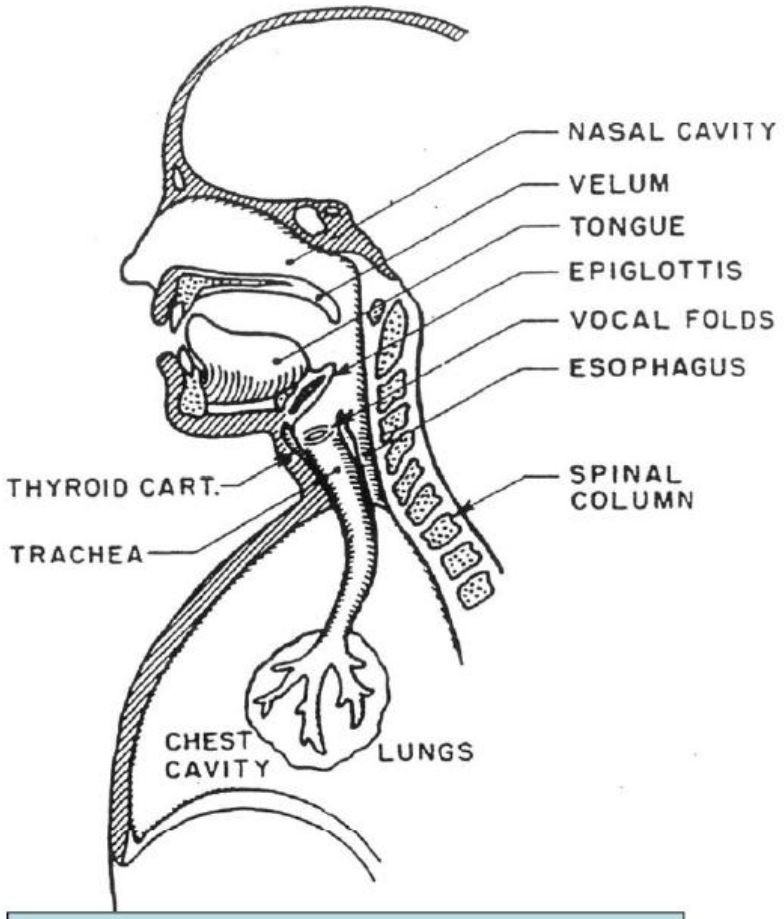
- from a Shannon's view of information:
  - message content/information – 2^6 symbols (phonemes) in the language; 10 symbols/sec for normal speaking rate → 60 bps is the equivalent information rate for speech (need to consider phoneme probabilities, phoneme correlations)
- from a communications point of view:
  - speech bandwidth is between 4 (telephone quality) and 8 kHz (wideband hi-fi speech) – need to sample speech at between 8 and 16 kHz, and need about 8 (log encoded) bits per sample for high quality encoding → 8,000x8=64 kbps (telephone) to 16,000x8=128 kbps (wideband)
- 1000-2000 times change in rate from discrete message symbols to waveform encoding

Information Rate of Speech



SPEECH PRODUCTION

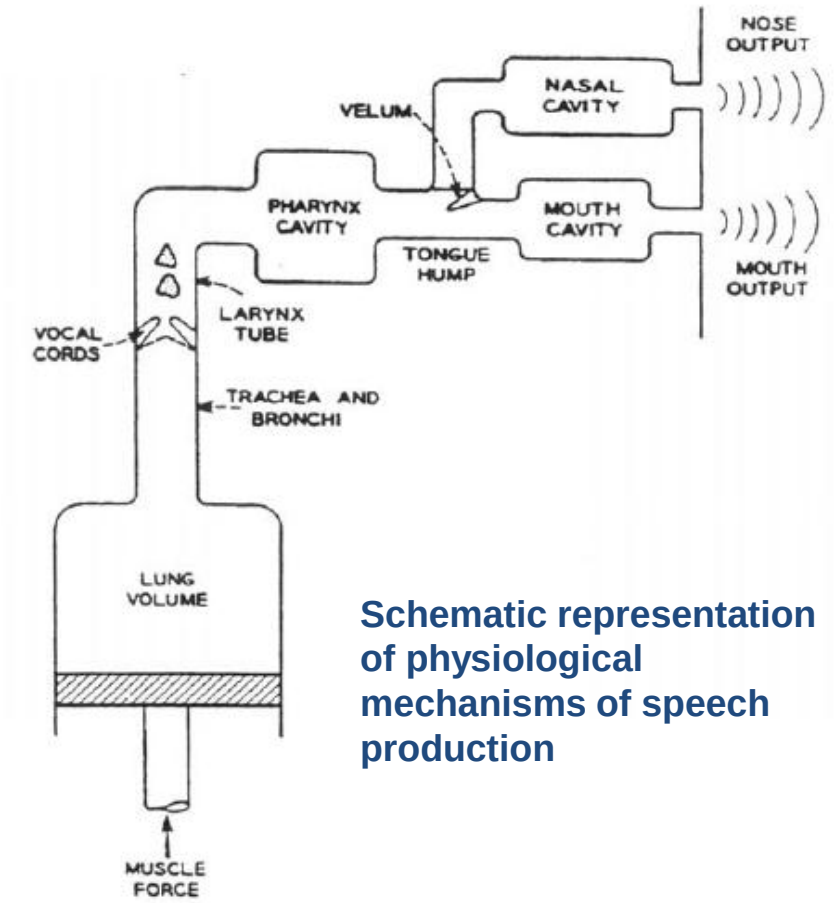
Schematic View of Vocal Tract



Acoustic Tube Models Demo

- Air enters the lungs via normal breathing and no speech is produced (generally) on in-take
- As air is expelled from the lungs, via the trachea or windpipe, the tensed vocal cords within the larynx are caused to vibrate (Bernoulli oscillation) by the air flow
- Air is chopped up into quasi-periodic pulses which are modulated in frequency (spectrally shaped) in passing through the pharynx (the throat cavity), the mouth cavity, and possibly the nasal cavity; the positions of the various articulators (jaw, tongue, velum, lips, mouth) determine the sound that is produced

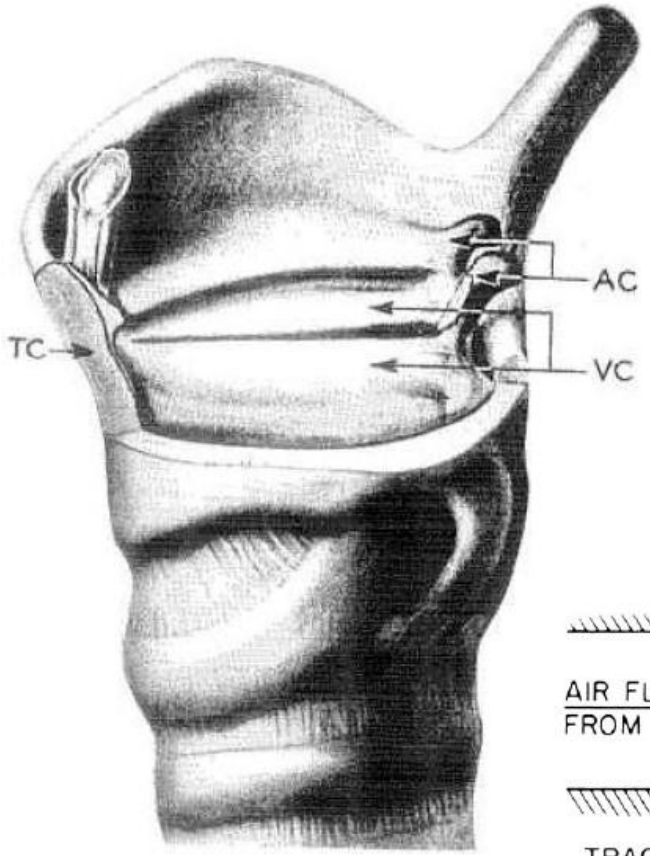
Schematic Production Mechanism



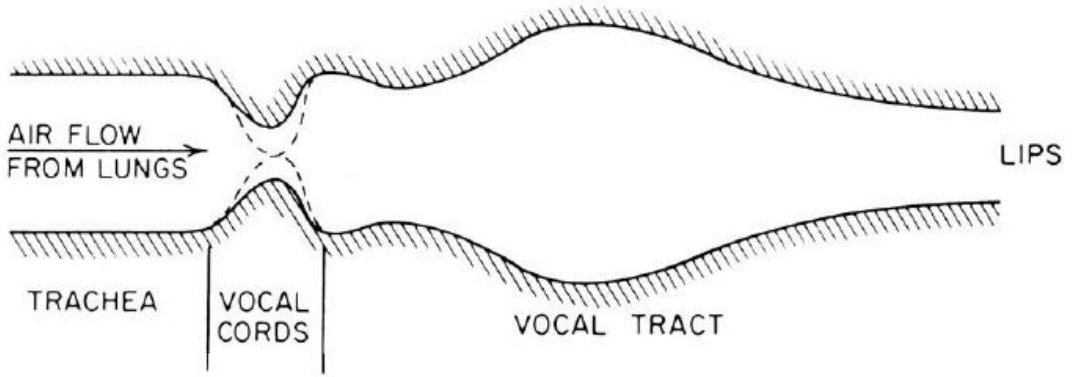
Schematic representation of physiological mechanisms of speech production

- Lungs and associated muscles act as the source of air for exciting the vocal mechanism
- Muscle force pushes air out of the lungs (like a piston pushing air up within a cylinder) through bronchi and trachea
- Based on the behavior of vocal cords, the produced sound can be either **voiced (vibrant)** or **unvoiced (turbulent)**

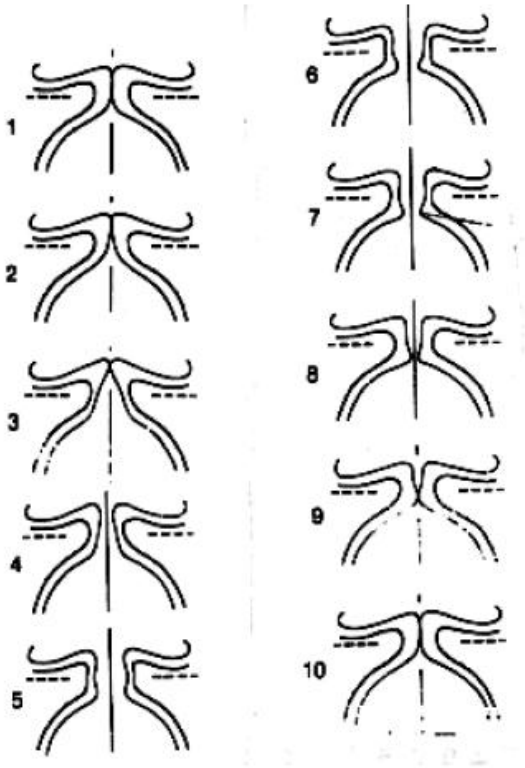
Vocal Cords



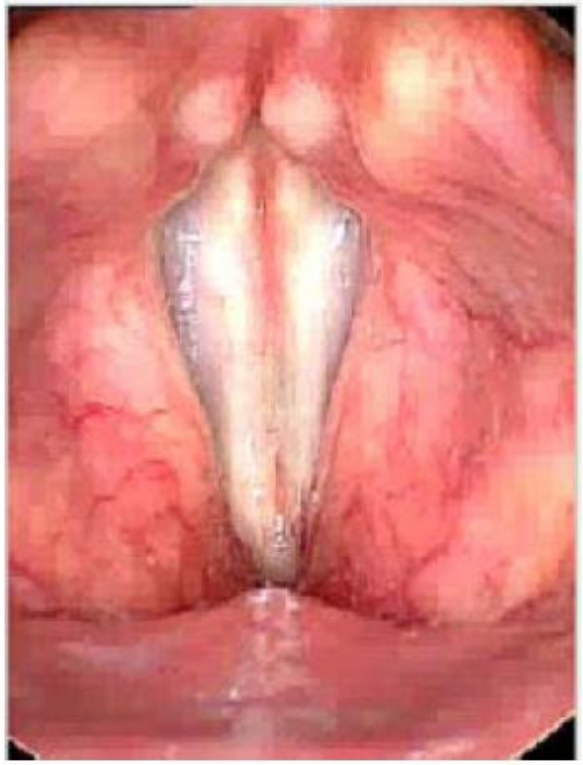
The vocal cords (folds) form a relaxation oscillator. Air pressure builds up and blows them apart. Air flows through the orifice and pressure drops allowing the vocal cords to close. Then the cycle is repeated.



