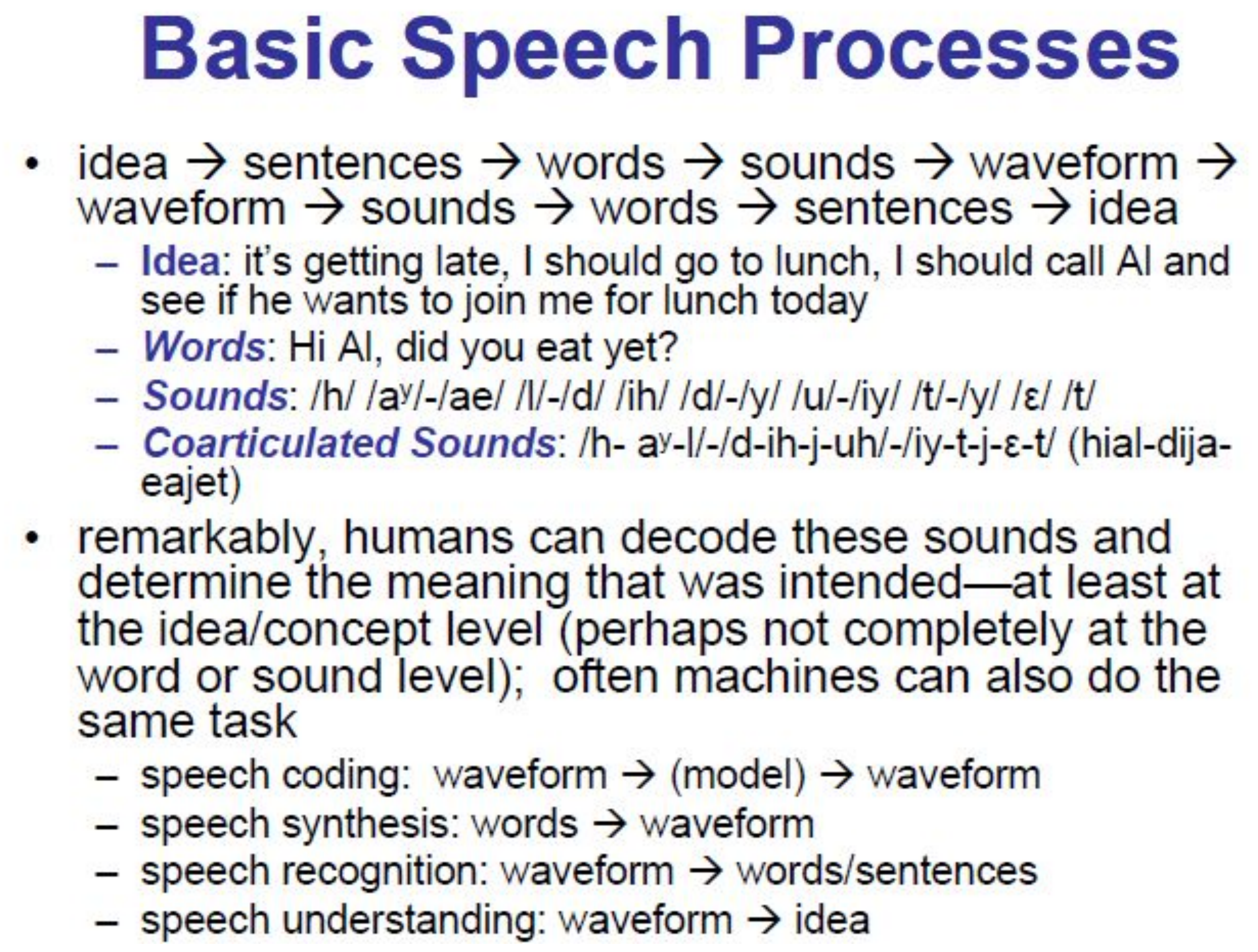
Lecture-4

- ❑ Human Speech production
- ❑ Acoustic Phonetics and Articulatory Phonetics
- ❑ Different categories speech sounds with example
- ❑ Location of sounds in the acoustic waveform and in spectrograms
- ❑ Sound propagation in the human vocal Tract
- ❑Time -varying linear system approaches
- ❑ Source Filter Model
- ❑ Conversion of text to sounds via letter-to-sound rules and PLS

Farther Reading : Chapter-3 of book□Discrete-Time Speech Signal Processing: Principles and Practice by Thomas F. Quatieri



Basic Speech Processes

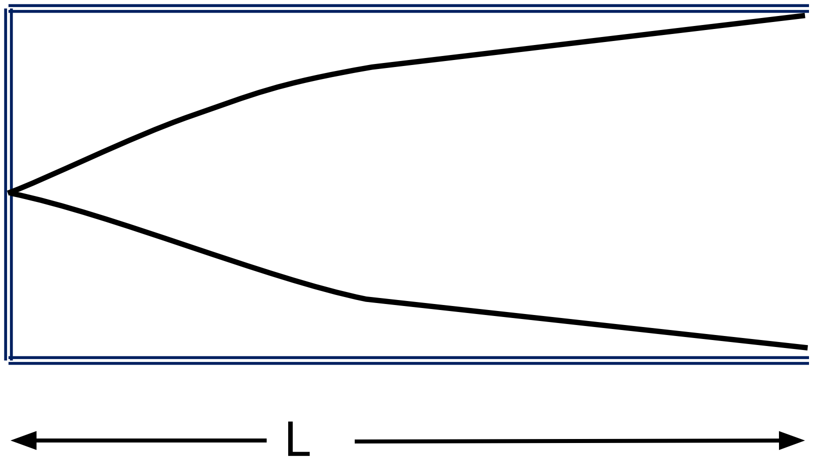
- idea → sentences → words → sounds → waveform → waveform → sounds → words → sentences → idea
  - **Idea**: it's getting late, I should go to lunch, I should call Al and see if he wants to join me for lunch today
  - ***Words***: Hi Al, did you eat yet?
  - ***Sounds***: /h/ /a^y/-/ae/ /l/-/d/ /ih/ /d/-/y/ /u/-/iy/ /t/-/y/ /ε/ /t/
  - ***Coarticulated Sounds***: /h- a^y-l/-/d-ih-j-uh/-/iy-t-j-ε-t/ (hial-dija-eajet)
- 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
  - speech coding: waveform → (model) → waveform
  - speech synthesis: words → waveform
  - speech recognition: waveform → words/sentences
  - speech understanding: waveform → idea

Basics

- ***speech*** is composed of a sequence of sounds
- ***sounds*** (and transitions between them) serve as a symbolic representation of information to be shared between humans (or humans and machines)
- arrangement of sounds is governed by rules of ***language*** (constraints on sound sequences, word sequences, etc)--/spl/ exists, /sbk/ doesn’t exist
- ***linguistics*** is the study of the rules of language
- ***phonetics*** is the study of the sounds of speech

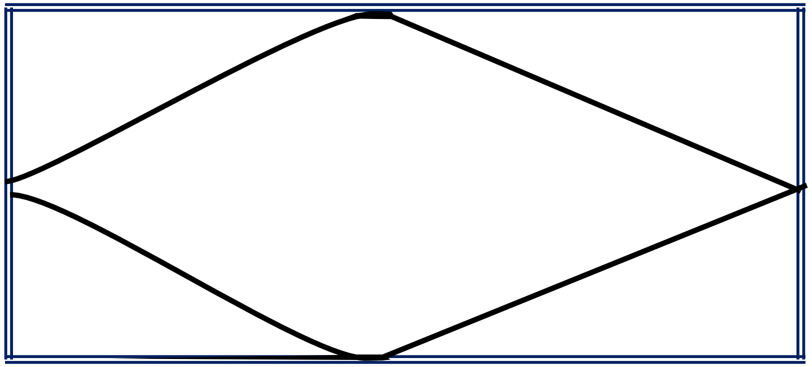
can exploit ***knowledge*** about the structure of sounds and language—and how it is encoded in the signal—to do speech analysis, speech coding, speech synthesis, speech recognition, speaker recognition, etc.

quarter-wavelength-resonances



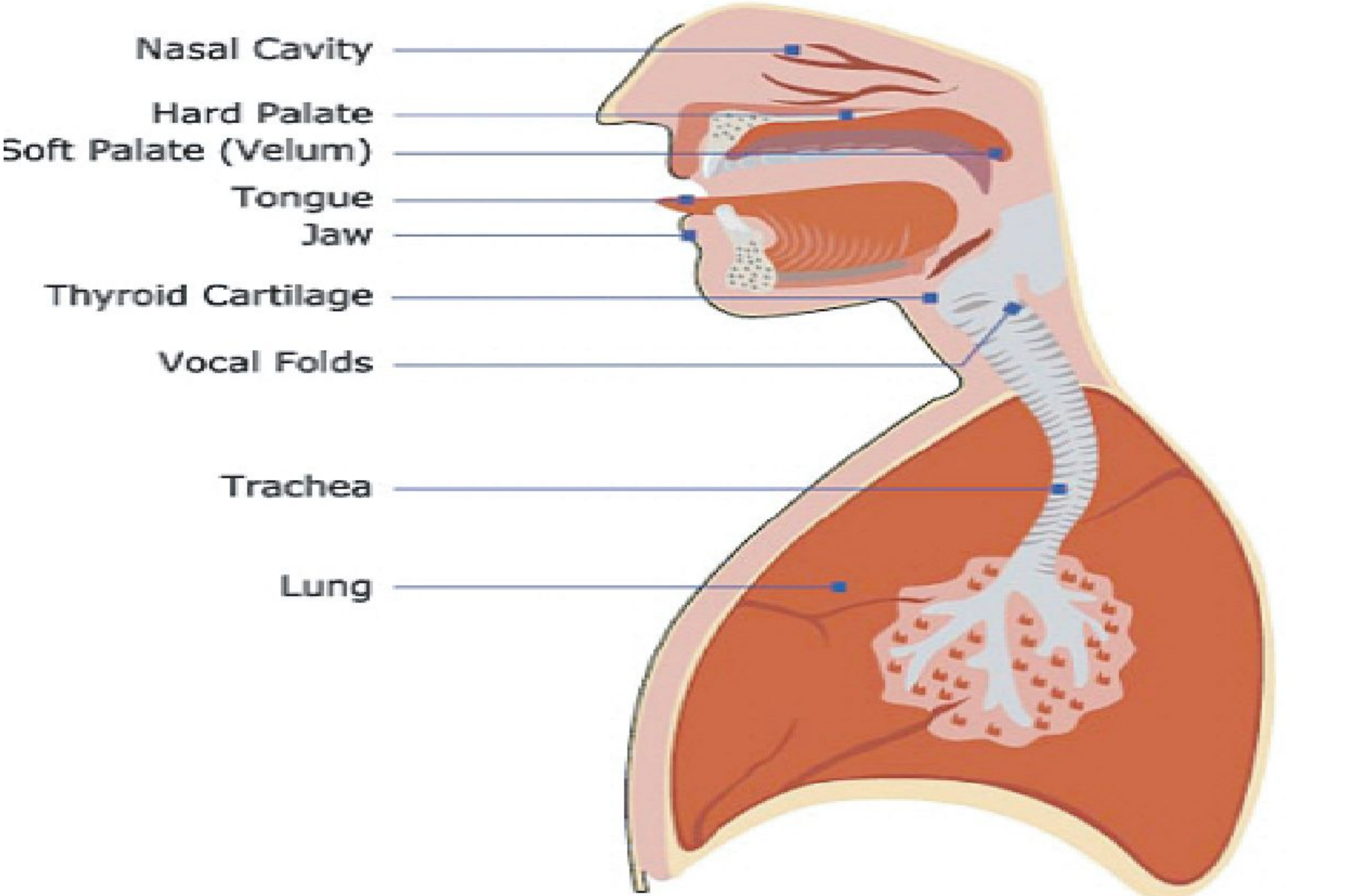
f=(2k+1).c/4L

L□ length of the tube,
k□ a positive integer,
c□ the sound velocity



f=k.c/2L

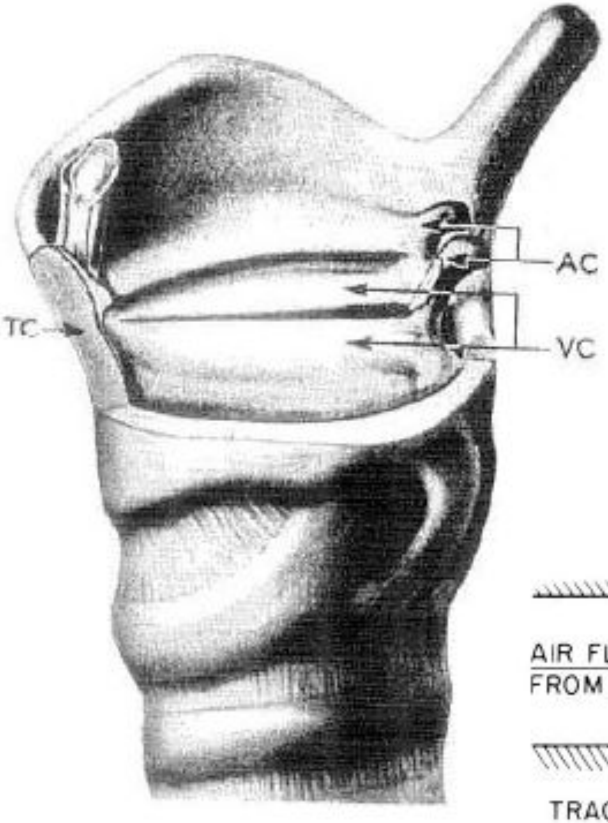
half-wavelength resonances


Nasal Cavity
Hard Palate
Soft Palate (Velum)
Tongue
Jaw
Thyroid Cartilage
Vocal Folds
Trachea
Lung

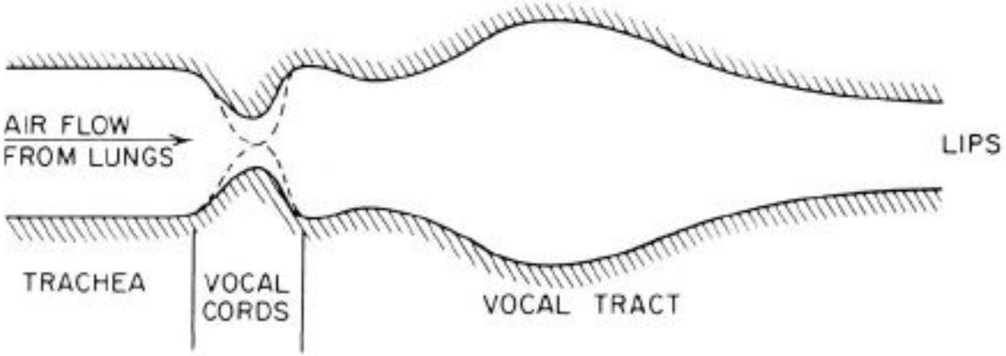
The Respiratory System and Speech:

- **The respiratory system and speech are interconnected:**
  - The average fundamental frequency and intensity increase with higher lung volume initiation levels in untrained voices.
  - The Lombard Effect: The voice naturally raises in intensity level when given the condition of noise. When wearing headphones at a loud level (70 dB), the listener's voice will raise unless the listener consciously controls his volume level.