Vocal Cord Views and Operation



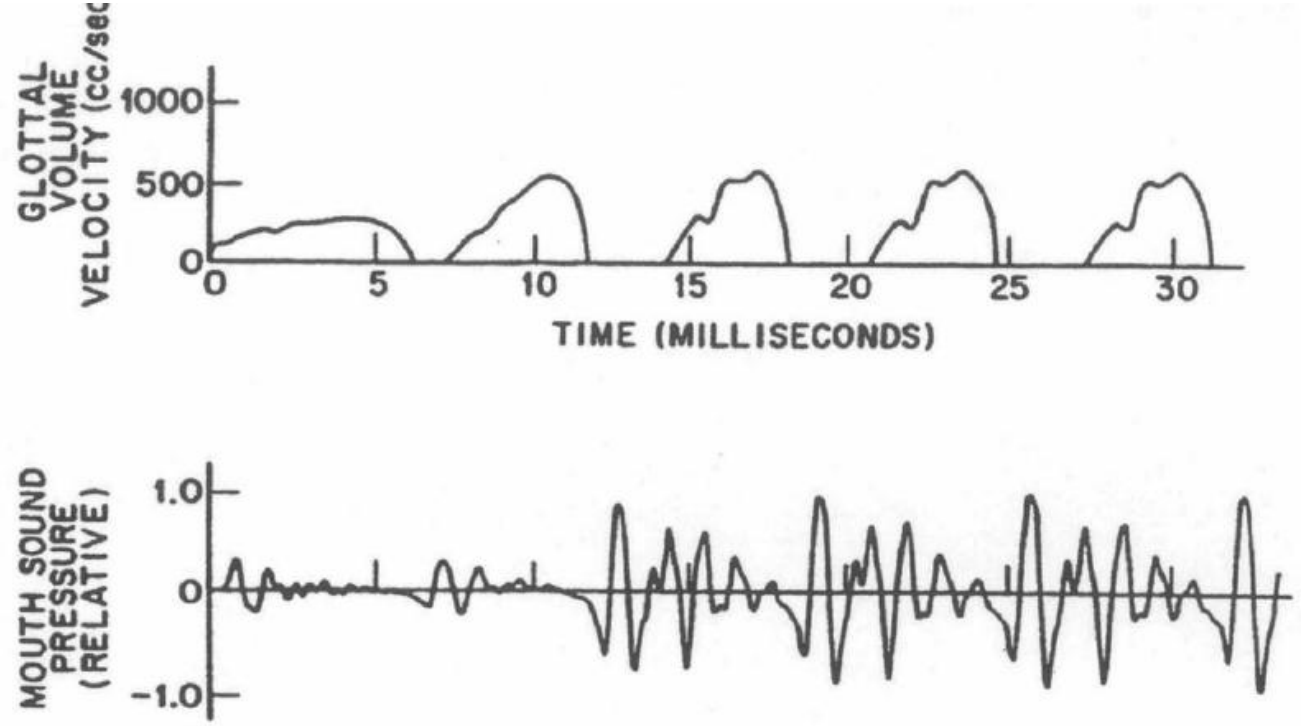
Bernoulli Oscillation



Tensed Vocal Cords – Ready to Vibrate

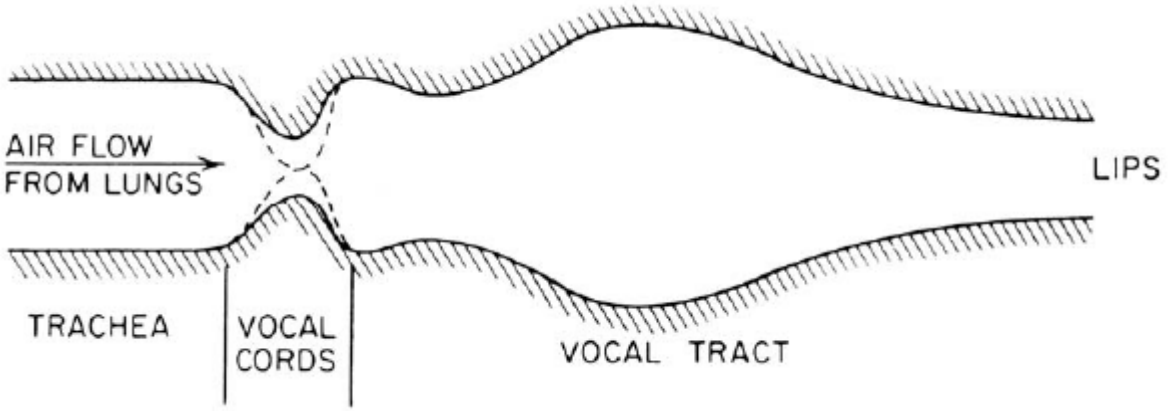
Lax Vocal Cords – Open for Breathing

Glottal Flow



-Glottal volume velocity and resulting sound pressure at the mouth for the first 30 milliseconds of a voiced sound
-15 milliseconds buildup to periodicity ➔ pitch detection issues at beginning and end of voicing; also voiced-unvoiced uncertainty for 15 millisecond

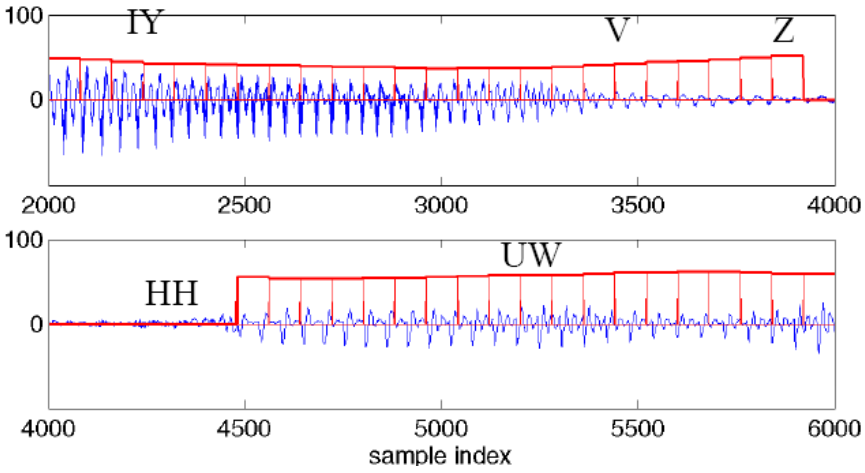
Excitation of Vocal Cords



- If vocal cords are **tensed**, air flow causes them to **vibrate**, producing **voiced** or quasi-periodic speech sounds (**musical notes**)
- If vocal cords are **relaxed**, air flow continues through vocal tract until it hits a constriction in the tract
  - causing it to become **turbulent**, producing **unvoiced** sounds (like **/s/**, **/sh/**)
  - or it hits a point of total closure in the vocal tract, building up pressure until the closure is opened and the pressure is suddenly and abruptly released, causing a brief **transient** sound, like at the beginning of **/p/**, **/t/**, or **/k/**

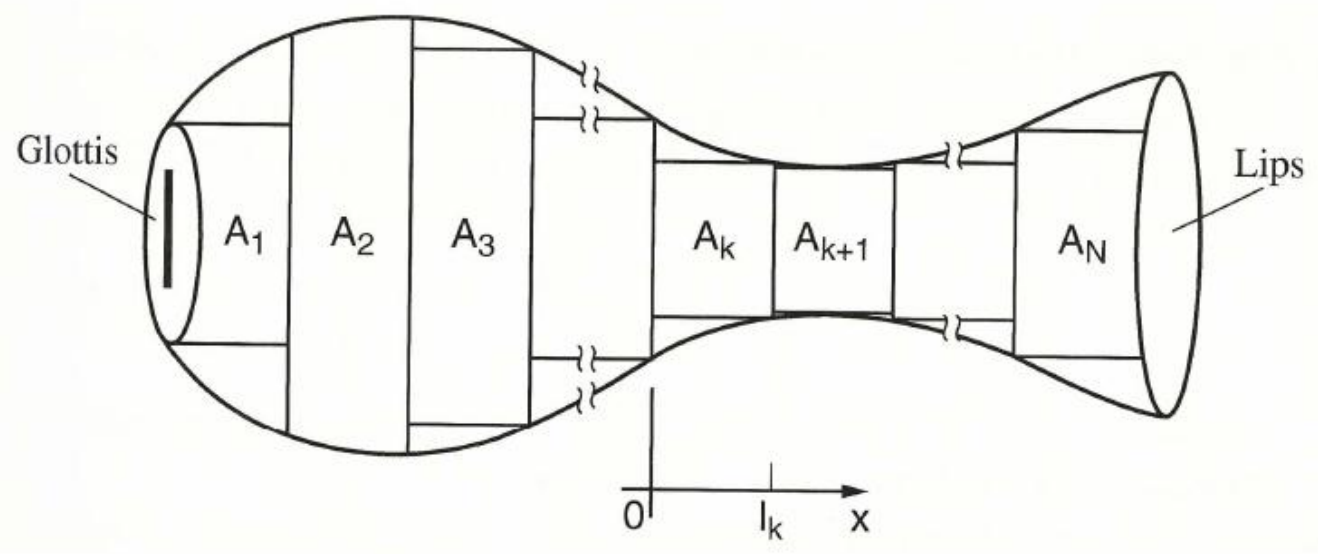
Pitch = 1/F0

- **F0** (Fundamental frequency)
  - the oscillating frequency of vocal cord for voiced sound; there is **no F0 for unvoiced sounds**
  - characterizes **personal identity**; males have relatively low F0 than females
  - analogous to pitch or pitch period, which is the time for a single period; males have relatively longer pitch



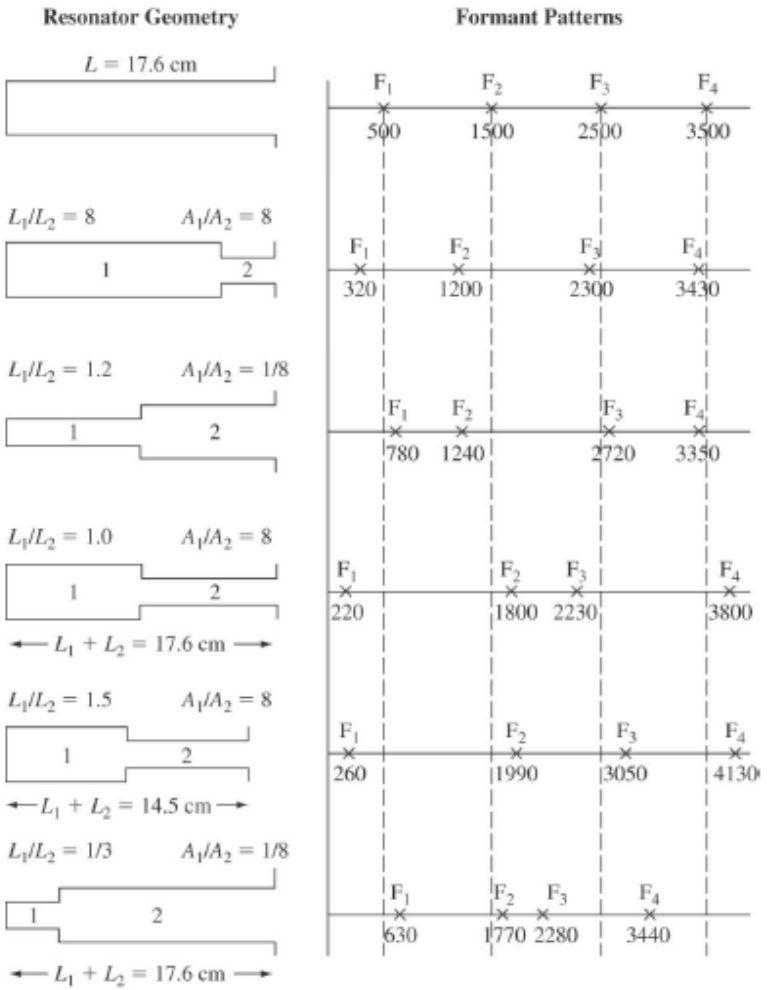
	F_0 ave (Hz)	F_0 min (Hz)	F_0 max (Hz)
Men	125	80	200
Women	225	150	350
Children	300	200	500

Concatenated Tube Models



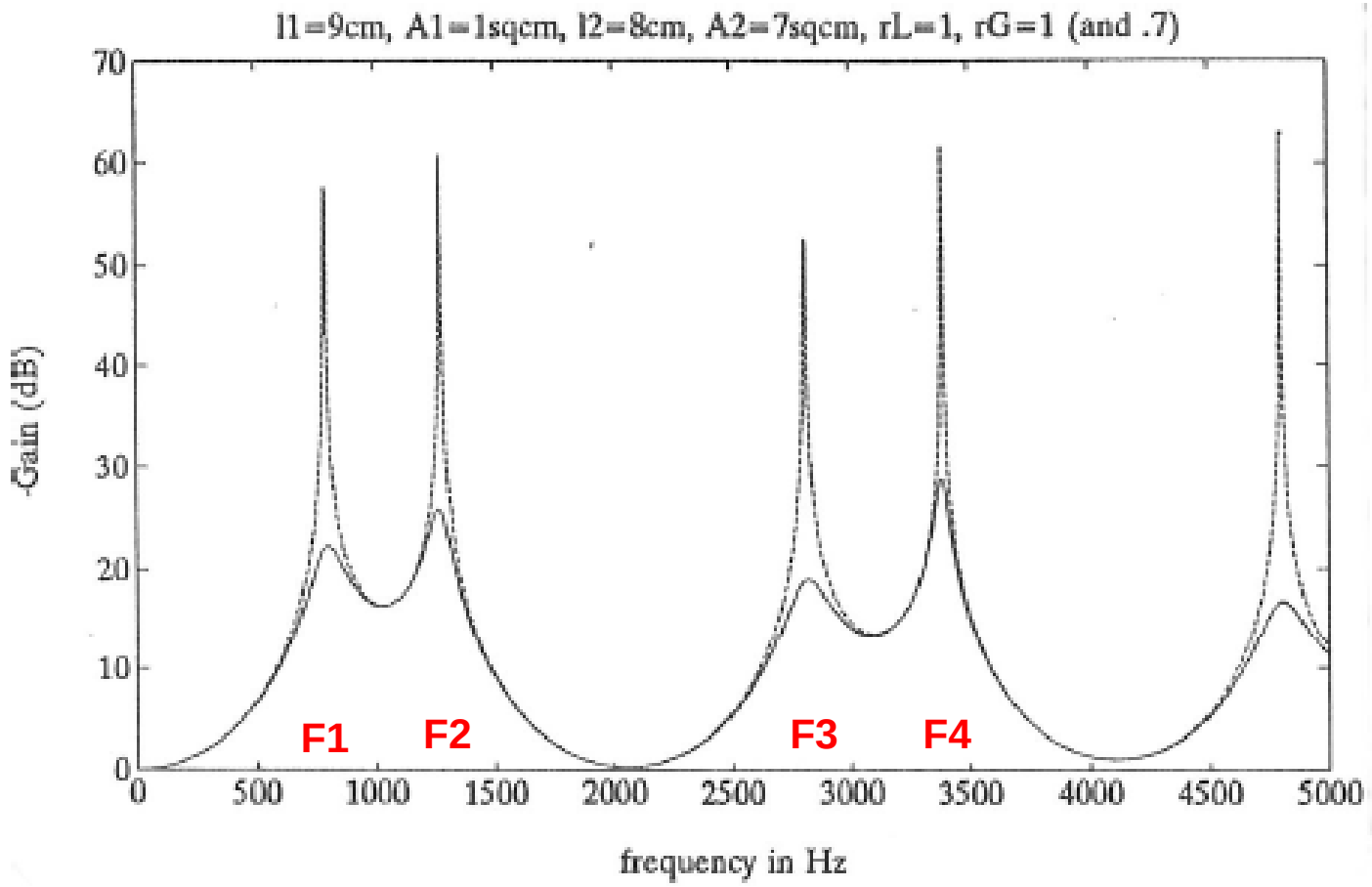
- human vocal tract is essentially a tube of varying cross sectional area, or can be approximated as a concatenation of tubes of varying cross sectional areas
- acoustic theory shows that the transfer function of energy from the excitation source to the output can be described in terms of the natural frequencies or resonances of the tube