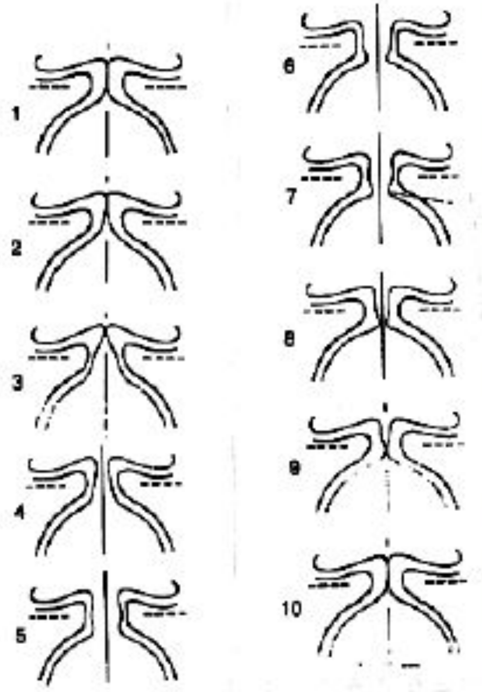
Vocal Cords



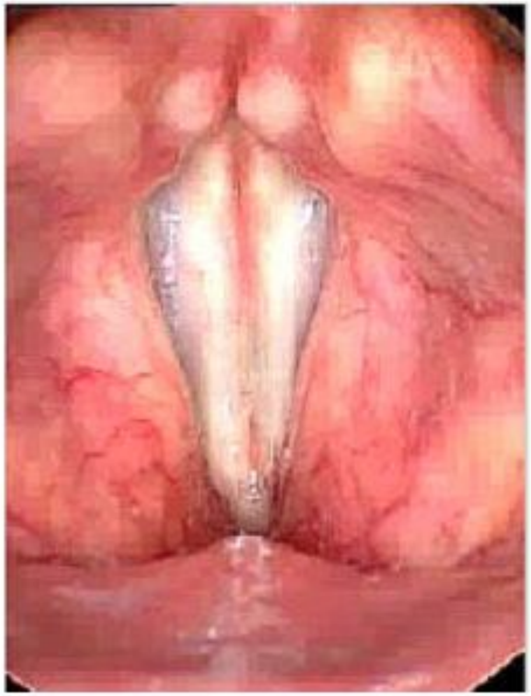
The vocal cords (folds) form a relaxation oscillator. Air pressure builds up and blows them apart. Air flows through the orifice and pressure drops allowing the vocal cords to close. Then the cycle is repeated.