Two Tube Model Resonances

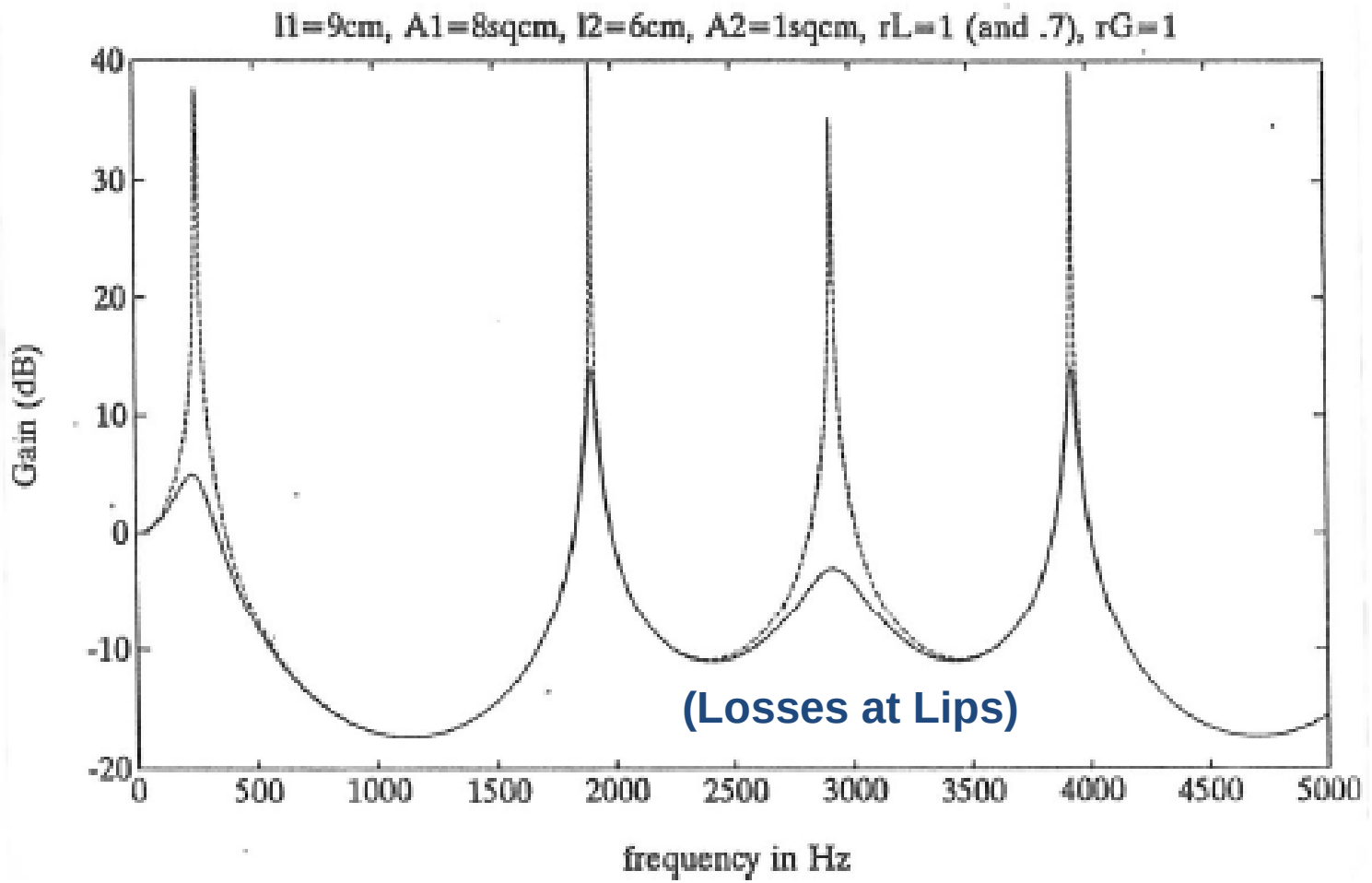


- (F1, F2, …): **formants** or **formant frequencies**
  - a set of resonating frequencies that pass the most acoustic energy from the source to output
  - typically there are 3 significant formants below about 3500 Hz, depending on the vocal tract area function
  - characterizes **phones**; which sound is being pronounced
  - unvoiced sounds also have formants, since the turbulence passes through vocal tract

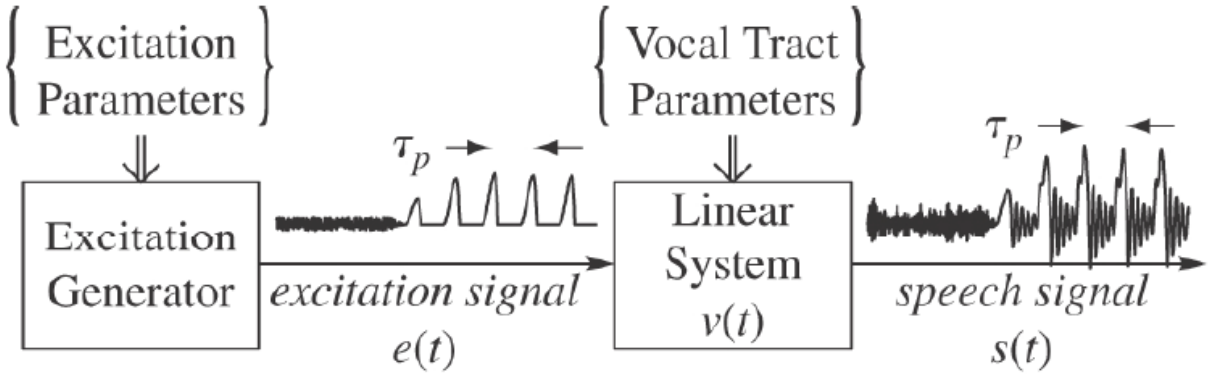
Two-Tube Model for Vowel /AA/

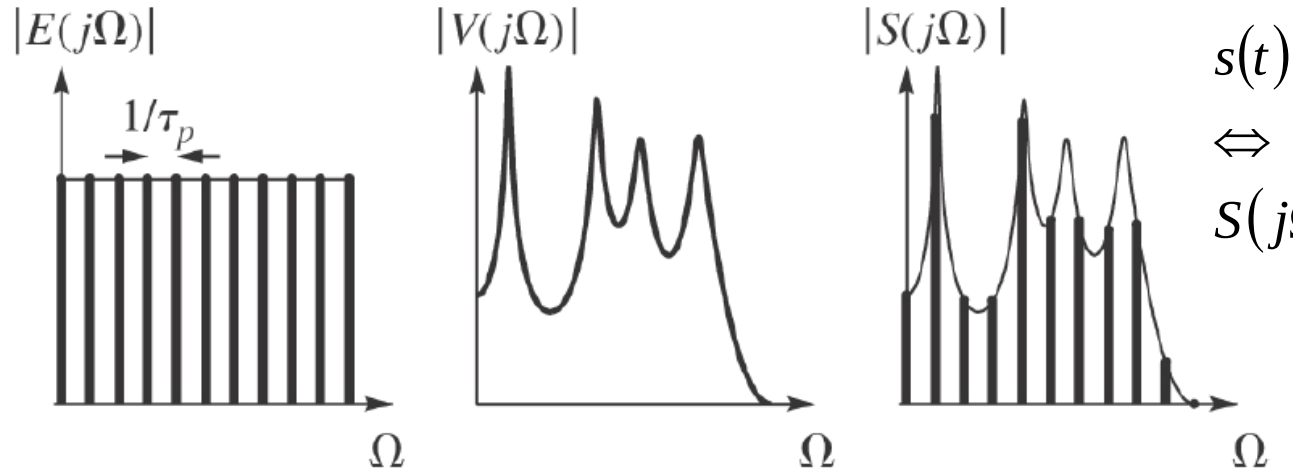


Two-Tube Model for Vowel /IY/



Source-System Model of Speech Production



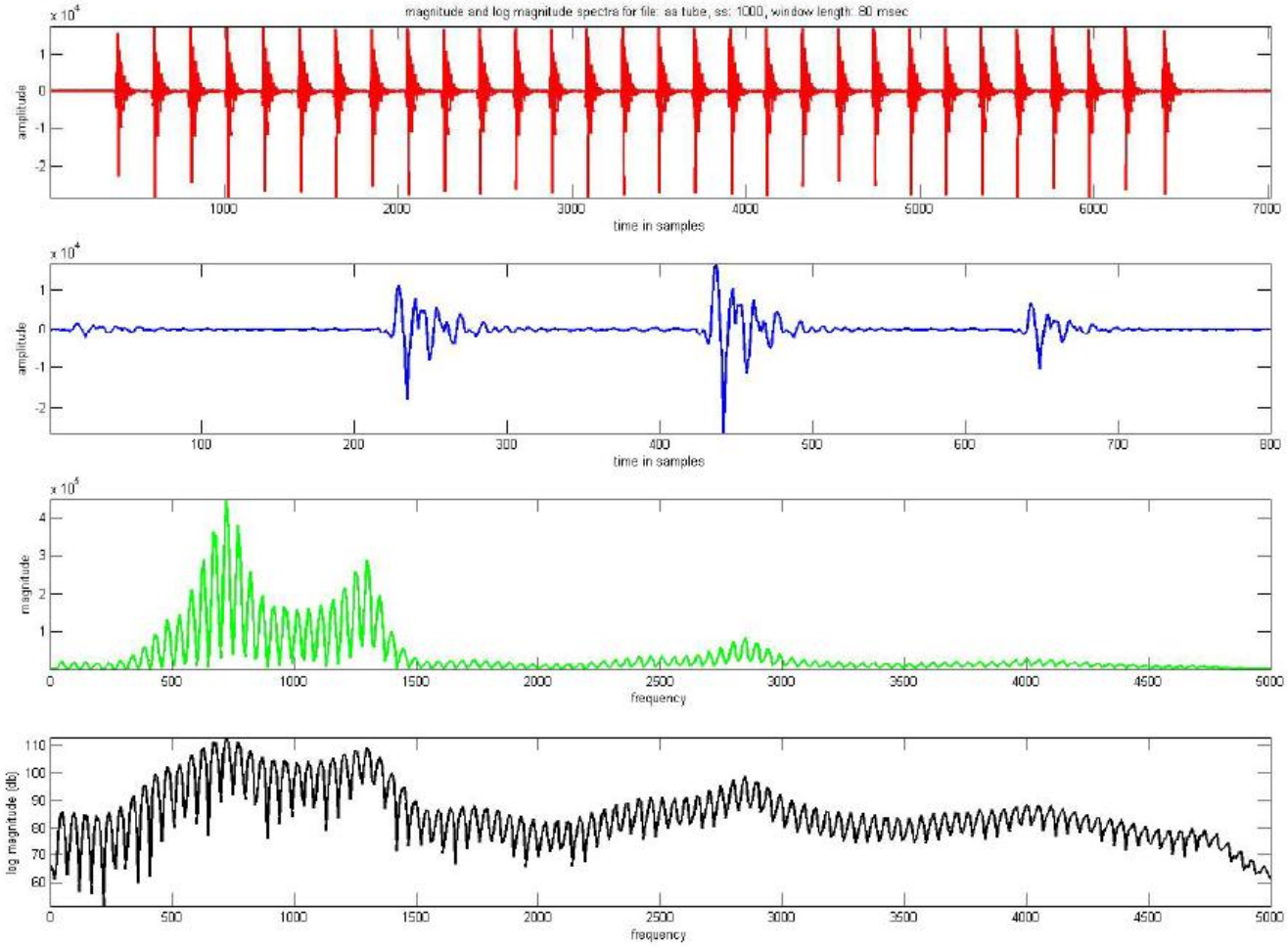


$$s(t) = e(t) * v(t)$$

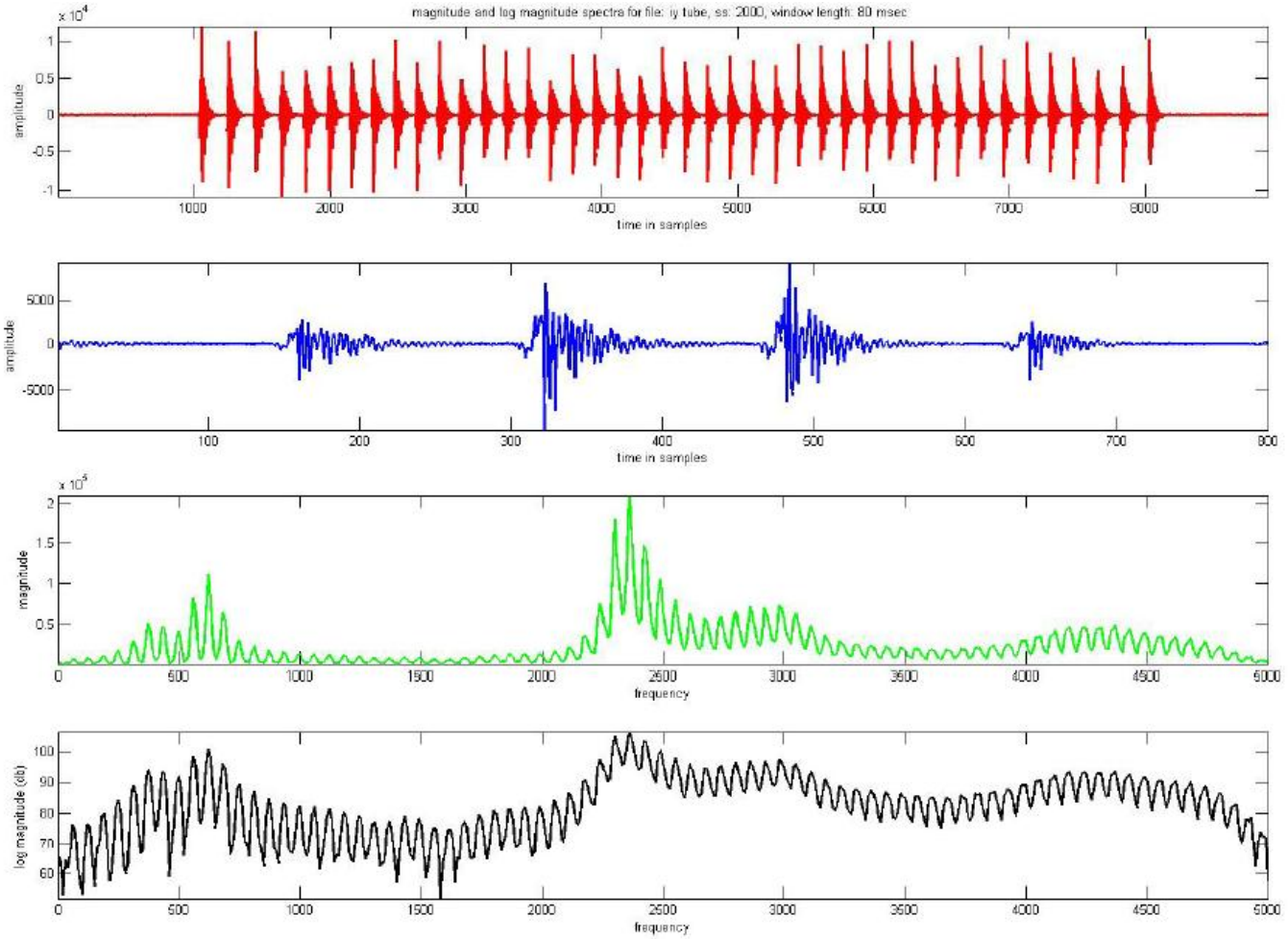
$$\Leftrightarrow$$

$$S(j\Omega) = E(j\Omega) \cdot V(j\Omega)$$

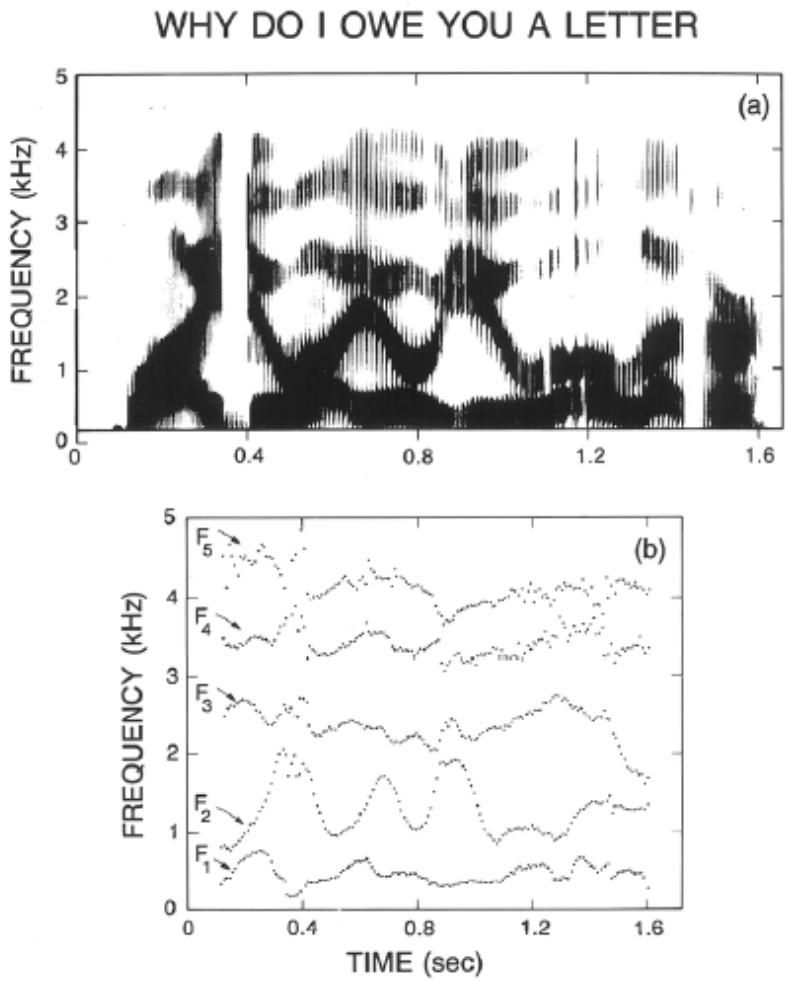
Tube Models – Low Frequency



Tube Models – High Frequency



Spectrogram and Formants



<u>Key Issue</u>: reliability in estimating formants from spectral data

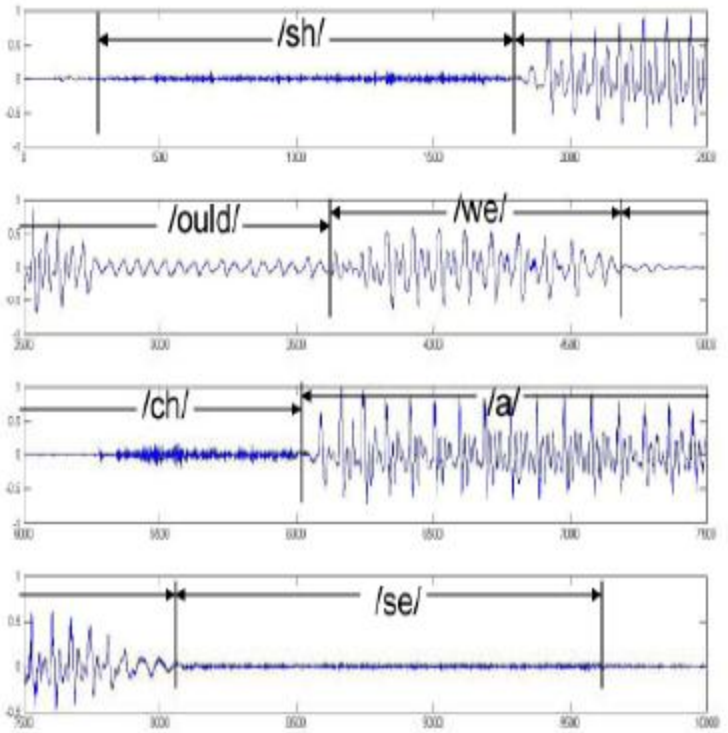
Sound Source for Unvoiced Sounds

- Turbulence noise is produced at a constriction in the vocal tract
  - Aspiration noise is produced at glottis
  - Frication noise is produced above the glottis

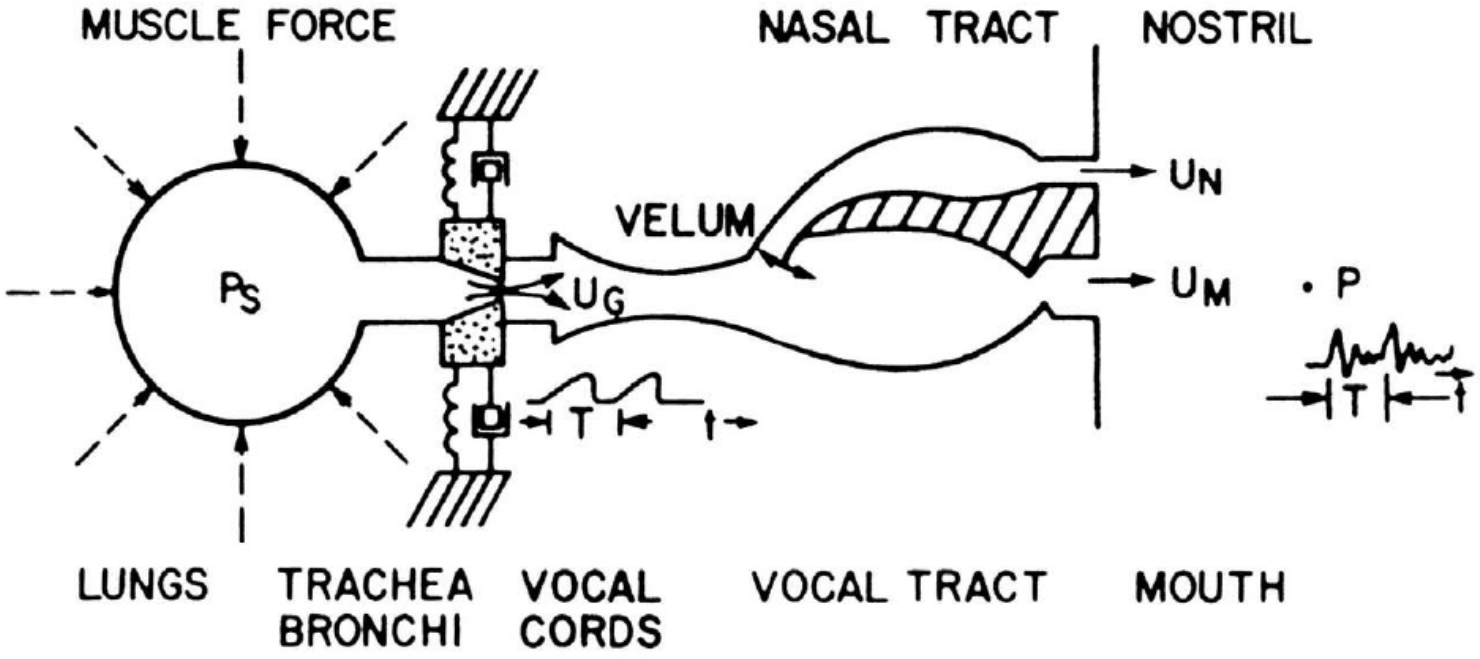
Speech Waveforms and Sounds

- "Should we chase"
  - /sh/ sound
  - /ould/ sounds
  - /we/ sounds
  - /ch/ sound
  - /a/ sound
  - /s/ sound

  - hard to distinguish weak sounds from silence



Review: Abstractions of Physical Model



General Synthesis Model

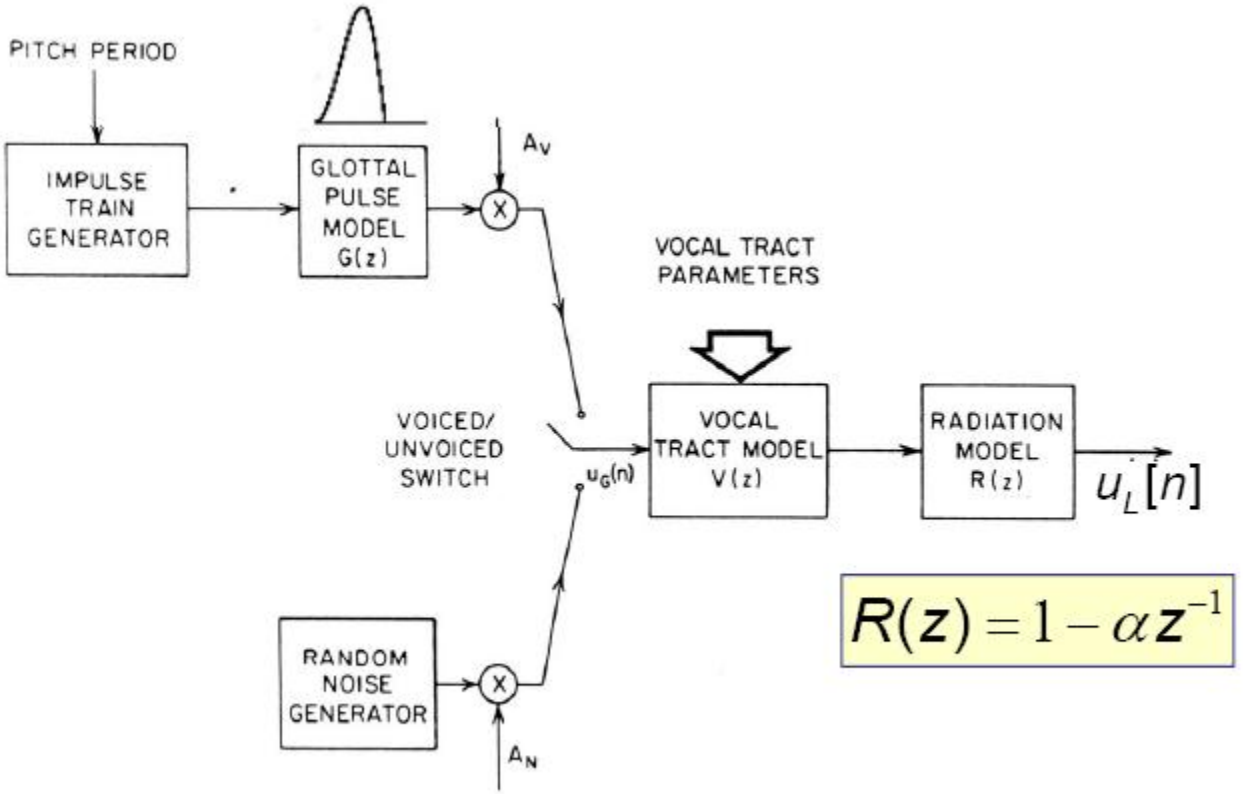


Fig. 3.50 General discrete-time model for speech production.

ACOUSTIC PHONETICS

English Speech Sounds

A condensed list of phonetic symbols for American English

Phoneme	ARPAbet	Example	Phoneme	ARAPAbet	Example
/i/	IY	beat	/ŋ/	NX	sing
/ɪ/	IH	bit	/p/	P	pet
/e/ (eʸ)	EY	bait	/t/	T	ten
/ɛ/	EH	bet	/k/	K	kit
/æ/	AE	bat	/b/	B	bet
/ɑ/	AA	Bob	/d/	D	debt
/ʌ/	AH	but	/g/	G	get
/ɔ/	AO	bought	/h/	HH	hat
/o/ (oʷ)	OW	boat	/f/	F	fat
/ʊ/	UH	book	/θ/	TH	thing
/u/	UW	boot	/s/	S	sat
/ə/	AX	about	/š/	SH	shut
/ɨ/	IX	roses	/v/	V	vat
/ɝ/	ER	bird	/ð/	DH	that
/ɚ/	AXR	butter	/z/	Z	zoo
/ɑʷ/	AW	down	/ž/	ZH	azure
/ɑʸ/	AY	buy	/č/	CH	church
/ɔʸ/	OY	boy	/ǰ/	JH	judge
/y/	Y	you	/ʍ/	WH	which
/w/	W	wit	/ l̩ /	EL	battle
/ɾ/	R	rent	/ m̩ /	EM	bottom
/l/	L	let	/ n̩ /	EN	button
/m/	M	met	/Γ/	DX	batter
/n/	N	net	/ʔ/	Q	(glottal stop)

Phonetic symbol representation

- Phoneme by IPA (international phonetic alphabet)
- ARPAbet representation
  - by ARPA (advanced research project agency)
  - 48 sounds
    - 18 vowels/diphthongs
    - 4 vowel-like consonants
    - 21 standard consonants
    - 4 syllabic sounds
    - 1 glottal stop

Phonemes—Link Between Orthography and Speech

- **Orthography** → sequence of sounds
  - Larry → /l/ /ae/ /r/ /iy/ (/L/ /AE/ /R/ /IY/)
- **Speech Waveform** → sequence of sounds
  - based on acoustic properties (temporal) of phonemes
- **Spectrogram** → sequence of sounds
  - based on acoustic properties (spectral) of phonemes