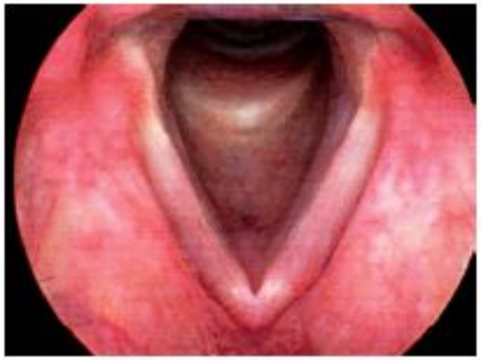
Vocal Cord Views and Operation



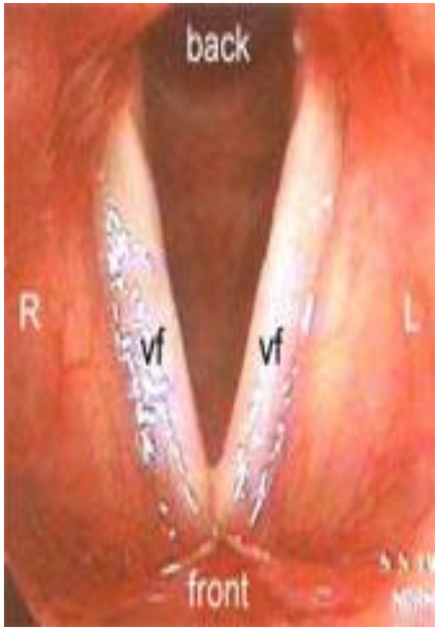
Bernoulli Oscillation

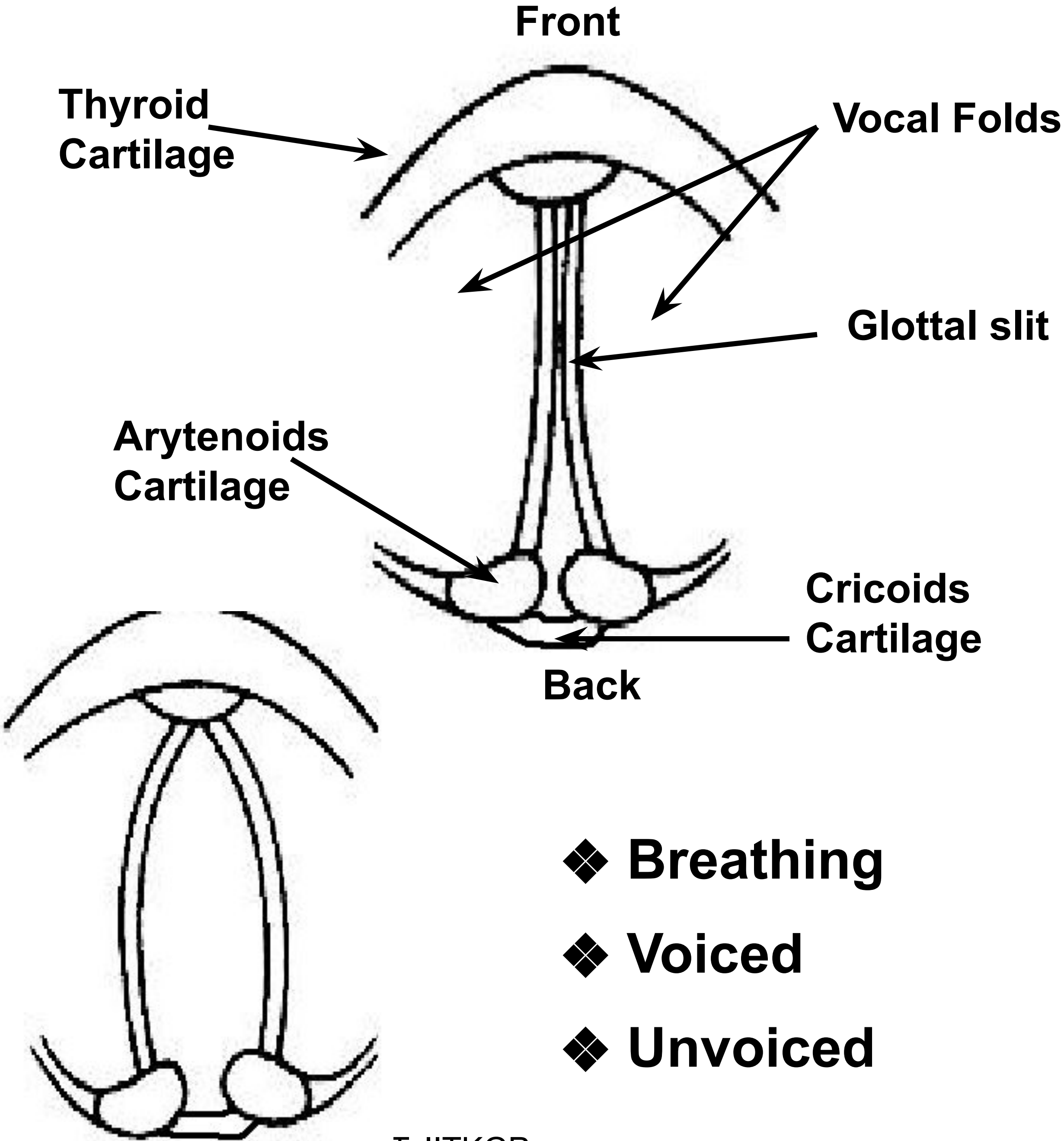


Tensed Vocal Cords – Ready to Vibrate



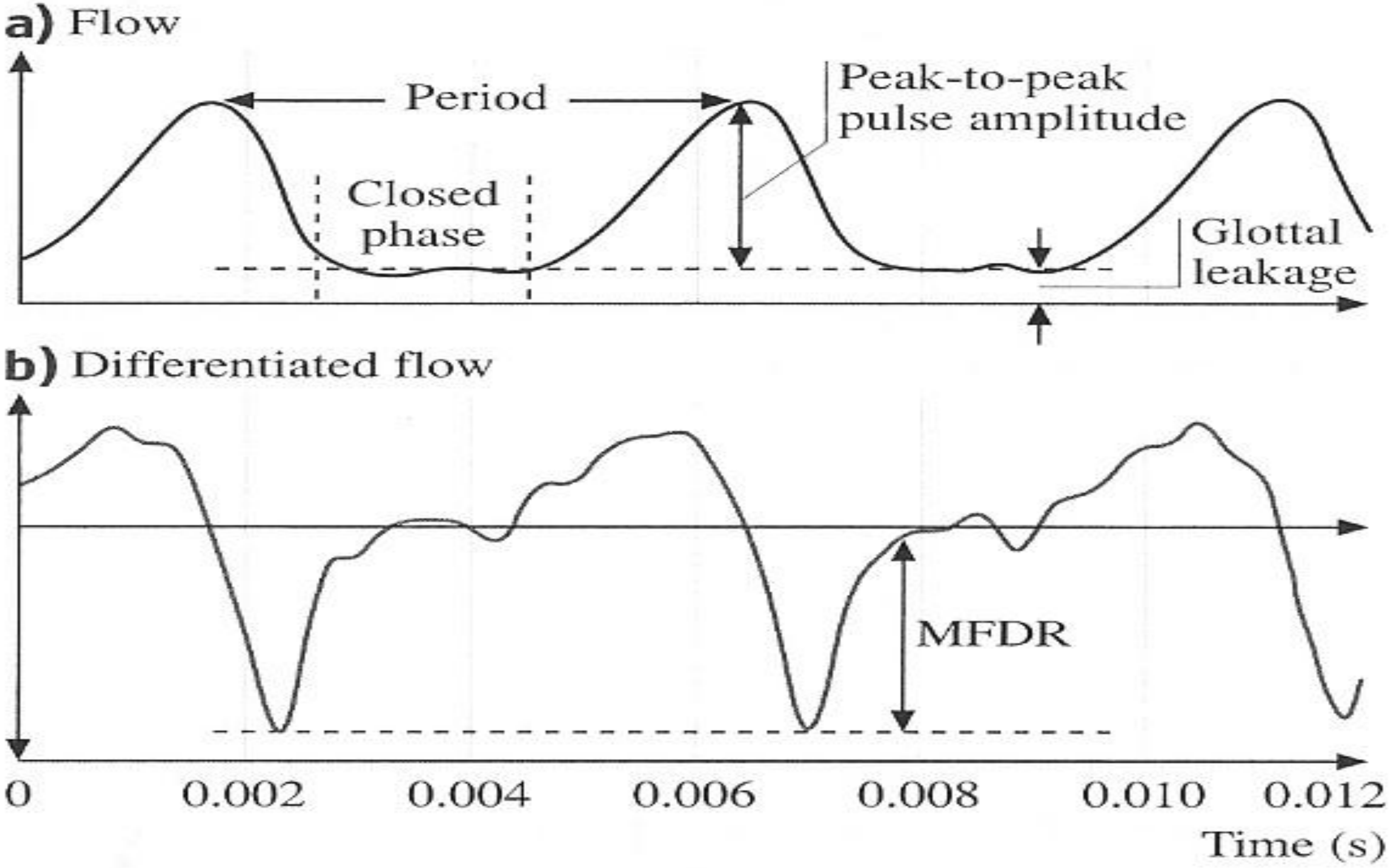
Vocal Cords – Open for Breathing



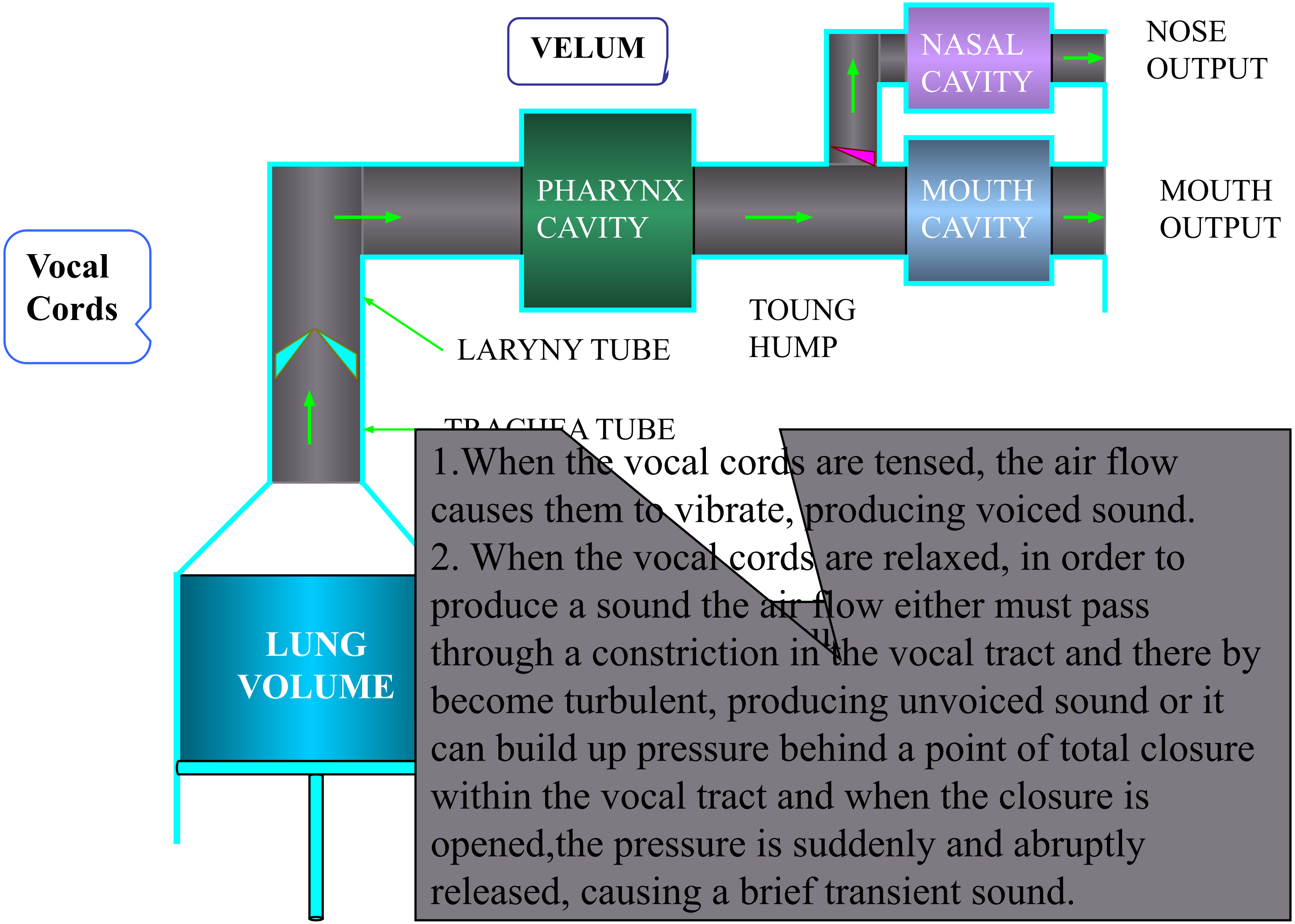


- Breathing
- Voiced
- Unvoiced

GLOTTAL FLOW



Schematic representation of the physiological mechanism of speech production

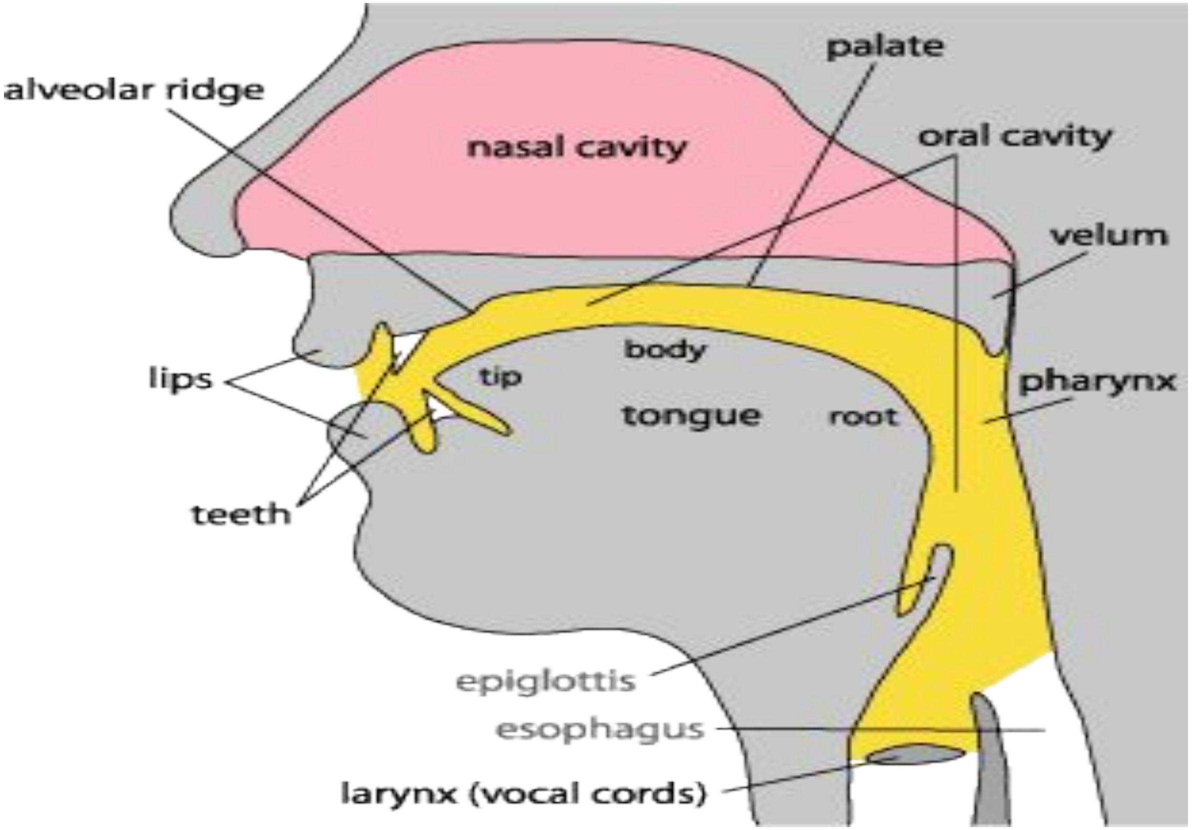


1.When the vocal cords are tensed, the air flow causes them to vibrate, producing voiced sound.

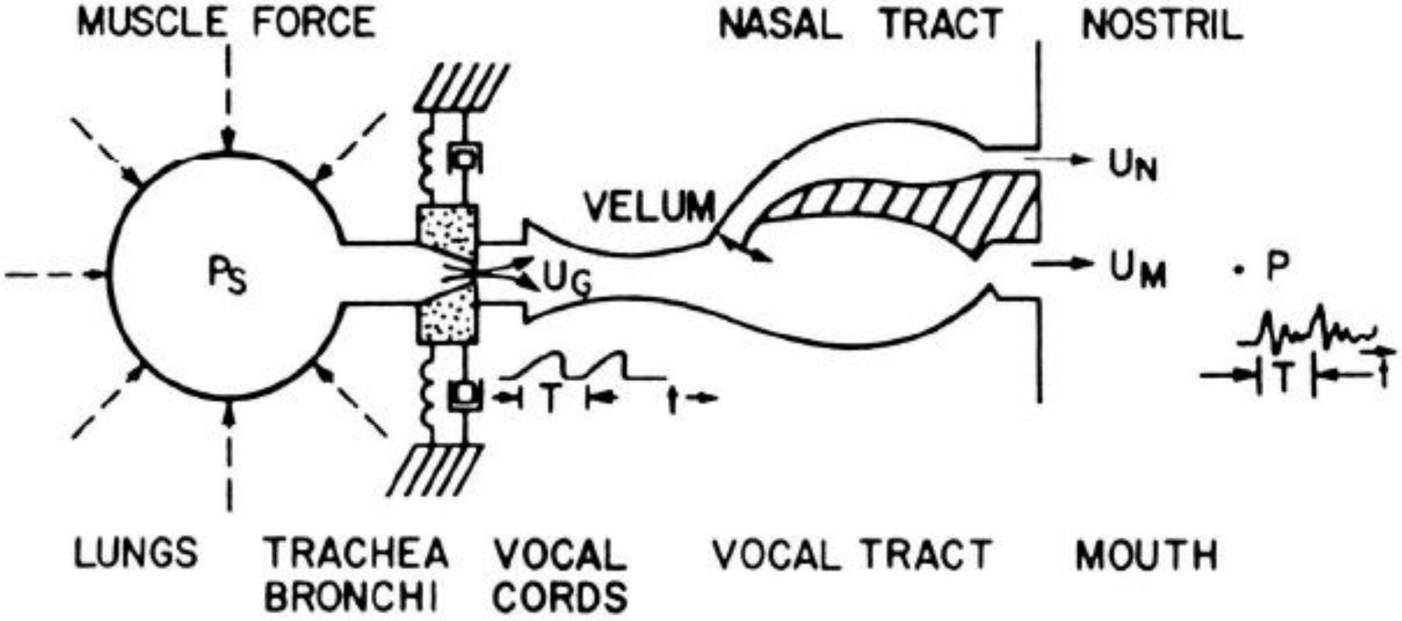
2. When the vocal cords are relaxed, in order to produce a sound the air flow either must pass through a constriction in the vocal tract and there by become turbulent, producing unvoiced sound or it can build up pressure behind a point of total closure within the vocal tract and when the closure is opened,the pressure is suddenly and abruptly released, causing a brief transient sound.

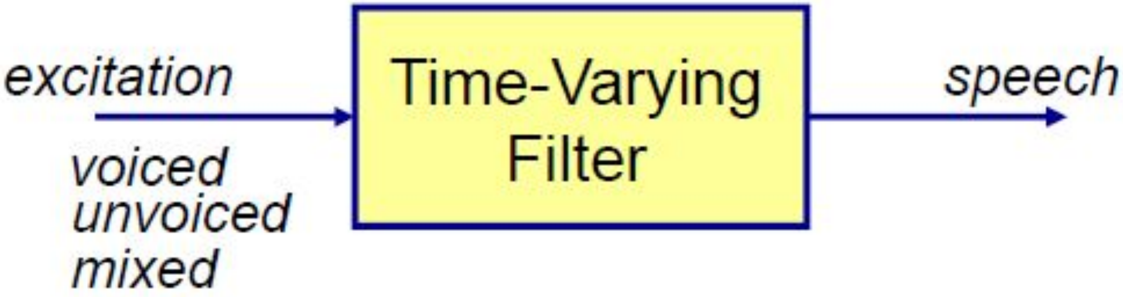
The Vocal Tract

- **The shape of the vocal tract transforms raw sound from the vocal folds into recognizable sounds.**



Abstractions of Physical Model





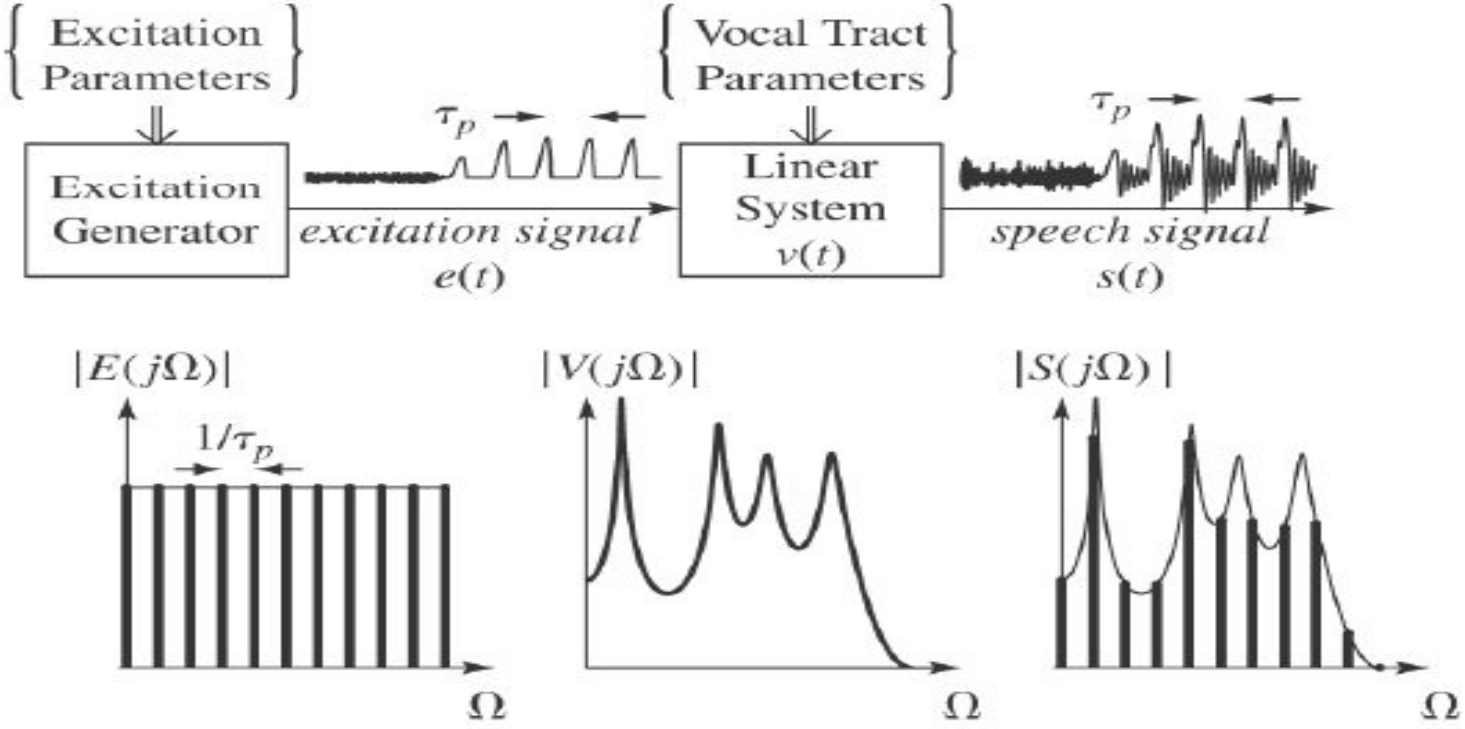
excitation

voiced
unvoiced
mixed

Time-Varying Filter

speech

Source-System Model of Speech Production



Women and Men

- The acoustics of male and female vowels differ reliably along two different dimensions:

1. Sound **Source**
2. Sound **Filter**

- Source--F0: depends on length of vocal folds

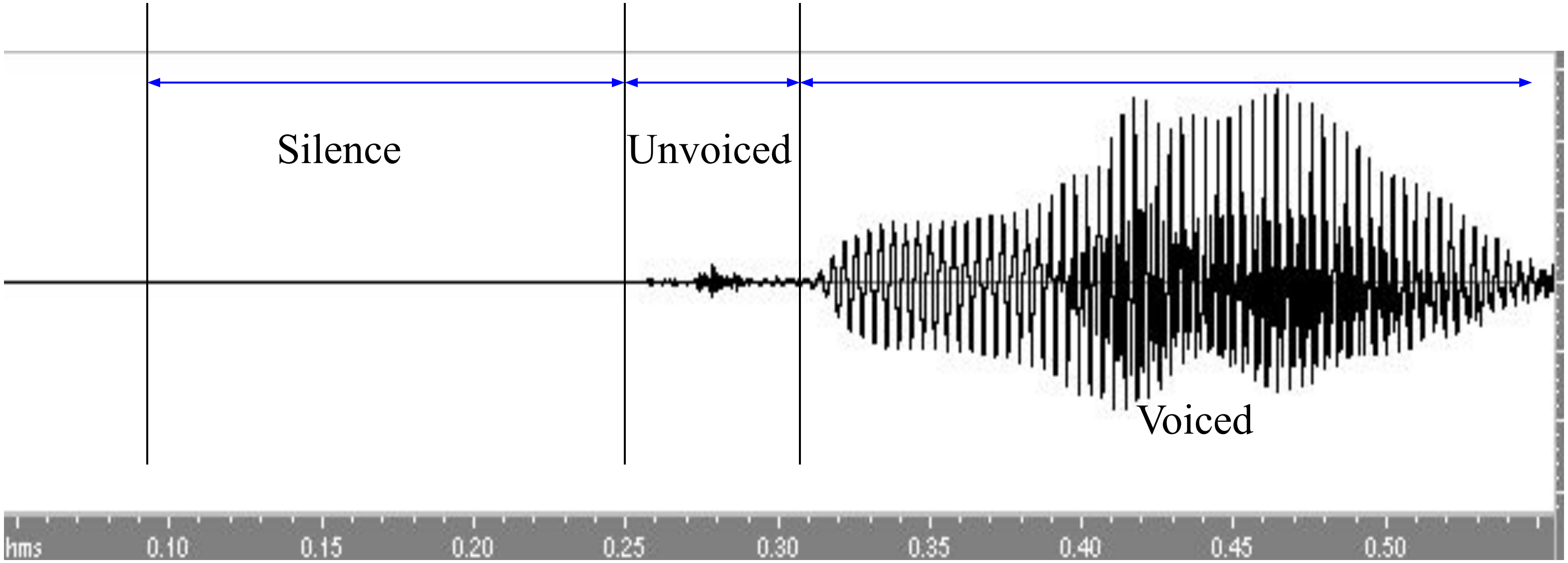
  shorter in women ⇒ higher average F0

  longer in men ⇒ lower average F0

- Filter--Formants: depend on length of vocal tract

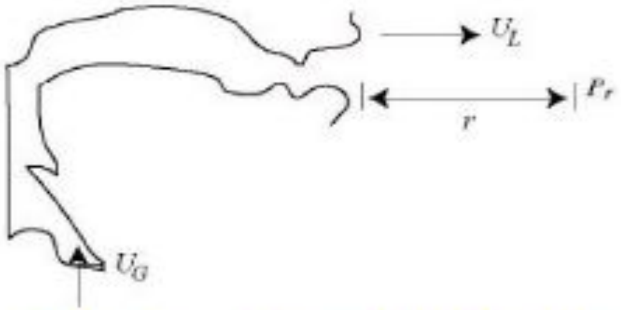
  shorter in women ⇒ higher formant frequencies

  longer in men ⇒ lower formant frequencies


Silence
Unvoiced
Voiced
hms
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50