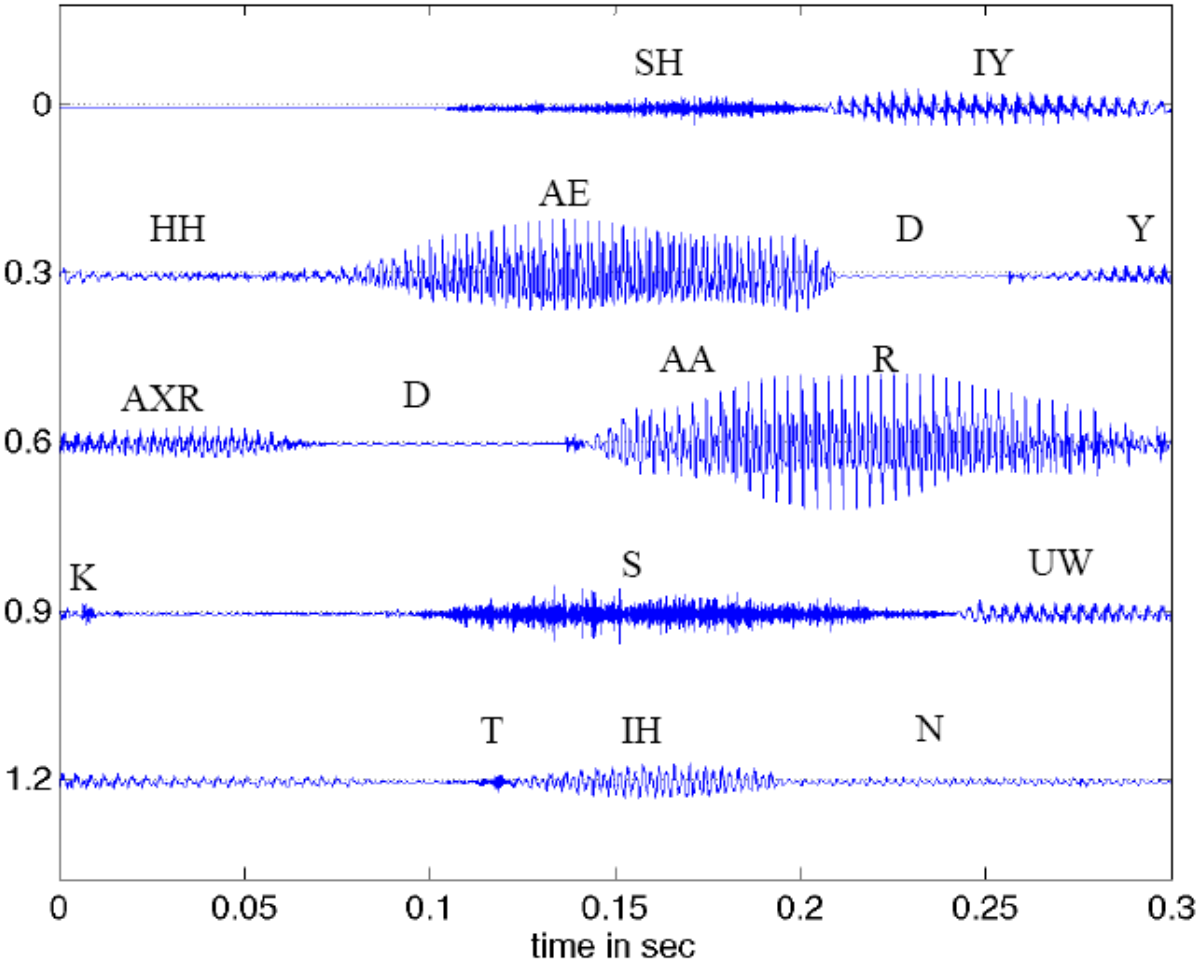
- The bottom line is that we use a phonetic code as an intermediate representation of language, from either orthography or from waveforms or spectrograms
- we have to learn how to recognize sounds within speech utterances

Phonetic Transcriptions

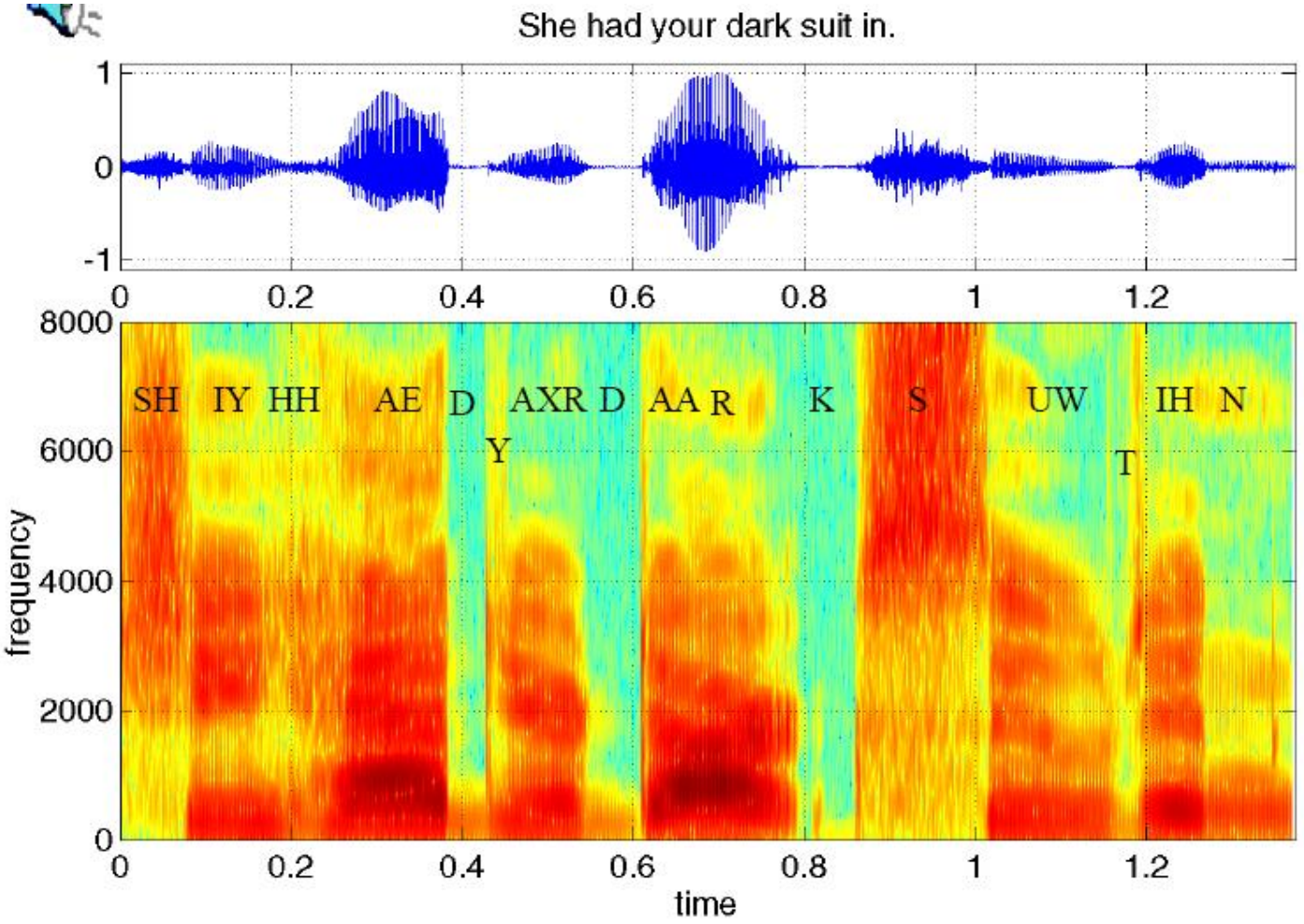
- based on ***ideal*** (dictionary-based) pronunciations of all words in sentence
  - 'My name is Larry'-/M/ /AY/-/N/ /EY/ /M/-/IH/ /Z/-/L/ /AE/ /R/ /IY/
  - 'How old are you'-/H/ /AW/-/OW/ /L/ /D/-/AA/ /R/-/Y/ /UW/
  - 'Speech processing is fun'-/S/ /P/ /IY/ /CH/-/P/ /R/ /AH/ /S/ /EH/ /S/ /IH/ /NG/-/IH/ /Z/-/F/ /AH/ /N/

- word ***ambiguity*** abounds
  - 'lives'-/L/ /IH/ /V/ /Z/ (he lives here) versus /L/ /AY/ /V/ /Z/ (a cat has nine lives)
  - 'record'-/R/ /EH/ /K/ /ER/ /D/ (he holds the world record) versus /R/ /IY/ /K/ /AW/ /D/ (please record my favorite show tonight)

She had your dark suit in …



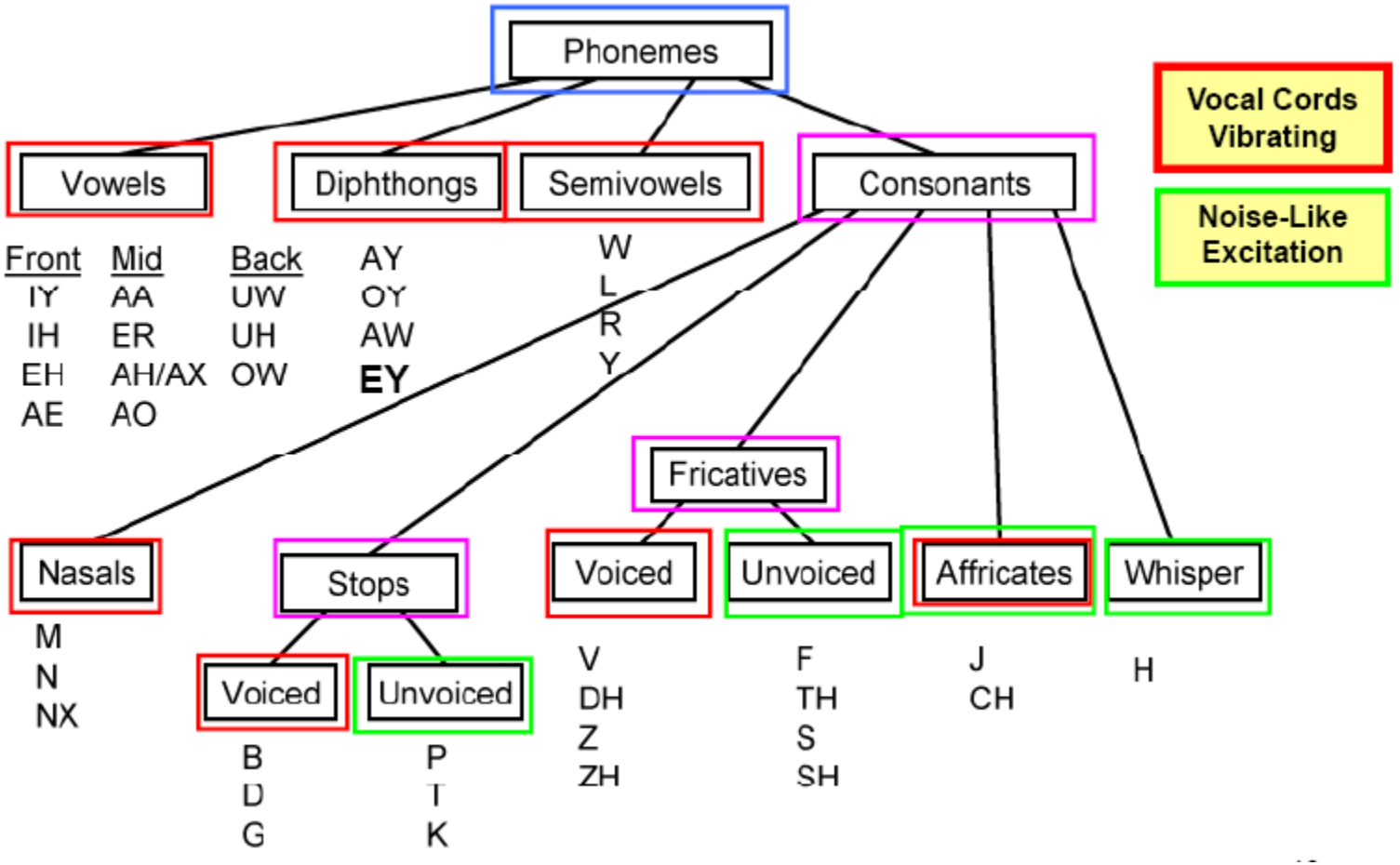
Wideband Spectrogram



Reduced Set of English Sounds

- Total 39 sounds
  - 11 vowels (front, mid, back) classification based on tongue hump position
  - 4 diphthongs (vowel-like combinations)
  - 4 semi-vowels (liquids and glides)
  - 3 nasal consonants
  - 6 voiced and unvoiced stop consonants
  - 8 voiced and unvoiced fricative consonants
  - 2 affricate consonants
  - 1 whispered sound
- look at each ***class of sounds*** to characterize their acoustic and spectral properties

Phoneme Classification Chart



Vowels

- longest duration sounds – least context sensitive
- can be held indefinitely in singing and other musical works (opera)
- carry very little linguistic information (some languages don't display vowels in text-Hebrew, Arabic)

Text 1: all vowels deleted

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 _nd m_n_g_m_nt.

Text 2: all consonants deleted

A__i_u_e_ _o_a__ _a_ __a_e_ e__e__ia___ __e _a_e, _i__ __e __i_e_ o_ o__u_a_io_a_ e___o_ee_ __i_____ _e__ea_i__.

Vowels and Consonants

Text 1: all vowels deleted

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 _nd m_n_g_m_nt.

(They noted significant improvements in the company's
image, supervision and management.)

Text 2: all consonants deleted

A__i_u_e_ _o_a__ _a_ __a_e_ e__e__ia___ __e _a_e,
_i__ __e __i_e_ o_ o__u_a_io_a_ e___o_ee_ __i_____
_e__ea_i__.

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

More Textual Examples

Text (all vowels deleted):

_n th_ n_xt f_w d_c_d_s, _dv_nc_s _n c_mm_n_c_t_ _ns w_ll r_d_c_lly ch_ng_ th_ w_y w_ l_v_ _nd w_rk.

Text (all consonants deleted):

_ _e _o_ _e_ _ o_ _oi_ _ _o _o_ _ _i_ _ _ _a_ _e _ _o_ _o_ _u_i_ _ …

More Textual Examples

Text (all vowels deleted):

_n th_ n_xt f_w d_c_d_s, _dv_nc_s _n c_mm_n_c_t_ _ns w_ll r_d_c_lly ch_ng_ th_ w_y w_ l_v_ _nd w_rk.

(In the next few decades, advances in communications will radically change the way we live and work.)

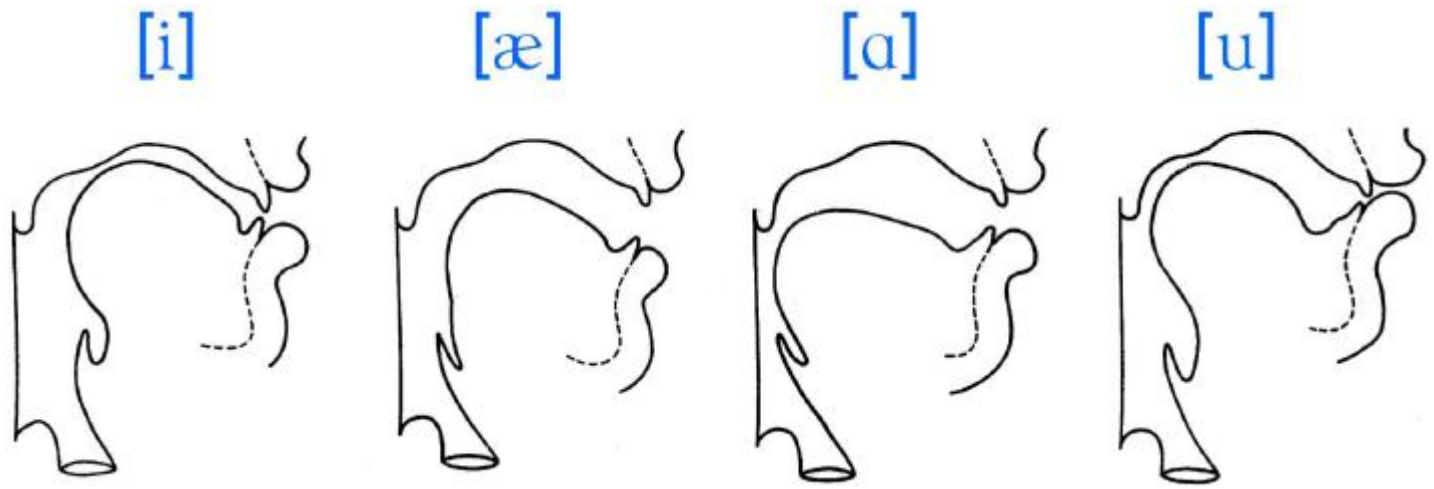
Text (all consonants deleted):

_ _e _o_ _e_ _ o_ _oi_ _ _o _o_ _ _i_ _ _ _a_ _e _ _o_ _o_ _u_i_ _ …

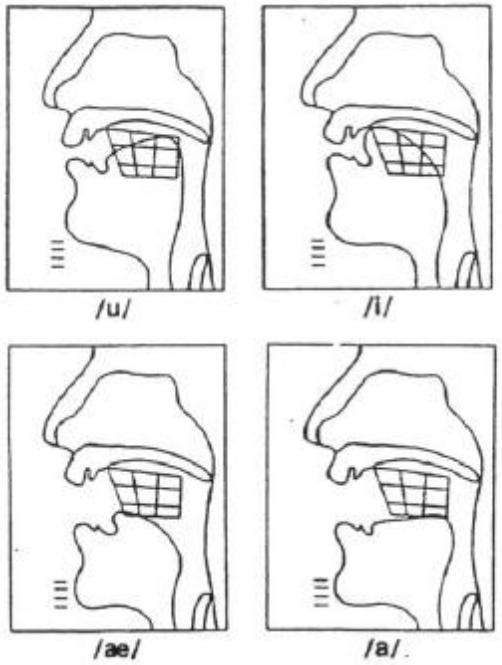
(The concept of going to work will change from commuting…)

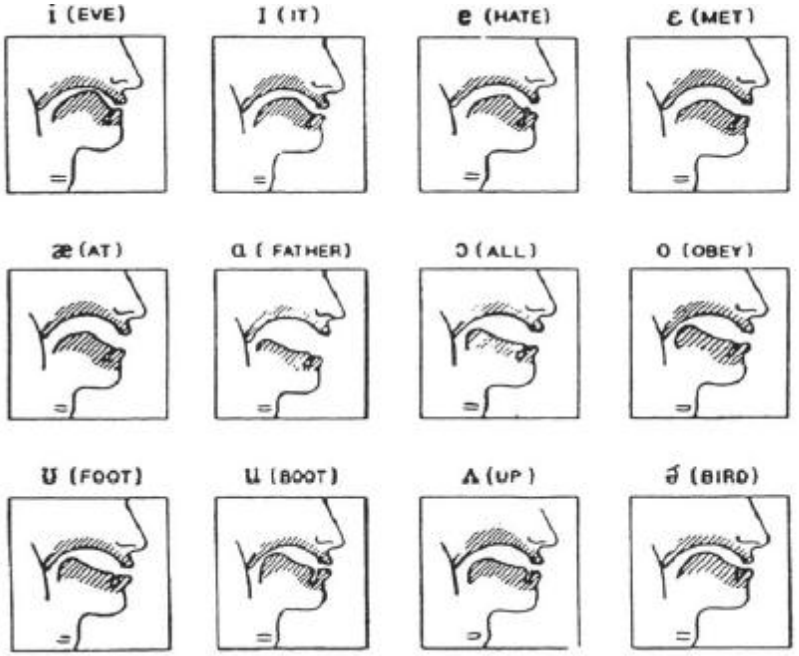
Vowel Production

- produced using ***fixed vocal tract shape***
- ***vocal cords are vibrating*** ➔ voiced sounds
- ***cross-sectional area*** of vocal tract determines vowel resonance frequencies and vowel sound quality
- ***tongue position*** (height, forward/back position) most important in determining vowel sound
- usually relatively ***long in duration*** (can be held during singing) and are spectrally well formed



Vowel Articulatory Shapes



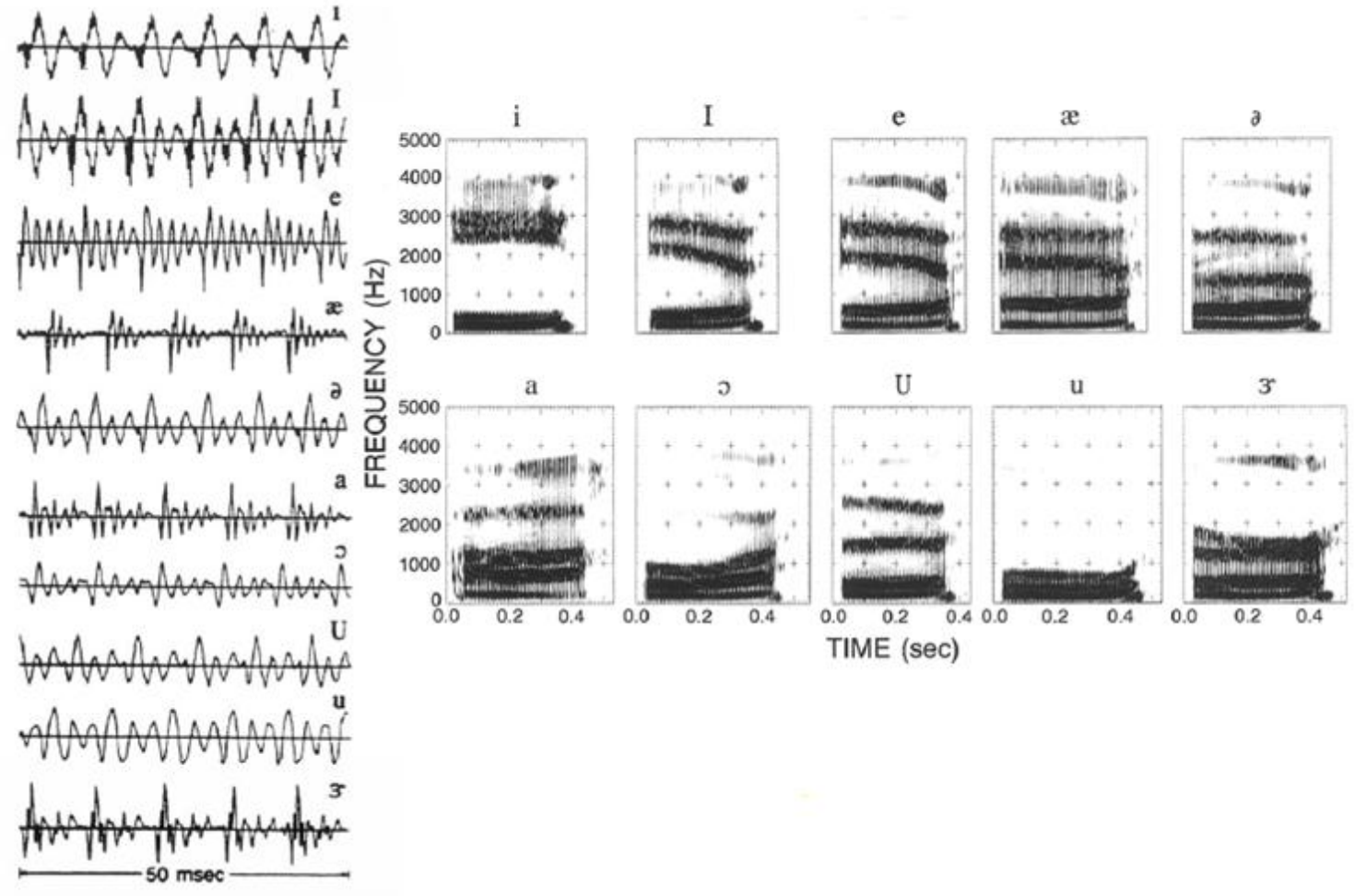


TONGUE POSITION

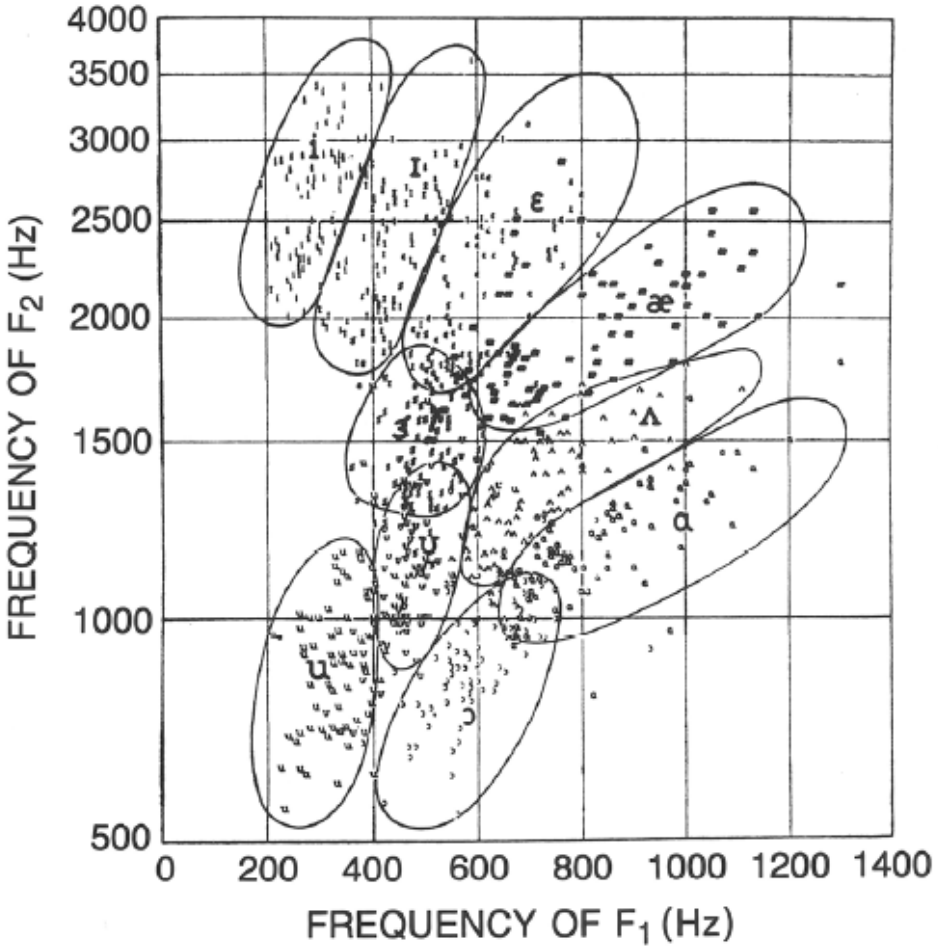
TONGUE HEIGHT	FRONT	BACK
HIGH	1. i	
MID	2. I	7 u
	3. ɛ	6 U
LOW	4. ae	5 a

- tongue hump position (front, mid, back)
- tongue hump height (high, mid, low)
- /IY/, /IH/, /AE/, /EH/ => front => high resonances
- /AA/, /AH/, /AO/ => mid => energy balance
- /UH/, /UW/, /OW/ => back => low frequency resonances

Vowel Waveforms & Spectrograms



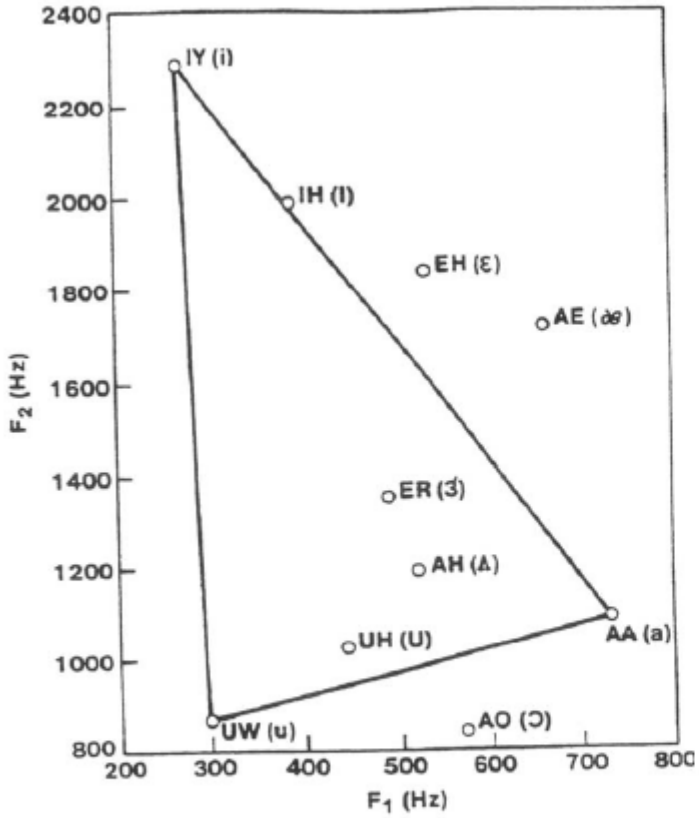
Vowel Formants



Clear pattern of variability of vowel pronunciation among men, women and children

Strong overlap for different vowel sounds by different talkers => no unique identification of vowel strictly from resonances => need context to define vowel sound