Acoustic Theory of Speech Production

- The acoustic characteristics of speech are usually modelled as a sequence of source, vocal tract filter, and radiation characteristics



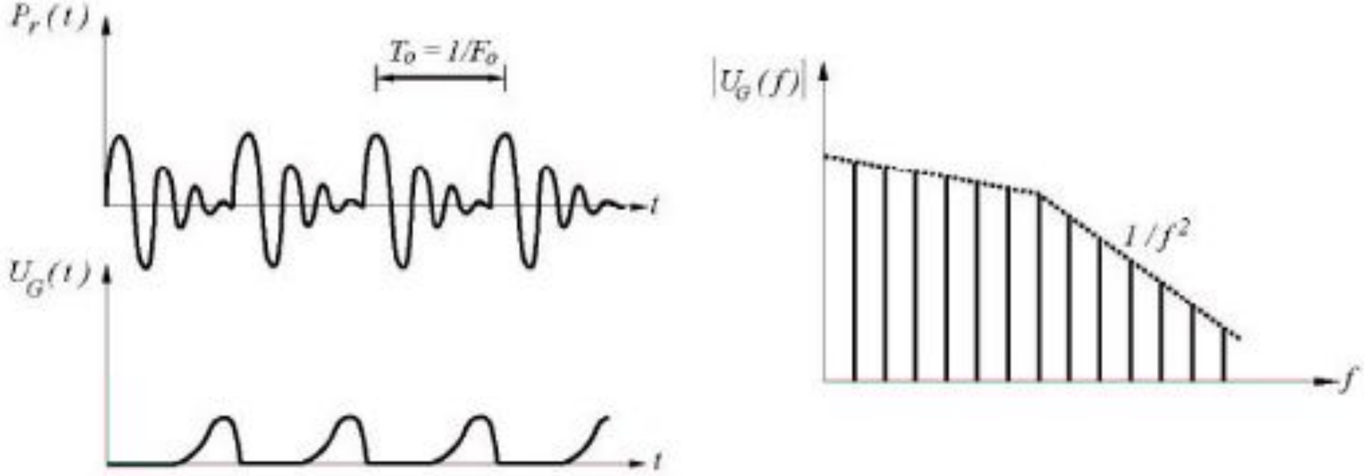
$$P_r(j\Omega) = S(j\Omega)\,T(j\Omega)\,R(j\Omega)$$

- For vowel production:

$$\begin{aligned} S(j\Omega) &= U_G(j\Omega) \\ T(j\Omega) &= U_L(j\Omega)\,/\,U_G(j\Omega) \\ R(j\Omega) &= P_r(j\Omega)\,/\,U_L(j\Omega) \end{aligned}$$

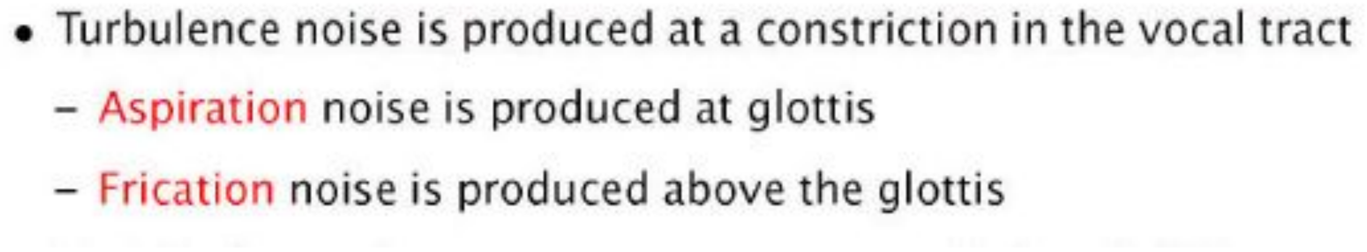
Sound Source for Voiced Sounds

Modelled as a volume velocity source at glottis, $U_G(j\Omega)$



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

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

Parametrization of Spectra

- human vocal tract is essentially a ***tube of varying cross sectional area***, or can be approximated as a ***concatentation of tubes*** of varying cross sectional areas

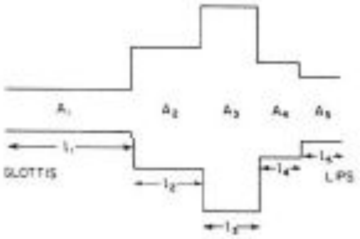


Fig. 3.32 Concatenation of 5 lossless acoustic tubes.

- 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
- resonances known as ***formants*** or ***formant frequencies*** for speech and they represent the frequencies that pass the most acoustic energy from the source to the output
- typically there are ***3 significant formants*** below about 3500 Hz
- formants are a highly efficient, ***compact representation of speech***

Manners and Place of Articulation

- During the articulation the airstreams through the vocal tract must be obstructed in some way. The place where the obstruction takes place is called the place of articulation
- Manner of articulation is concerned with airflow ; the paths it take and the degree to which it is impeded by vocal tract constrictions.

The consonants are classified depending on the place of obstruction and manner of articulation.

ক /*k*/ Velar Un-aspirated unvoiced stop

Vowel sound may be specified in terms of the position of the tongue and the position of the lips.

ই, ঈ /*I*/ High front Un-rounded

Manners of Articulation due to State of the Glottis

If the glottis are closed then it is voiced and if opened then it is unvoiced or voiceless.

Place of articulation

a. **Bilabial:** Bilabial sounds are produced when the two lips make the constriction

b. **Labiodentals:** These sounds are produced by contacting lower lip with the upper teeth.

c. **Dental:** Dental sounds are produced by the constriction of tip or blade of the tongue with the upper teeth.

d. **Alveolar:** The sound made by the tip or the blade of the tongue in contact against the alveolar ridge, which is the bony prominence immediately behind the upper teeth.

e. **Post alveolar:** The sound, which is articulated by the tip or the blade of the tongue with the back area of the alveolar ridge.

f. **Retroflex** : Retroflex sounds are made when the tip of the tongue curled back in the direction of the front part of the hard palate- in other words just behind the alveolar ridge. Depending on how far the tongue curls back, retroflexed could be apico-postalveolar or apico-palatal.

g. **Palatal:** This sound is produced when the constriction is made by the front part of the tongue with the hard palate.

h. **Velar:** It refers to a sound made by the back of the tongue against the soft palate.

i. **Uvular:** This sound is produced when the back of the tongue touches the uvula.

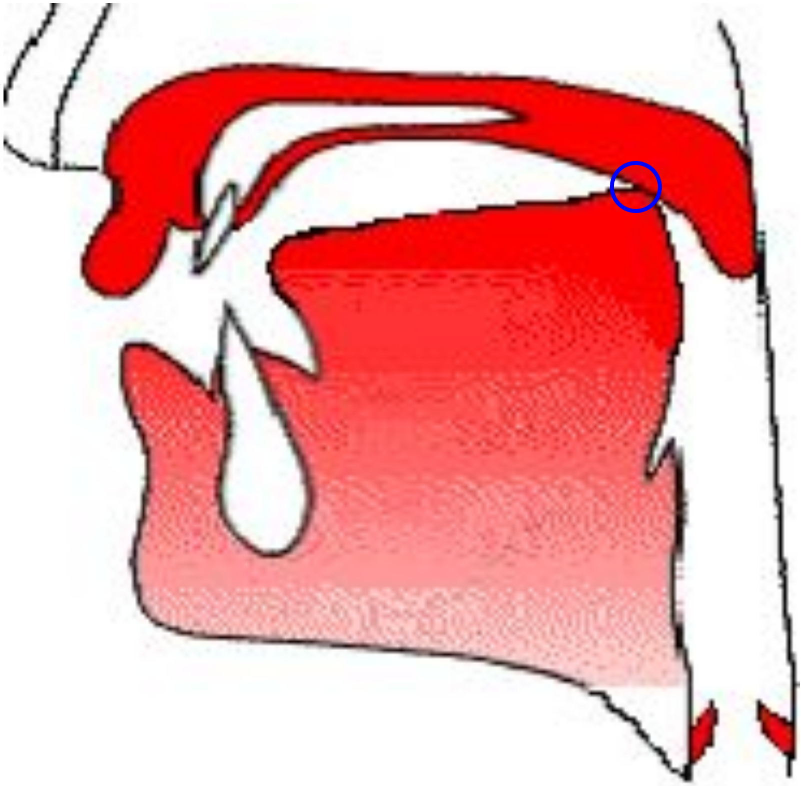
j. **Pharyngeal:** It refers to a sound produced in the pharynx, the tubular cavity, which constitutes the throat above the larynx.

k. **Glottal:** These are the sounds, which made in the larynx due to the closure or narrowing of the glottis.

Manner of articulation

a) Plosive, or oral stop
b) Nasal stop
c) Fricative
d) Affricate
e) Lateral
f) Approximant
g) Trill:
h) Flap and Tap

1. Voiced
2. Unvoiced
3. Aspiration



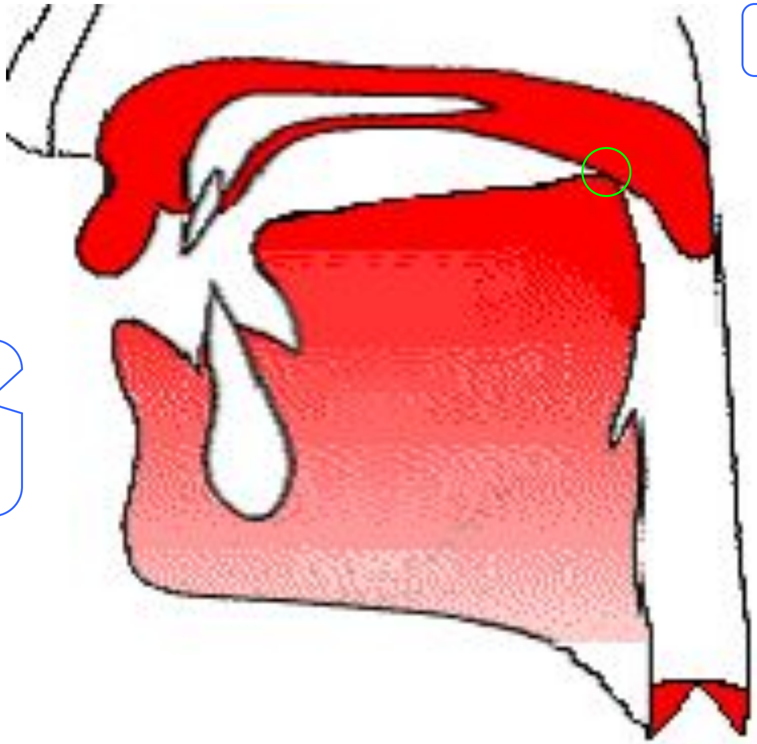
Place of Articulation (Velar)

Back of tongue (Articulator)

Nasal Passage

Velum closed

Vocal Cord Open

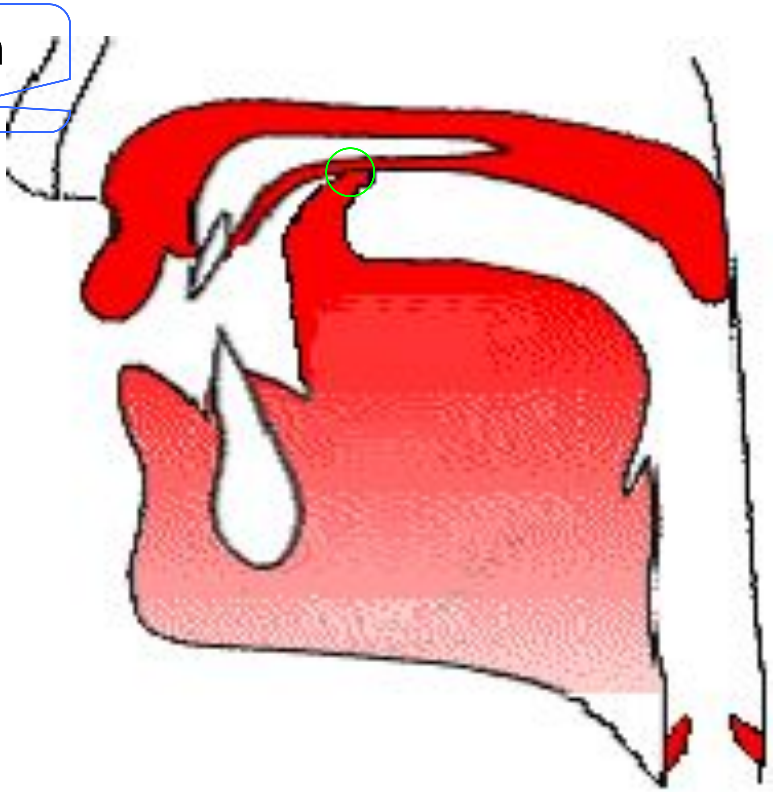


Place of Articulation (Velar)

Velum closed

Back of tongue (Articulator)

Vocal Cord Closed



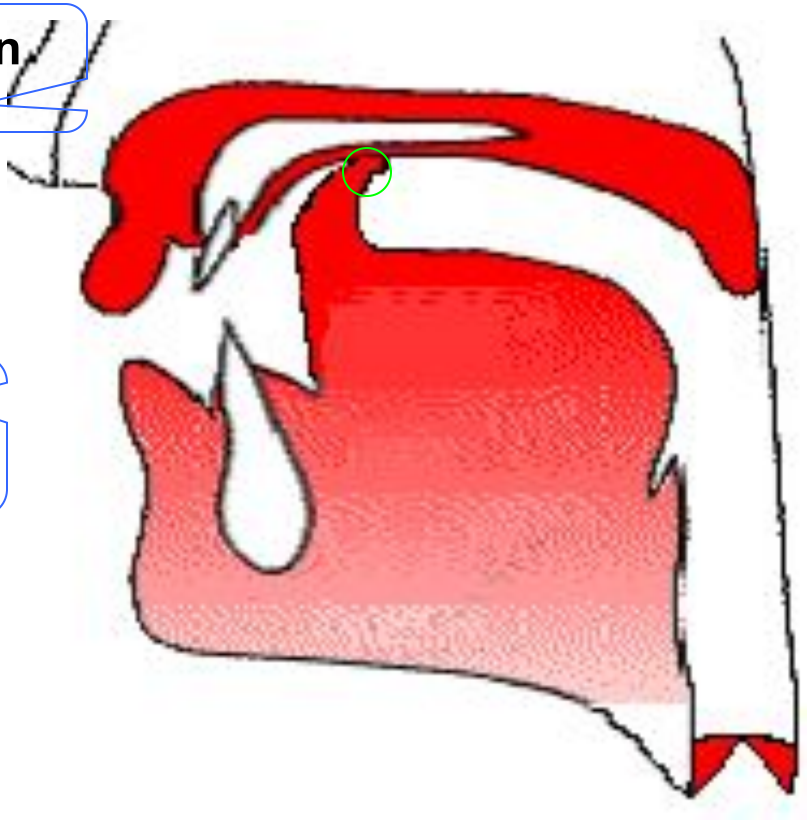
Place of Articulation (Post-alveolar)

Tongue tip curled back (Articulator)

Nasal Passage

Velum closed

Vocal Cord Open



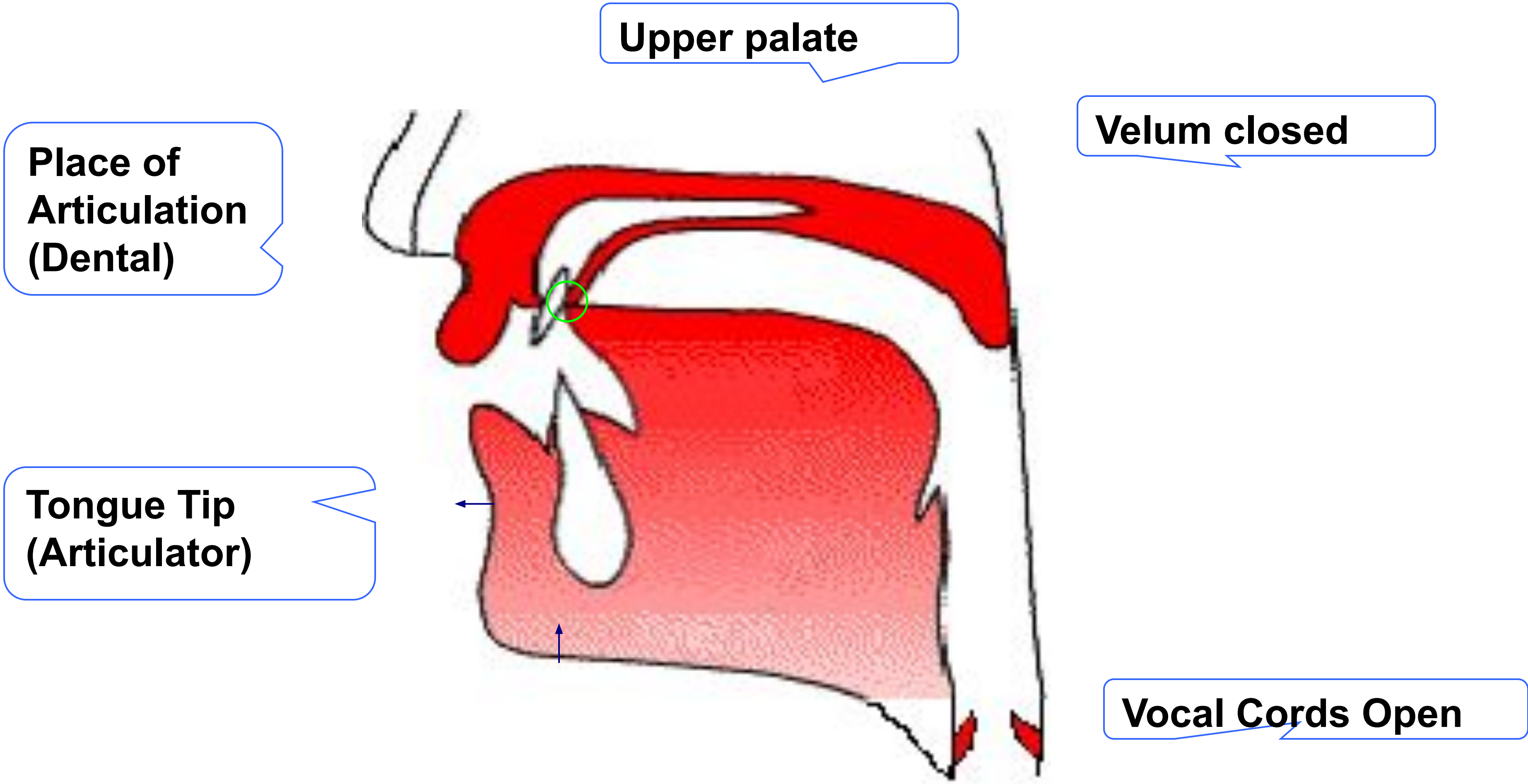
Place of Articulation (Post-alveolar)

Nasal Passage

Velum closed

Tongue tip curled back (Articulator)

Vocal Cord Closed



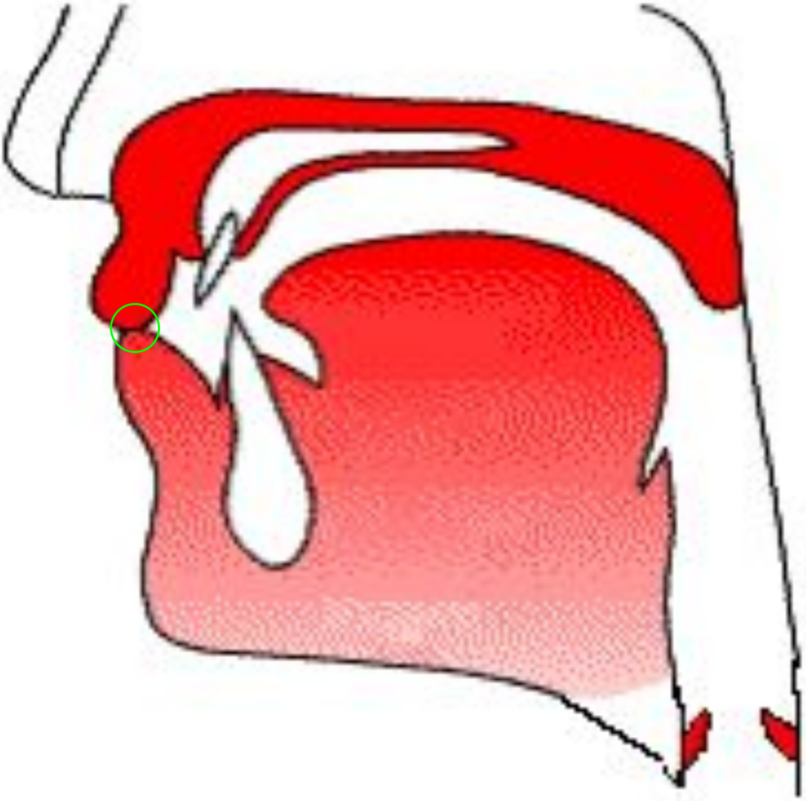
Upper palate

Velum closed

Place of Articulation (Dental)

Tongue Tip (Articulator)

Vocal Cords Open



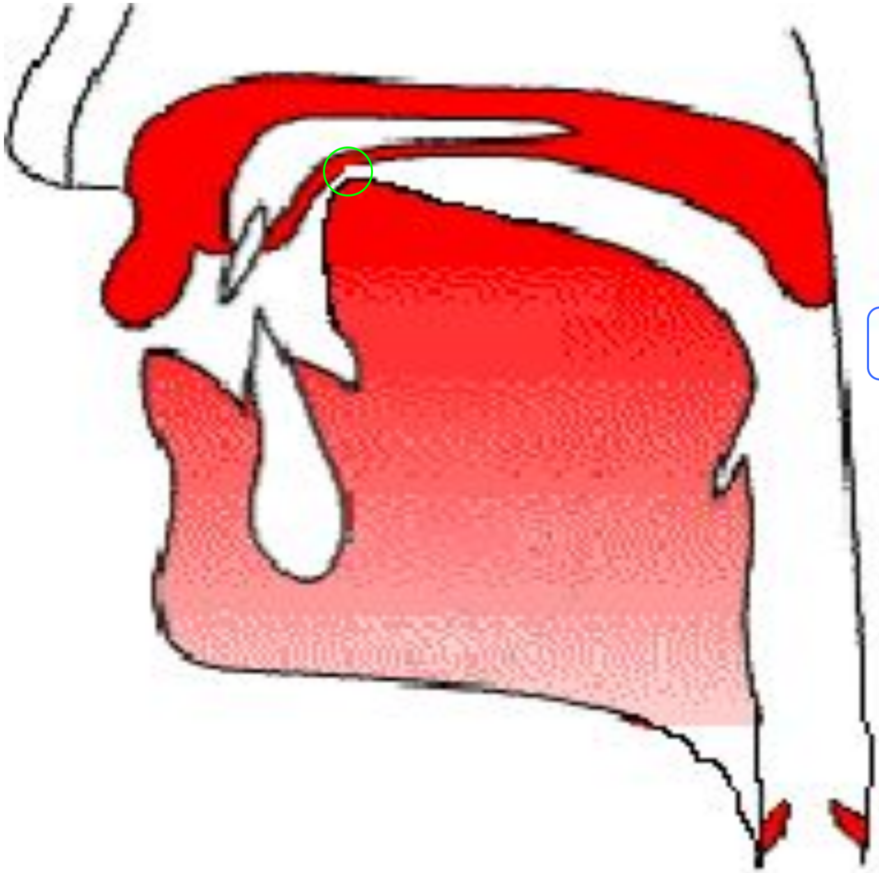
Upper palate

Place of Articulation (Bilabial)

Velum closed

Both the Lips (Articulator)

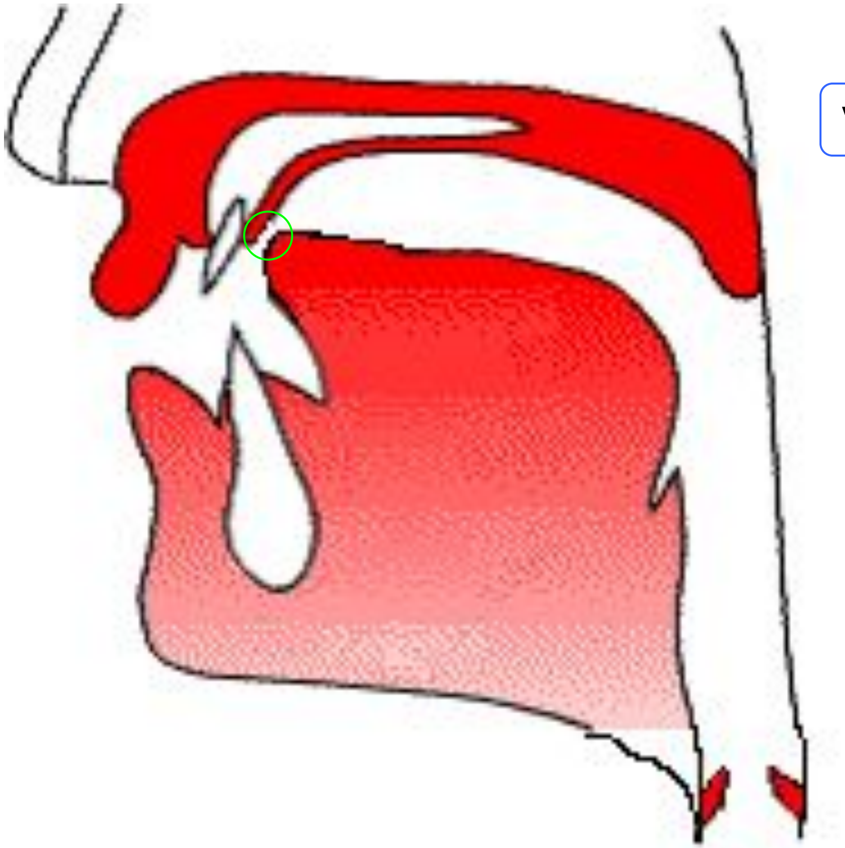
Vocal Cords Open



Place of Constriction (Post alveolar)

Velum closed

Vocal Cords Open



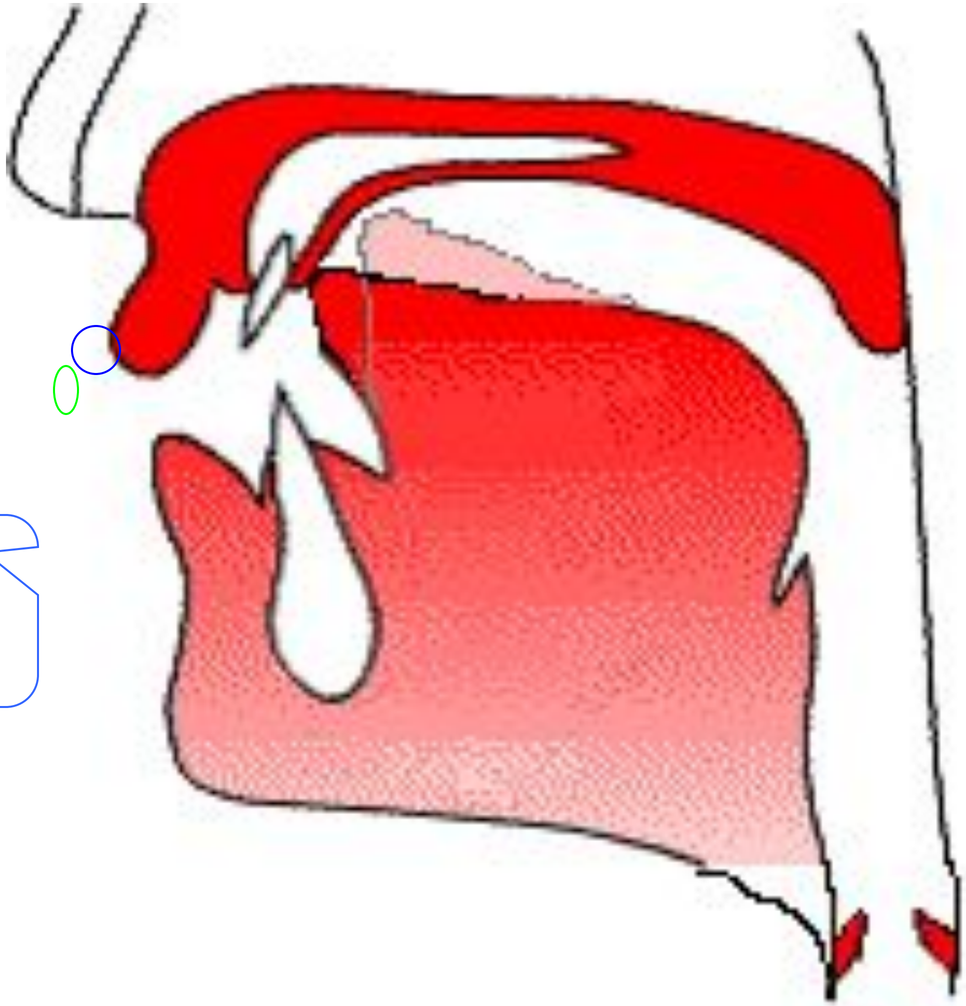
Place of Constriction (Post alveolar)