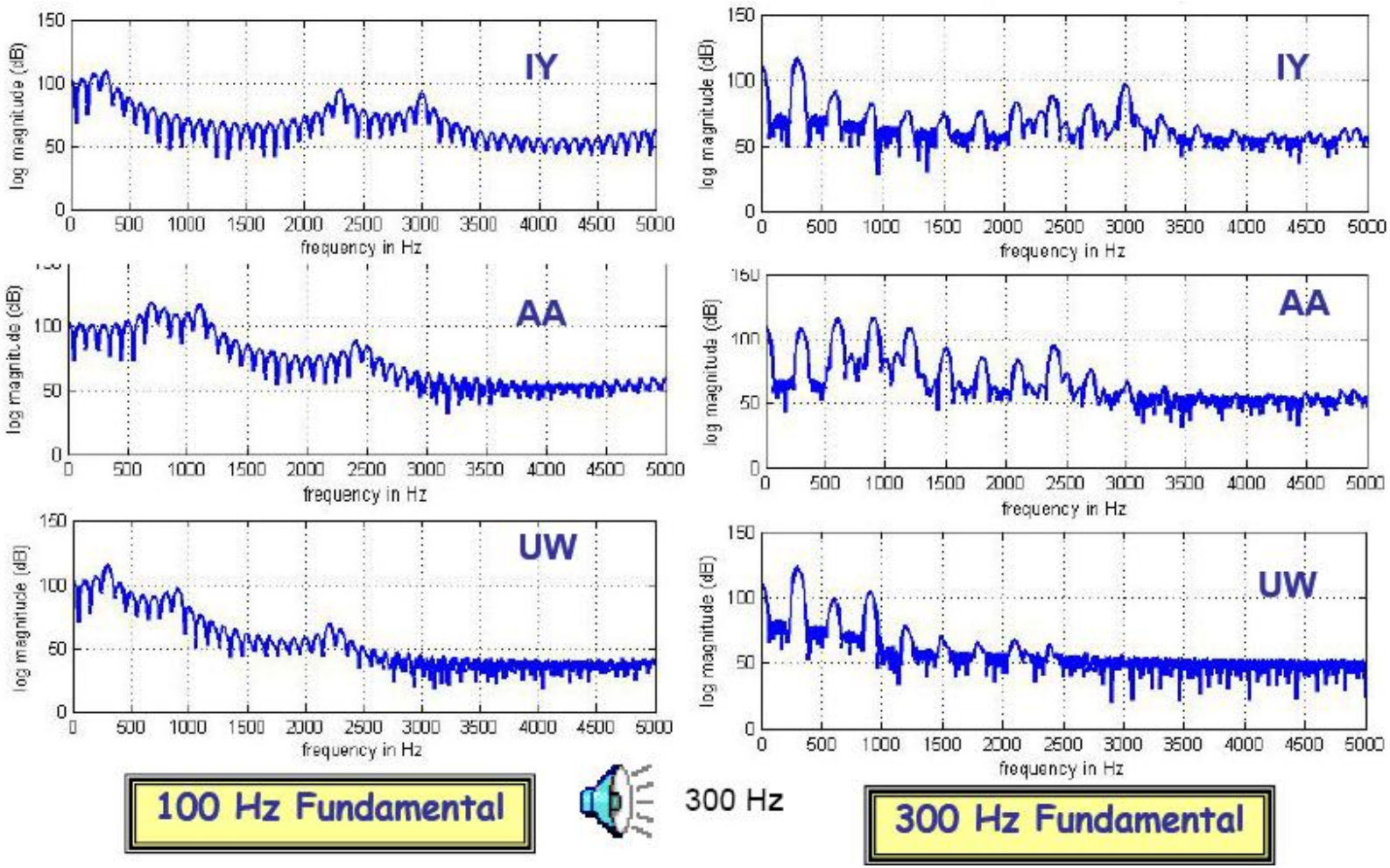
The Vowel Triangle



FORMANT FREQUENCIES FOR THE VOWELS					
Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	i	(beet)	270	2290	3010
IH	ɪ	(bit)	390	1990	2550
EH	ɛ	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
AH	ʌ	(but)	520	1190	2390
AA	ɑ	(hot)	730	1090	2440
AO	ɔ	(bought)	570	840	2410
UH	ʊ	(foot)	440	1020	2240
UW	u	(boot)	300	870	2240
ER	ɝ	(bird)	490	1350	1690

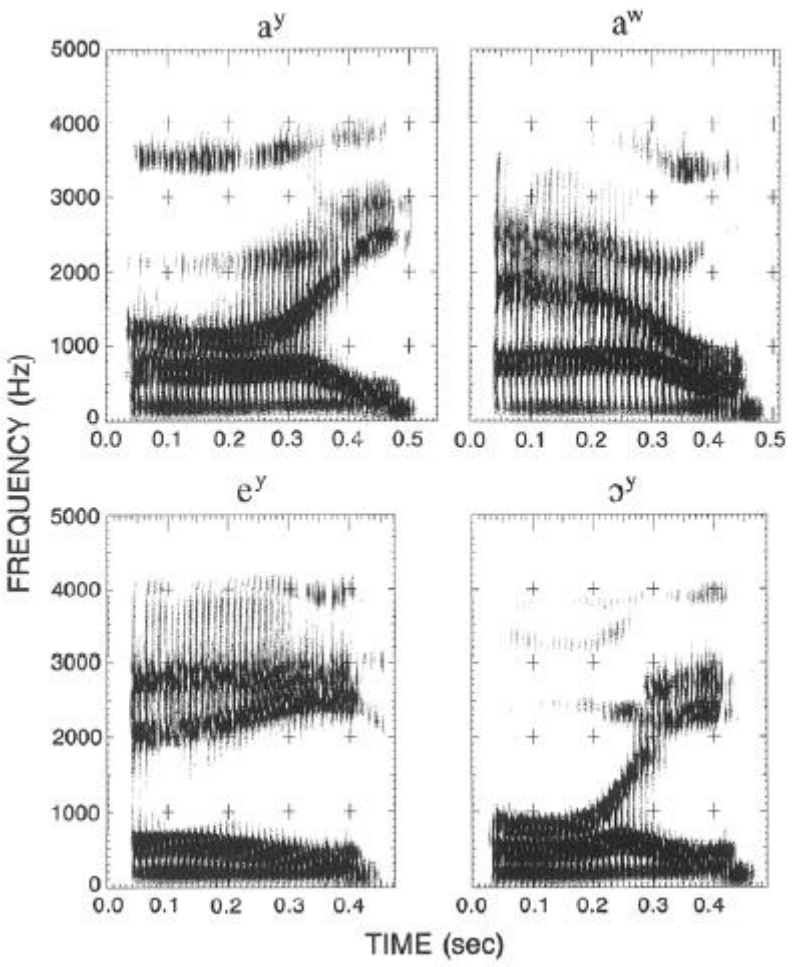
Centroids of common vowels form clear triangular pattern in F1-F2 space

Canonic Vowel Spectra



Diphthongs

- Gliding speech sound from one vowel to or toward another vowel
  - /AY/ in buy
  - /AW/ in down
  - /EY/ in bait
  - /OY/ in boy
  - /OW/ in boat (usually classified as vowel, not diphthong)
  - /Y/ in you (usually classified as glide)



Distinctive Features

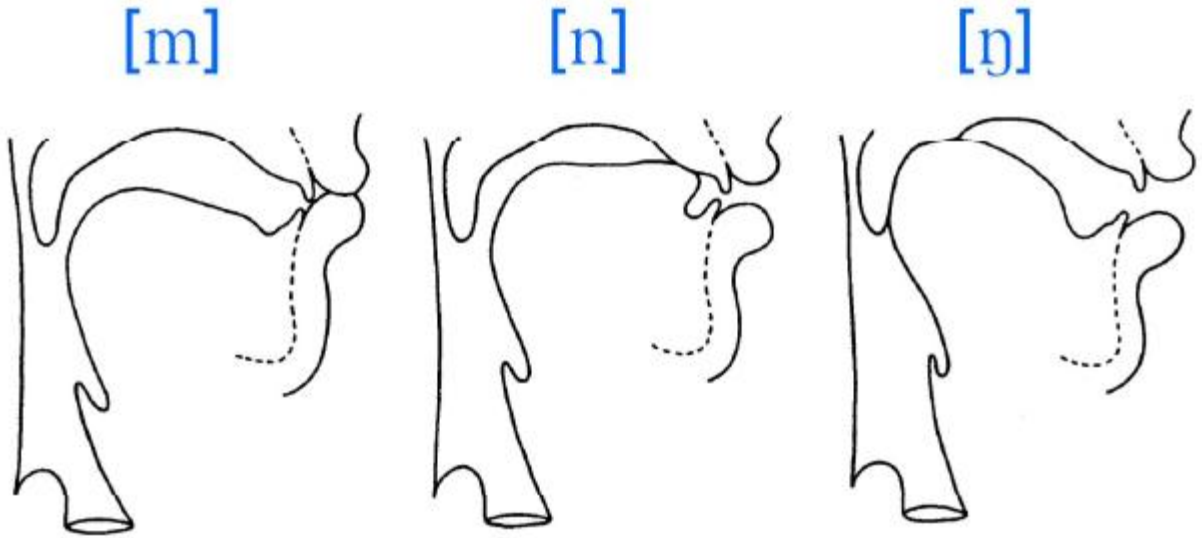
- Classify non-vowel/non-diphthong sounds in terms of distinctive features
  - place of articulation
    - Bilabial (lips)—p,b,m,w
    - Labiodental (between lips and front of teeth)-f,v
    - Dental (teeth)-th,dh
    - Alveolar (front of palate)-t,d,s,z,n,l
    - Palatal (middle of palate)-sh,zh,r
    - Velar (at velum)-k,g,ng
    - Pharyngeal (at end of pharynx)-h
  - manner of articulation
    - Glide—smooth motion-w,l,r,y
    - Nasal—lowered velum-m,n,ng
    - Stop—constricted vocal tract-p,t,k,b,d,g
    - Fricative—turbulent source-f,th,s,sh,v,dh,z,zh,h
    - Voicing—voiced source-b,d,g,v,dh,z,zh,m,n,ng,w,l,r
    - Mixed source—both voicing and unvoiced-j,ch
    - Whispered--h

Semivowels (Liquids and Glides)

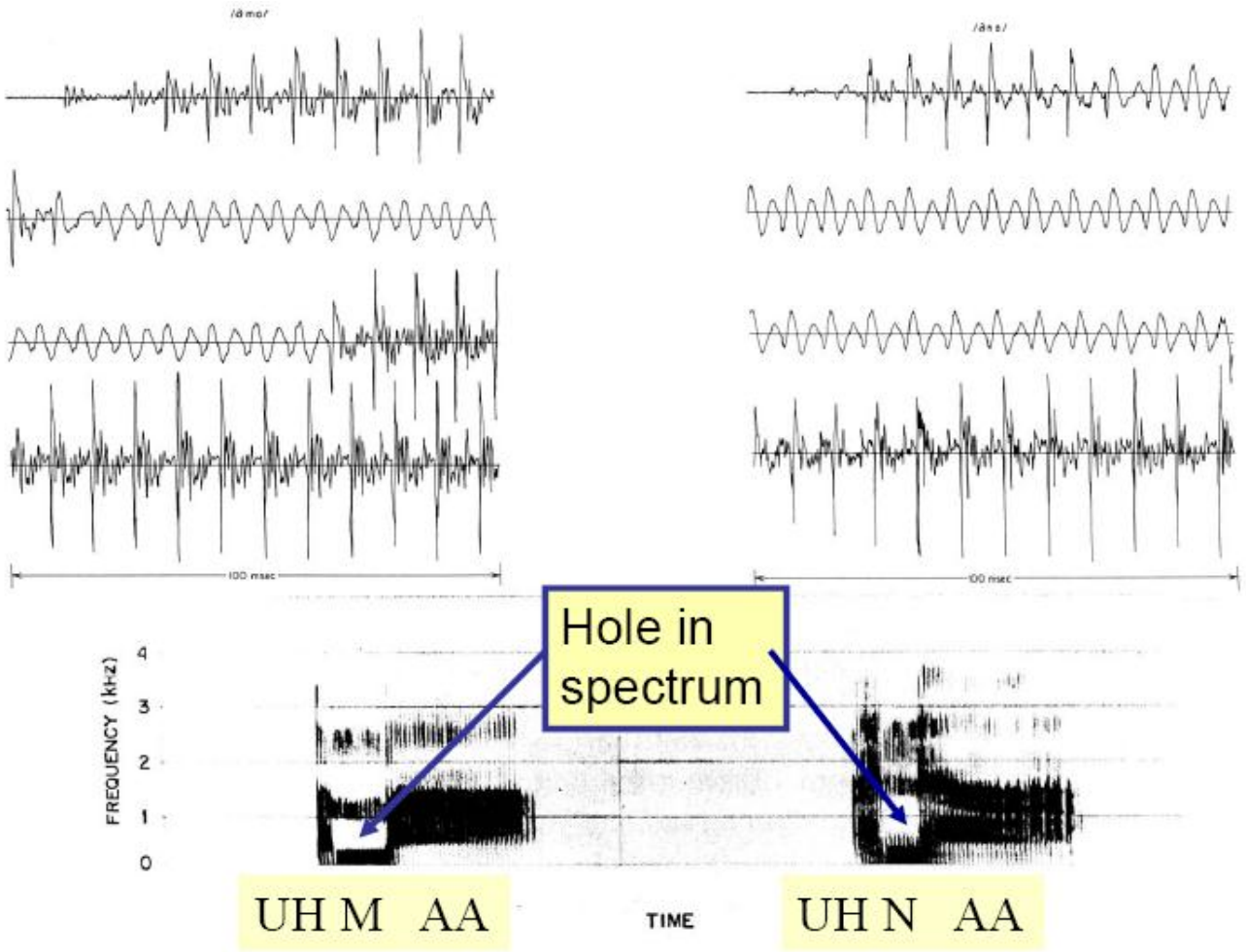
- vowel-like in nature (called semivowels for this reason)
- voiced sounds (w-l-r-y)
- acoustic characteristics of these sounds are strongly influenced by context—unlike most vowel sounds which are much less influenced by context

Nasal Consonants: /M/ /N/ /NG/

- vocal tract totally constricted at some point along the tract
- velum lowered so sound is radiated at nostrils
- constricted oral cavity serves as a resonant cavity that traps acoustic energy at certain natural frequencies (anti-resonances or zeros of transmission)
- /M/ is produced with a constriction at the lips → low frequency zero
- /N/ with a constriction just behind the teeth → higher zero
- /NG/ with a constriction just forward of the velum → even higher zero