Velum closed

Vocal Cords Open

Velum closed

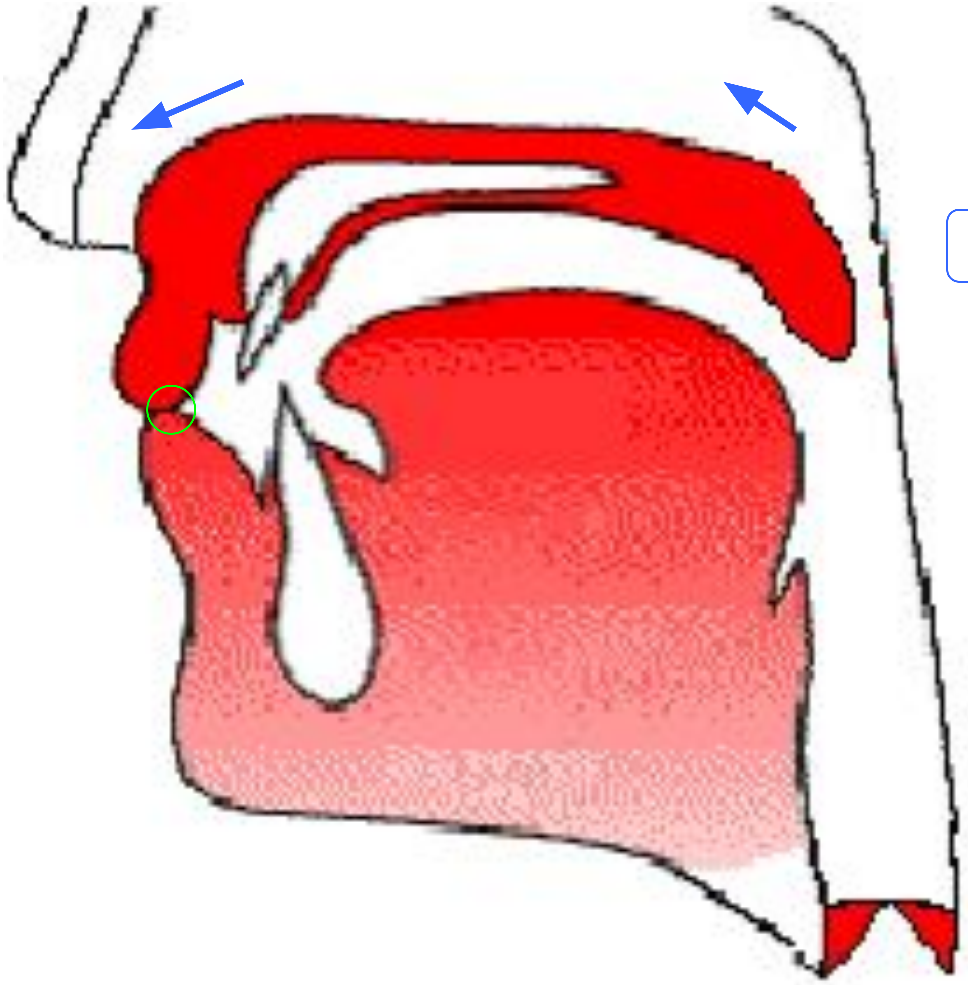
Place of Constriction (Post alveolar)



Place of release (Alveolar)

Vocal Cords Open

Place of Articulation (Bilabial)



Velum Open

Vocal Cords Closed

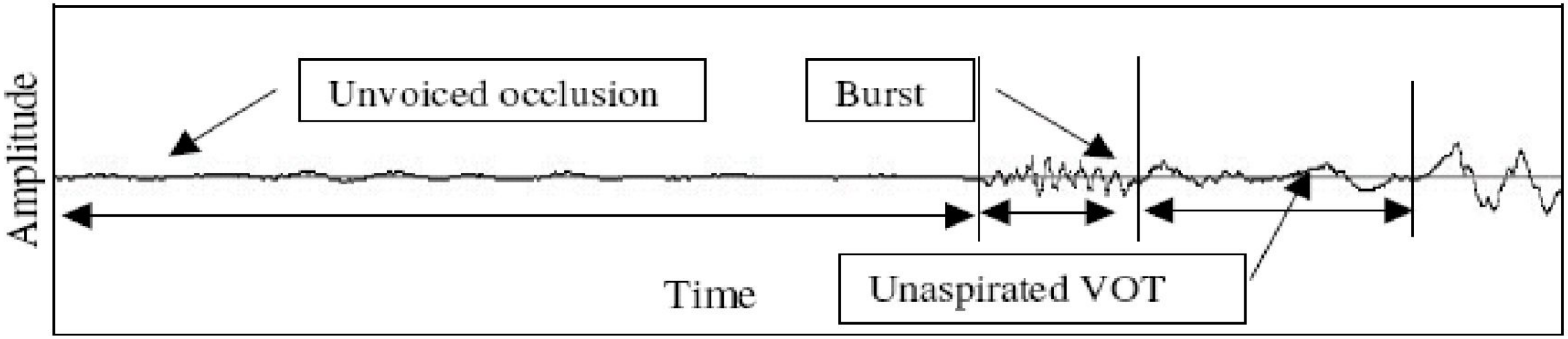


Figure 4.5 Example segment of unaspirated unvoiced stop /k/

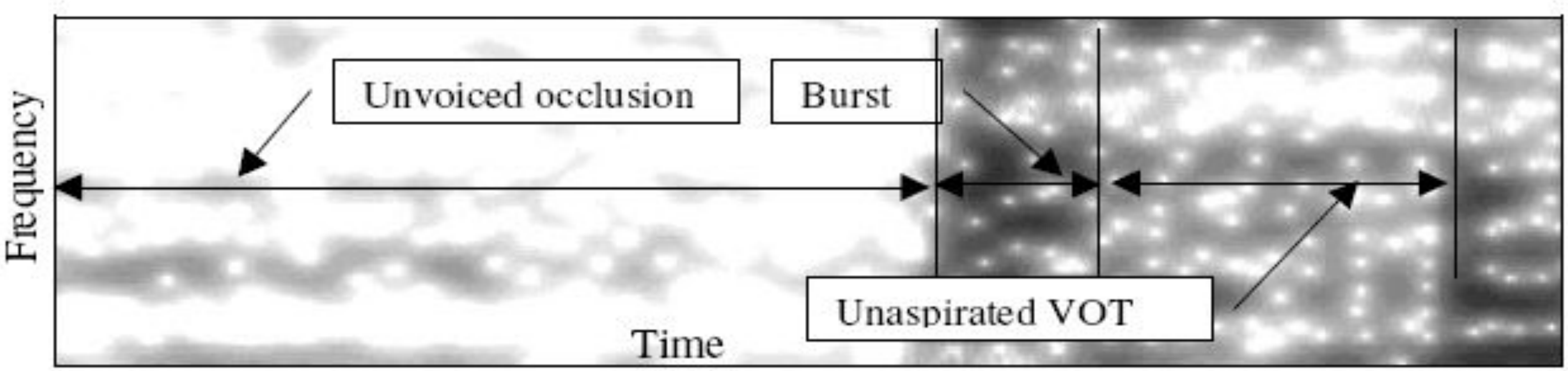


Figure 4.5a Spectrogram of the example unaspirated unvoiced stop /k/

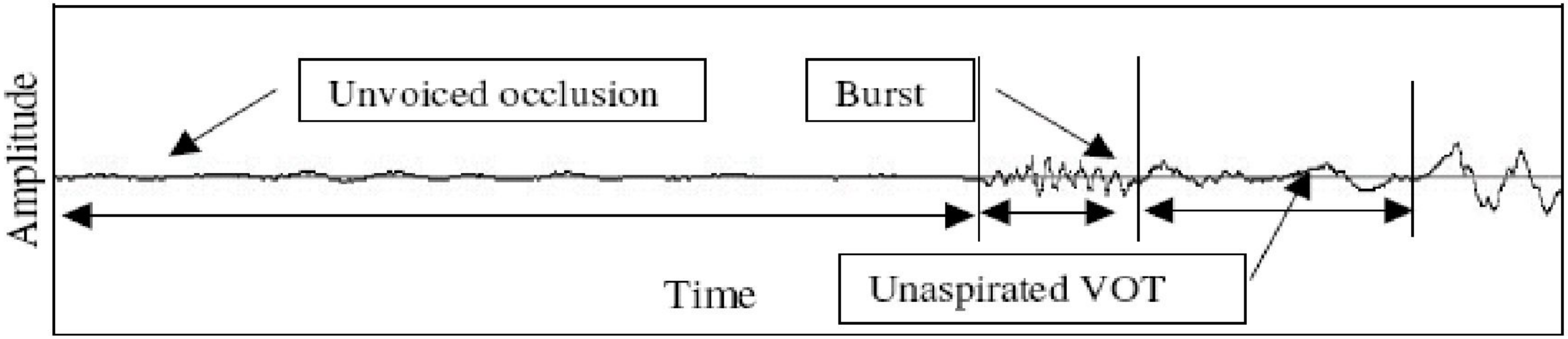


Figure 4.5 Example segment of unaspirated unvoiced stop /k/

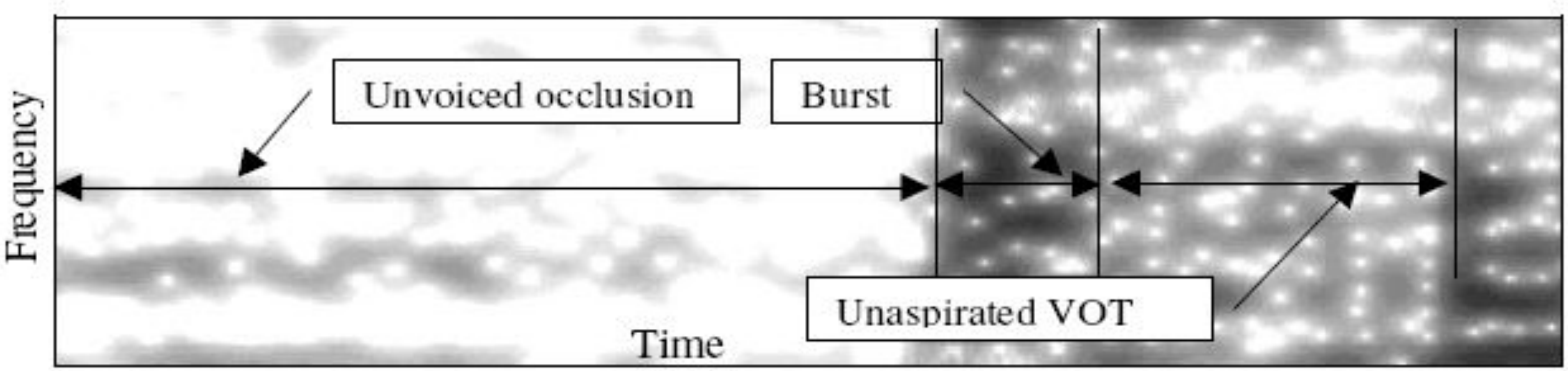


Figure 4.5a Spectrogram of the example unaspirated unvoiced stop /k/

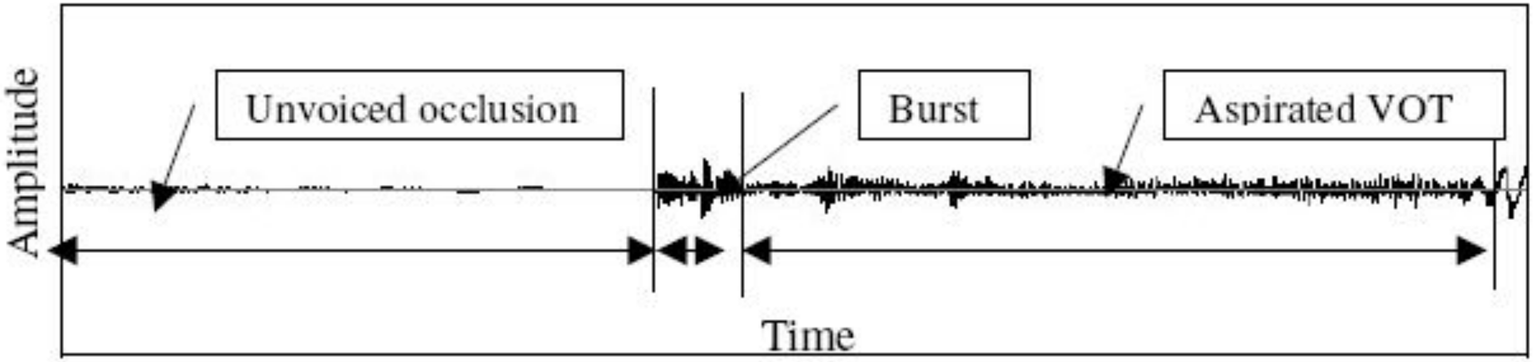


Figure 4.6 Example segment of an aspirated unvoiced stop /k^h/

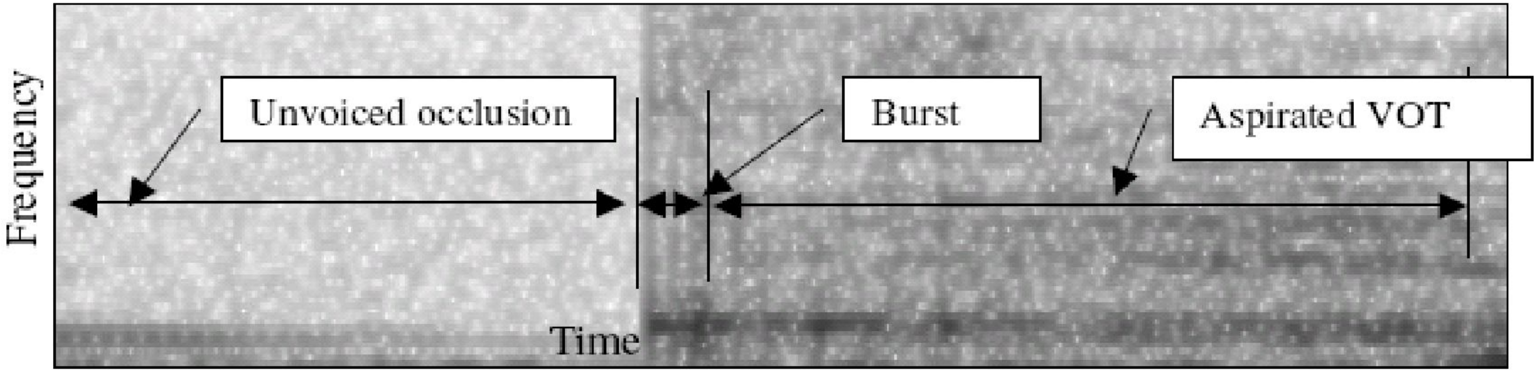


Figure 4.6a Spectrogram example segment of aspirated unvoiced stop /k^h/

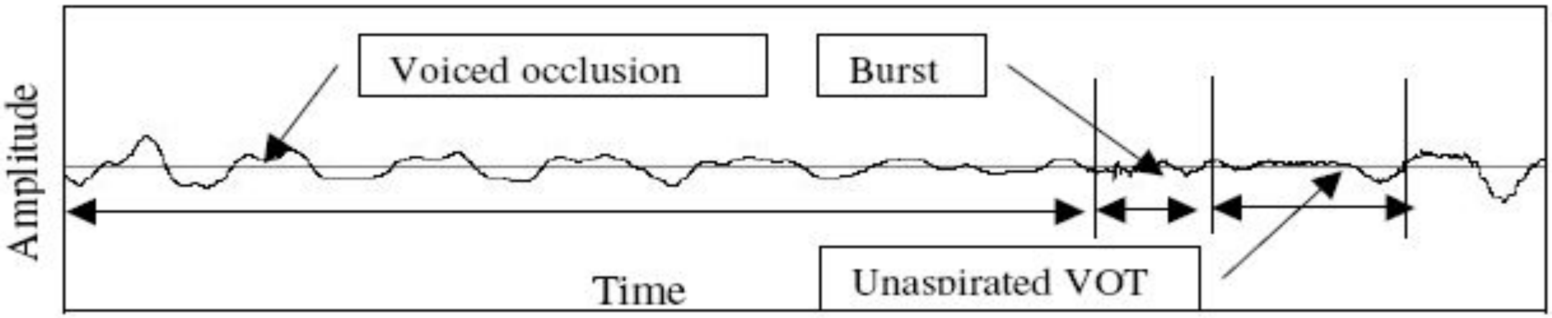


Figure 4.7 Example segment of unaspirated voiced stop /g/

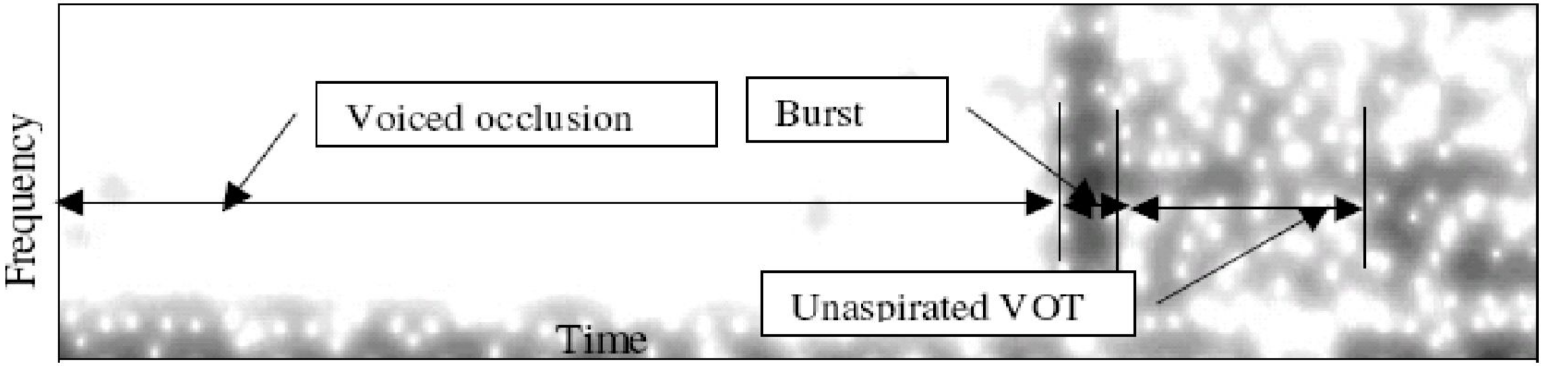


Figure 4.7a Spectrogram segment of unaspirated voiced stop /g/

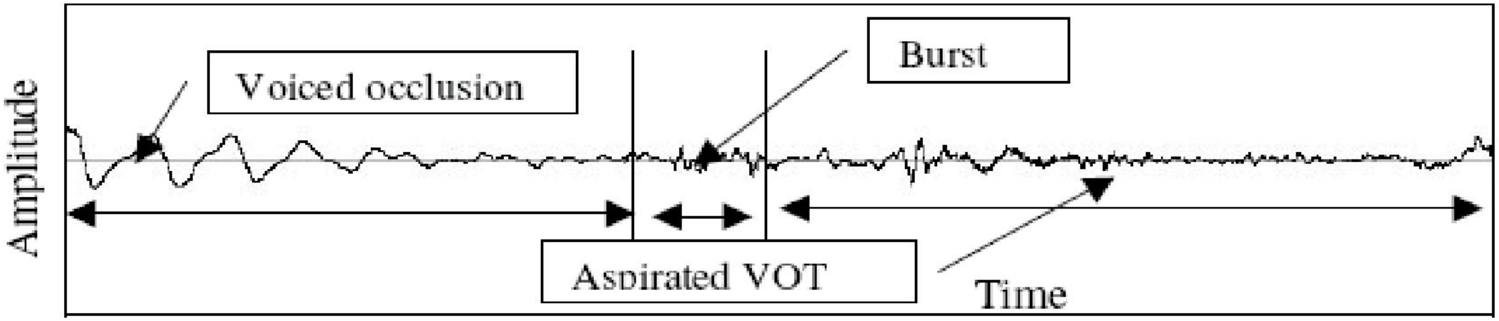


Figure 4.8 Example segment of aspirated voiced stop /g^h/

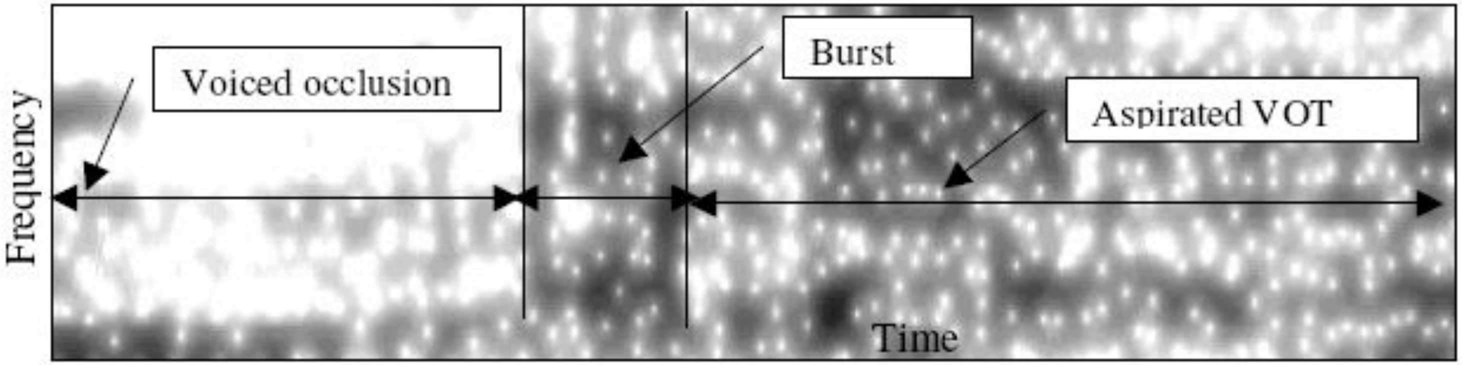


Figure 4.8a Spectrogram example segment of aspirated voiced stop /g^h/

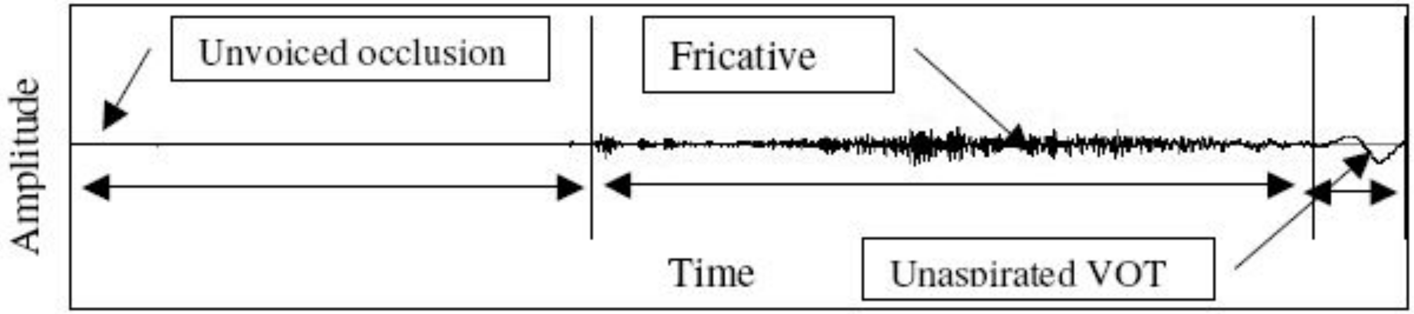


Figure 4.9 Example segment of unaspirated unvoiced affricates / ʧ/

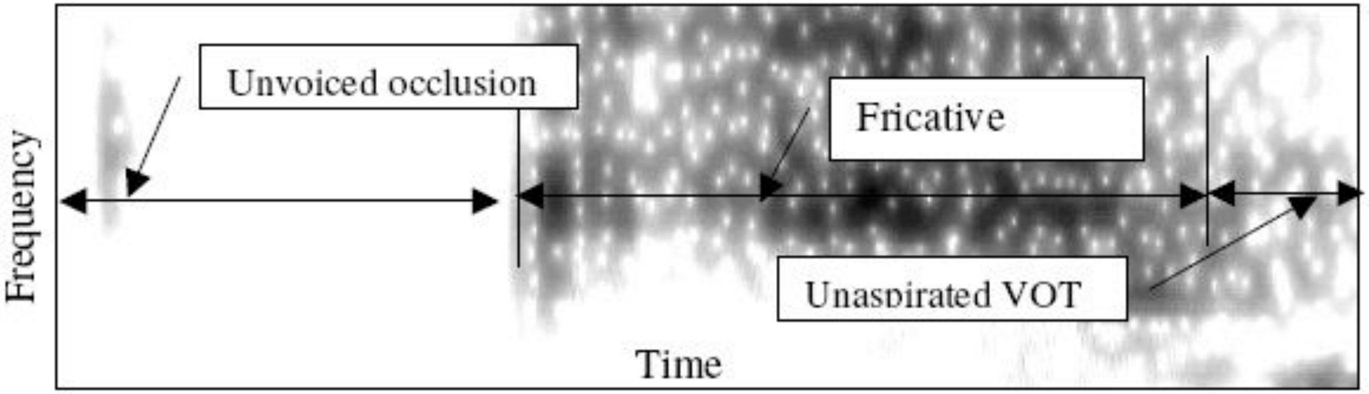


Figure 4.9a Spectrogram of the example segment of unaspirated unvoiced affricates /ʧ/

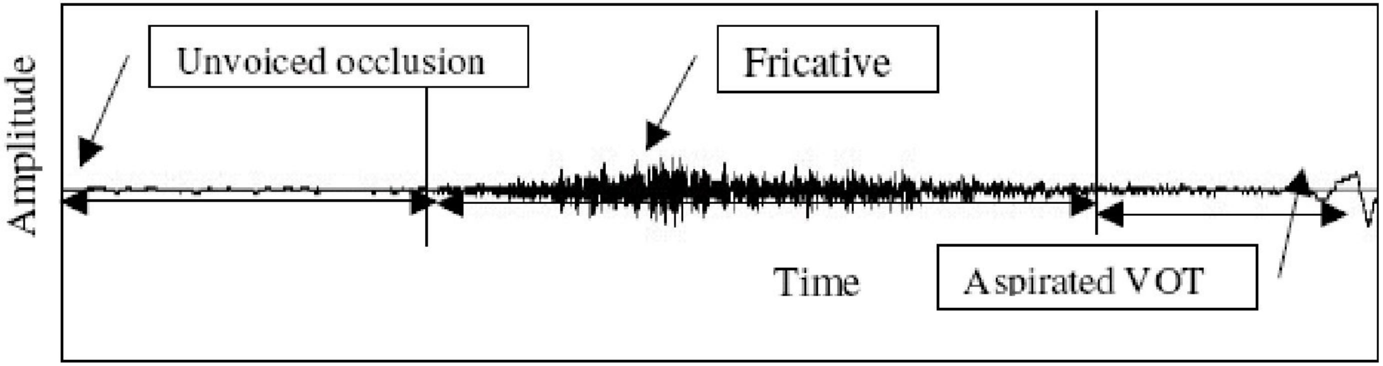


Figure 4.10 Example segment of aspirated unvoiced affricates / tʃʰ/

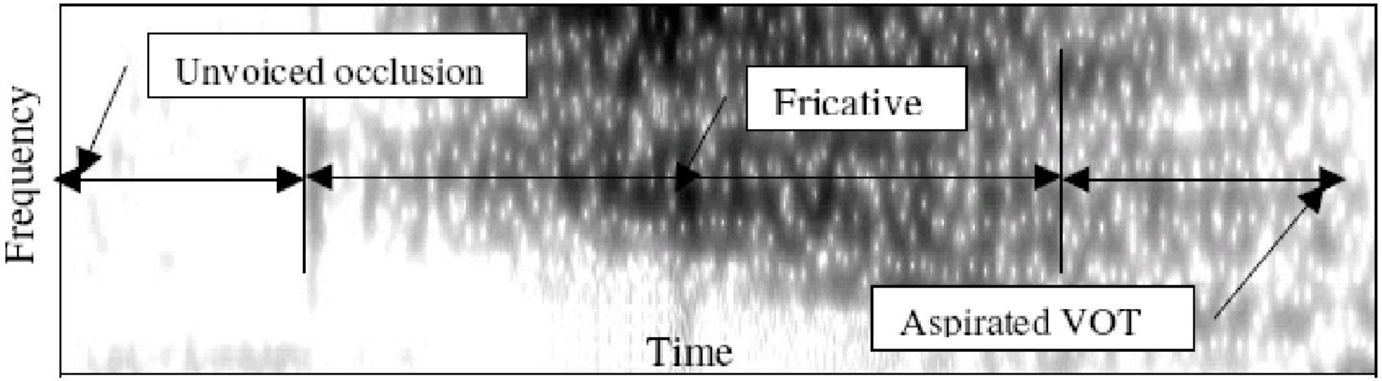


Figure 4.10a Spectrogram of the example segment of aspirated unvoiced affricates /tʃʰ/