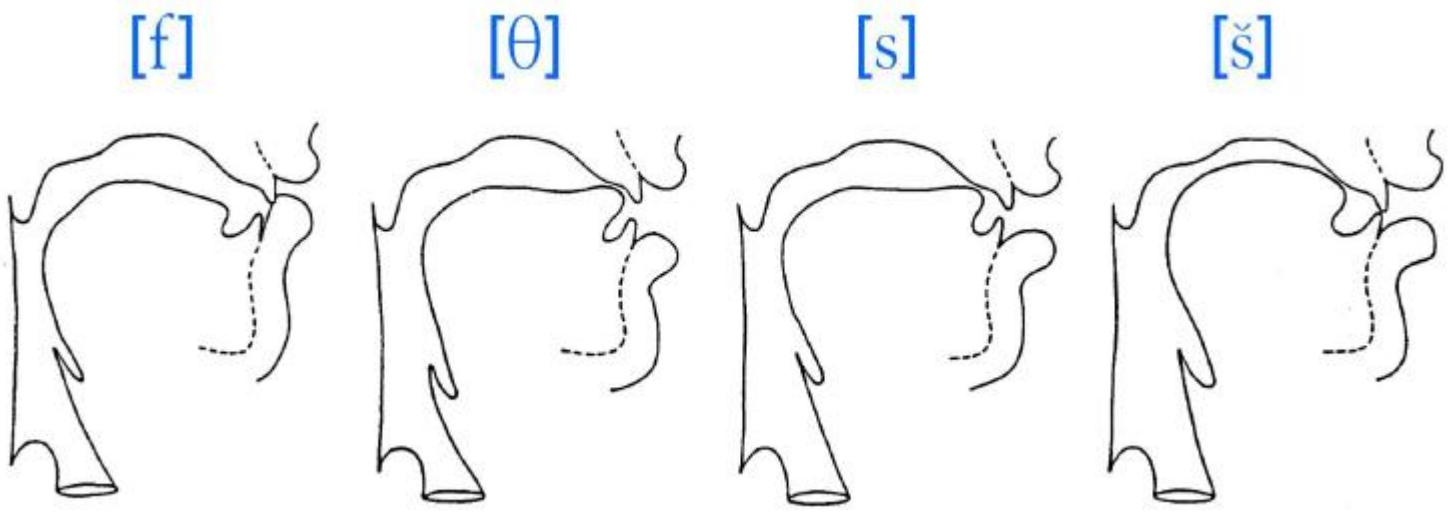
Nasal Sounds



Unvoiced Fricatives

- Consonant sounds /F/, /TH/, /S/, /SH/
  - produced by exciting vocal tract by steady air flow which becomes turbulent in region of a constriction in the vocal tract
    - /F/ constriction near the lips
    - /TH/ constriction near the teeth
    - /S/ constriction near the middle of the vocal tract
    - /SH/ constriction near the back of the vocal tract
  - noise source at constriction => vocal tract is separated into two cavities
  - sound radiated from lips – front cavity
  - back cavity traps energy and produces anti-resonances (zeros of transmission)

Unvoiced Fricative Production



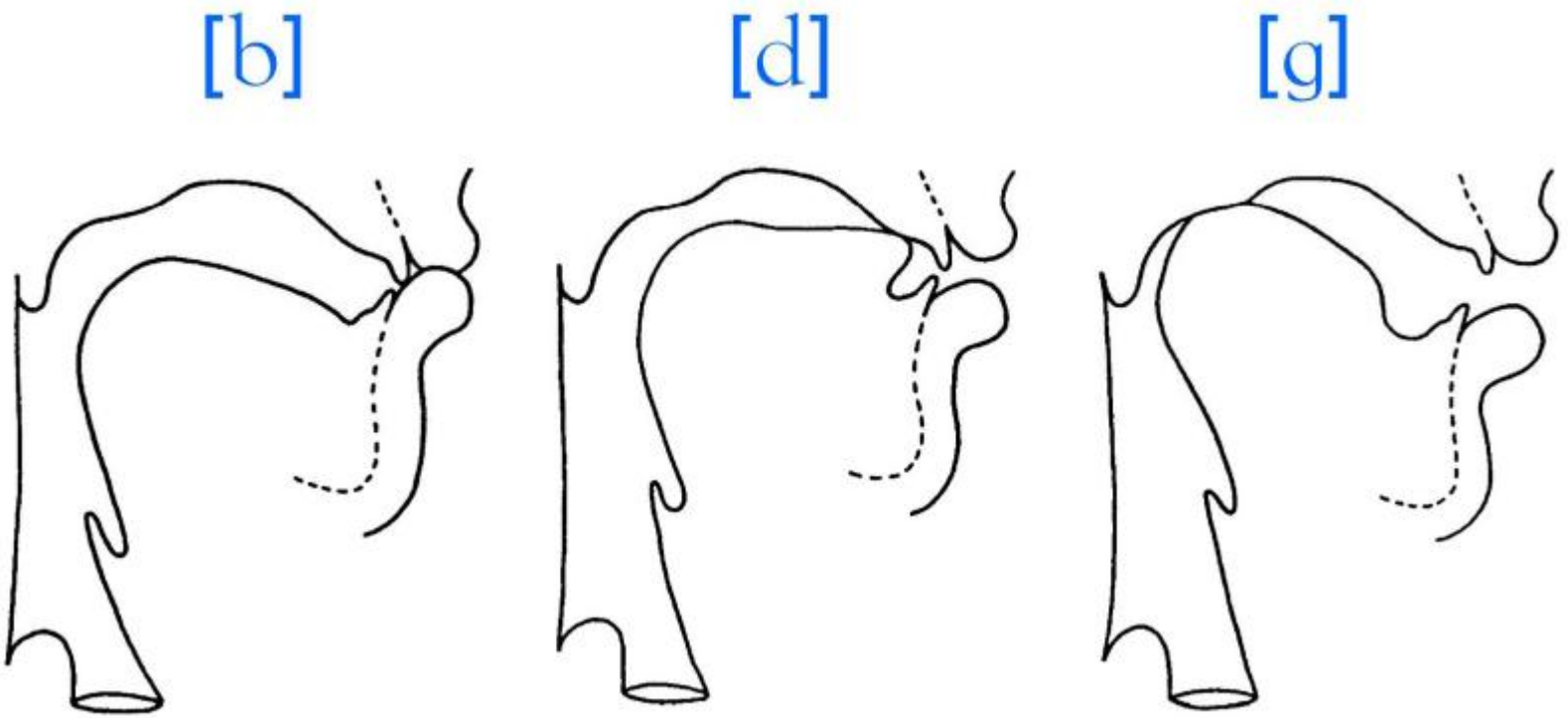
Voiced Fricatives

- Sounds /V/,/DH/, /Z/, /ZH/
  - place of constriction same as for unvoiced counterparts
  - two sources of excitation; vocal cords vibrating producing semi-periodic puffs of air to excite the tract; the resulting air flow becomes turbulent at the constriction giving a noise-like component in addition to the voiced-like component

Voiced and Unvoiced Stop Consonants

- sounds-/B/, /D/, /G/ (voiced stop consonants) and /P/, /T/, /K/ (unvoiced stop consonants)
  - voiced stops are transient sounds produced by building up pressure behind a total constriction in the oral tract and then suddenly releasing the pressure, resulting in a pop-like sound
    - /B/ constriction at lips
    - /D/ constriction at back of teeth
    - /G/ constriction at velum
  - no sound is radiated from the lips during constriction → sometimes sound is radiated from the throat during constriction (leakage through tract walls) allowing vocal cords to vibrate in spite of total constriction
  - stop sounds strongly influenced by surrounding sounds
  - unvoiced stops have no vocal cord vibration during period of closure → brief period of frication (due to sudden turbulence of escaping air) and aspiration (steady air flow from the glottis) before voiced excitation begins

Stop Consonant Production



Distinctive Phoneme Features

	p	k	t	b	d	g	f	thin	s	sh	v	the	z	azure	m	n	ng	l	r	w	h
Place																					
bilabial	+	−	−	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−
labiodental	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−
dental	−	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−
alveolar	−	−	+	−	+	−	−	−	+	−	−	−	+	−	−	+	−	+	−	−	−
palatal	−	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	+	−	−
velar	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−
pharyngeal	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+
Manner																					
glide	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	−
nasal	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	−	−	−	−
stop	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
fricative	−	−	−	−	−	−	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−
voicing	−	−	−	+	+	+	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+

FIGURE 17.7 *Binary distinctive feature set of Jakobson et al. From [10].*

- **the brain recognizes sounds by doing a distinctive feature analysis from the information going to the brain**
- **the distinctive features are somewhat insensitive to noise, background, reverberation => they are robust and reliable**

Distinctive Features

Place of articulation	Manner of articulation					
			Stop		Fricative	
	Glide	Nasal	Voiced	Unvoiced	Voiced	Unvoiced
Front						
Bilabial	w,ʍ	m	b	p		
Labiadental					v	f
Middle						
Dental					ð	θ
Alveolar	j,l	n	d	t	z	s
Palatal	r				ʒ	ʃ
Back						
Velar	w, ʍ	ŋ	g	k		
Pharyngeal						h
Glottal			?			

FIGURE 17.8 Articulatory classification of consonants. From [15].

- **place and manner of articulation completely define the consonant**
- **sounds, making speech perception robust to a range of external factors**

Summary

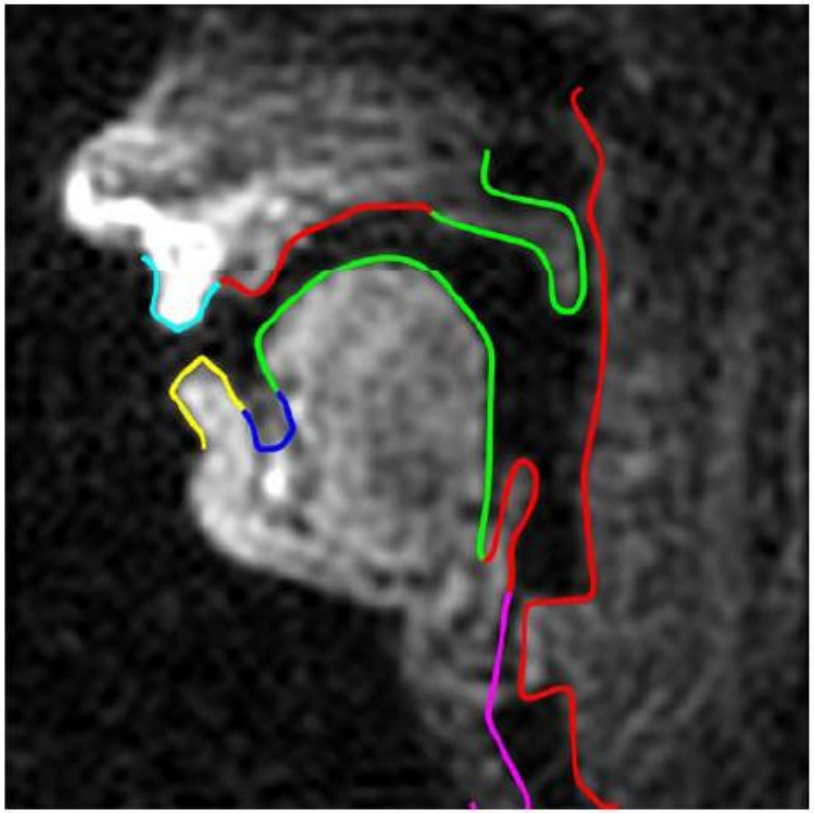
- ***sounds*** of the English language—phonemes, syllables, words
- ***phonetic transcriptions*** of words and sentences — coarticulation across word boundaries
- ***vowels and consonants*** — their roles, articulatory shapes, waveforms, spectrograms, formants
- ***distinctive feature*** representations of speech

ELEC747 Speech Signal Processing

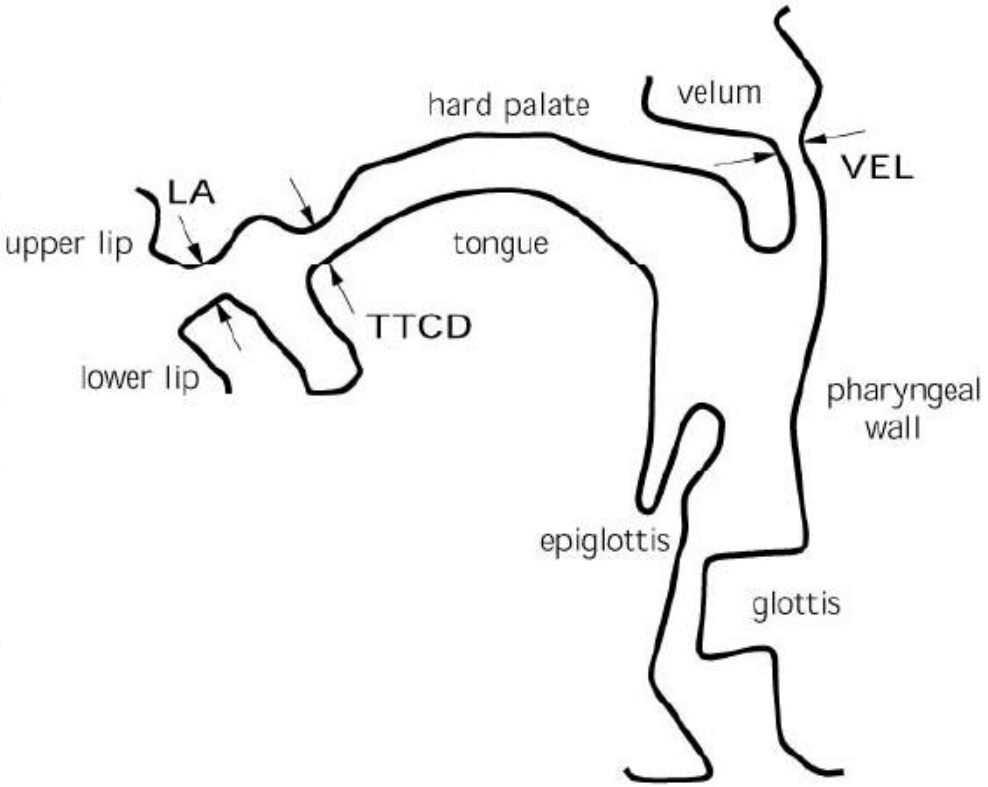
Gil-Jin Jang

END OF CHAPTER 3. FUNDAMENTALS OF HUMAN SPEECH PRODUCTION

MRI of Speech
Shri Narayanan, USC



(a)



(b)

Real Time MRI
Shri Narayanan, USC