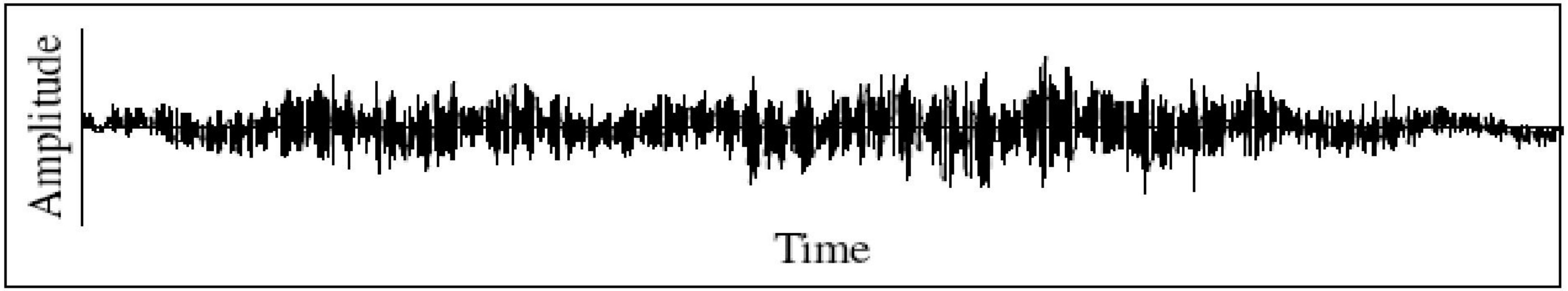


Figure 4.4 An example of sibilant sound segment /s/

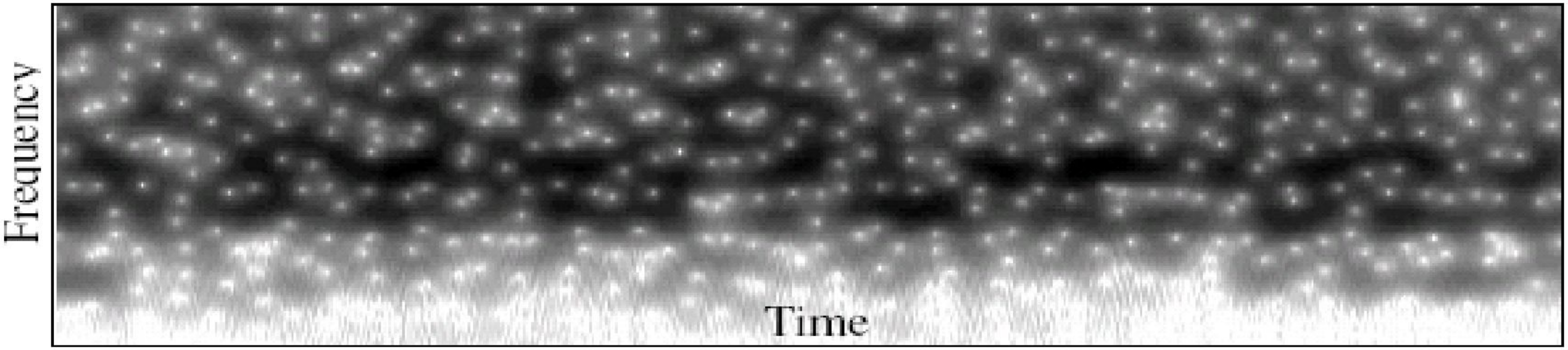


Figure 4.4a Spectrogram of the example sibilant sound segment /s/

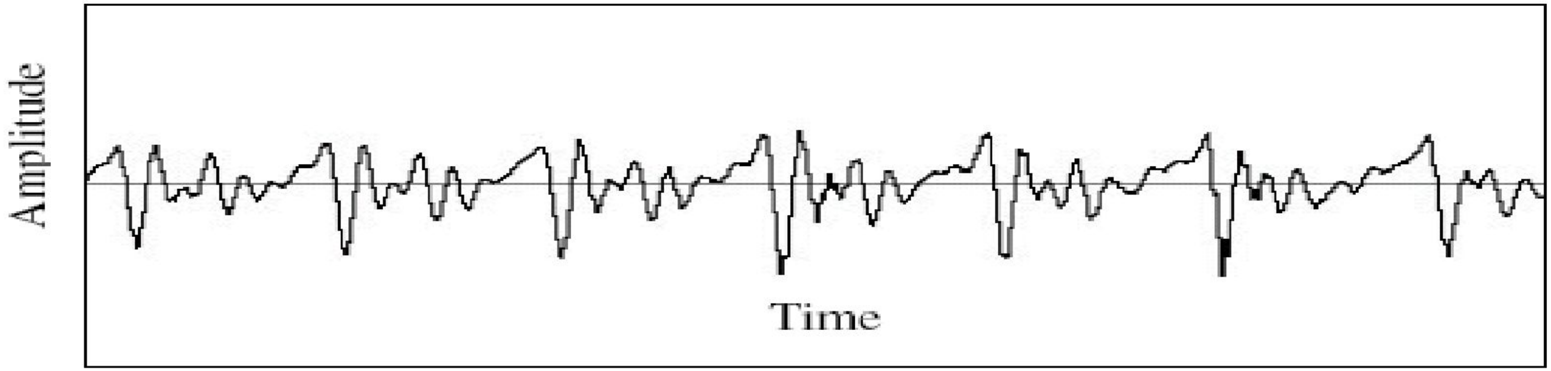


Figure 4.2 Segment of a voiced sound /ɔ/

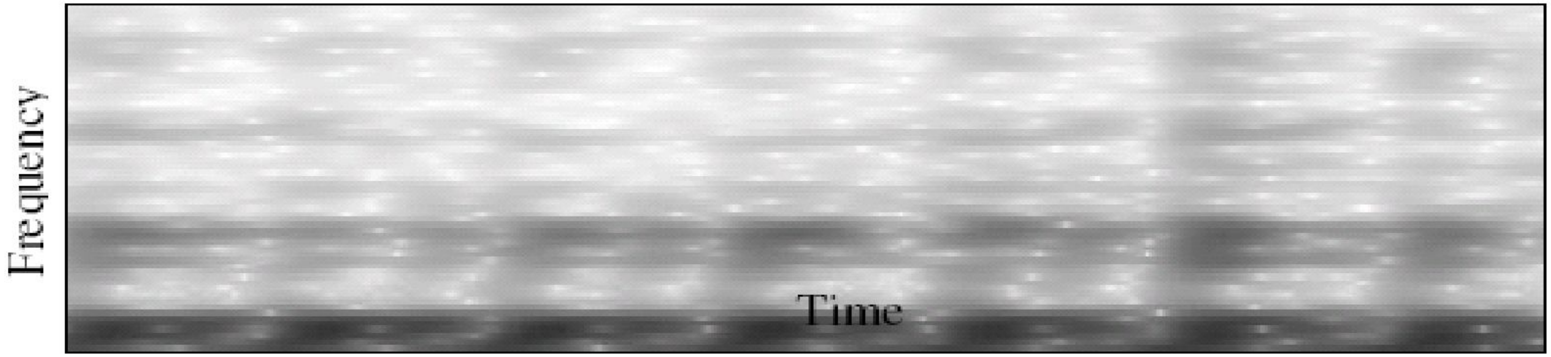


Figure 4.3 Spectrogram of the voice sound /ɔ/

Classification of sound in linguistically distinct speech (phonemes)

- Vowels: a) Oral vowels b) nasal vowels

- Dipthongs: Dipthongs is a gliding monosyllabic speech sound that start at or near the articulatory position for one vowel and moves to or toward the position for another

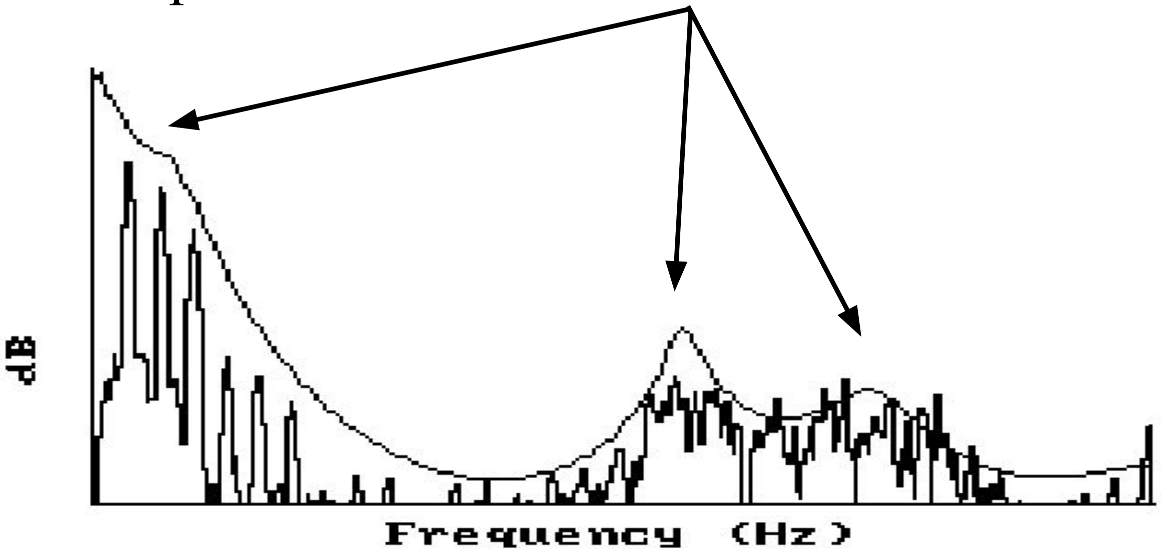
- Semivowels: Semivowels are vowel like nature. They are generally characterized by gluding transition in vocal tract area function between adjacent phonemes.

- Consonant: a. Nasal consonants. b. unvoiced fricatives.

  c. Voiced fricative d. voiced and unvoiced stop

What is Formant??

- To identify dissimilar sounds i.e., vowels, the ears are more sensitive to peaks in the signal spectrum. These resonant peaks in the spectrum are called formants.



Spectrographic view of vowel /i/

- Formants are the characteristics partial that identify vowels to the listeners.

- Formant with lowest frequency is called F1, the second F2 & the third F3. F1 & F2 are enough to disambiguate the vowel.

Different Vowels, Different Formants

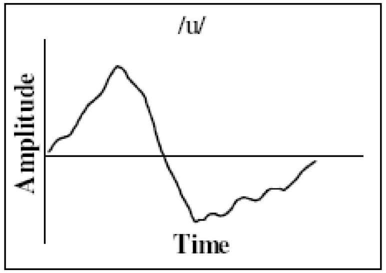
- The formant frequencies of [ə]resemble the resonant frequencies of a tube that is open at one end.
- For the average man (ref: Peter Ladefoged):
  - F1 = 500 Hz
  - F2 = 1500 Hz
  - F3 = 2500 Hz
- However, we can change the shape of the vocal tract to get different resonant frequencies.
- Vowels may be defined in terms of their characteristic resonant frequencies (**formants**).

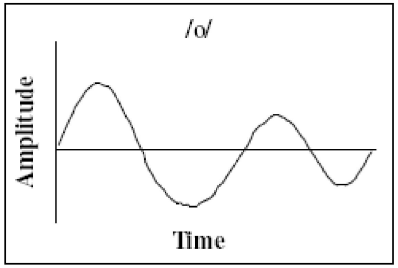
Articulatory description of Vowels

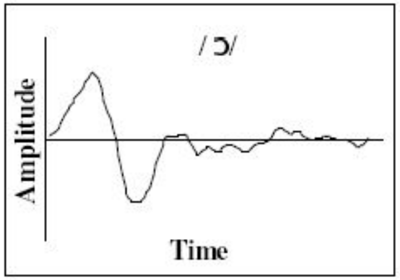
Vowels have traditionally been described according to following pseudo-articulatory parameters:

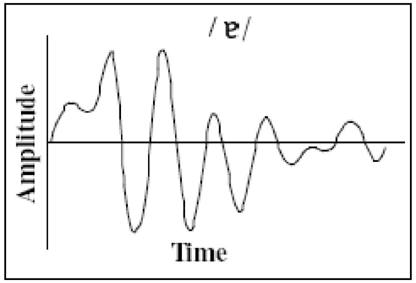
1. Height (of tongue) (F1)
2. Front/Back (of tongue)(F2)
3. Rounding (of lips)

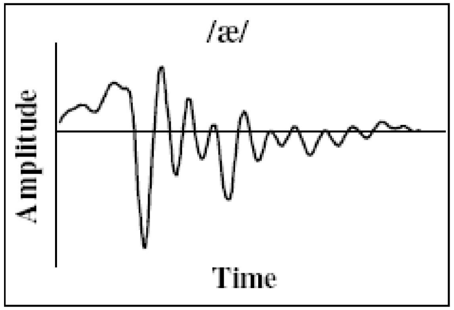
Time Domain Shape

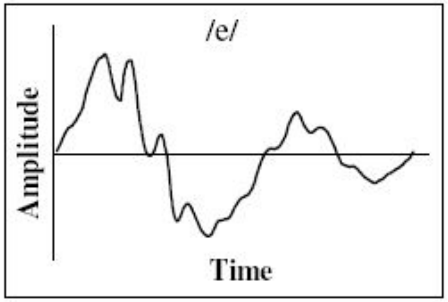


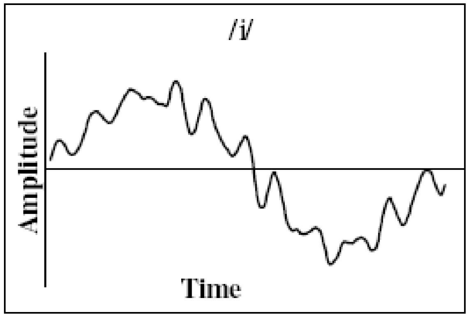




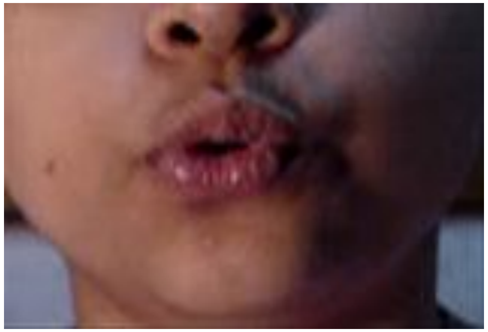




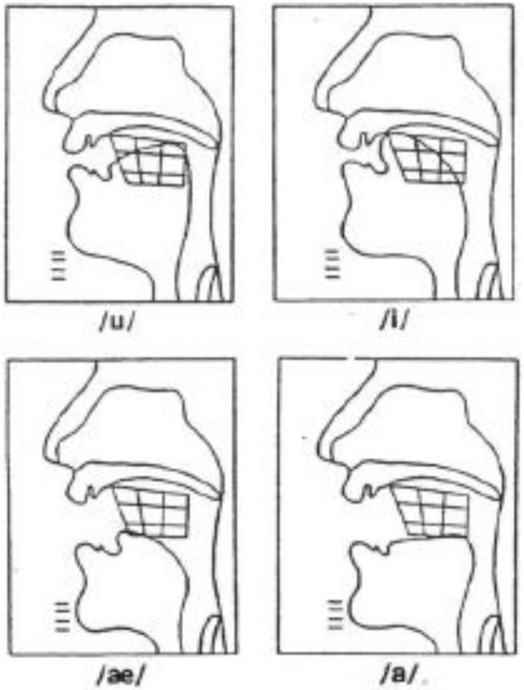


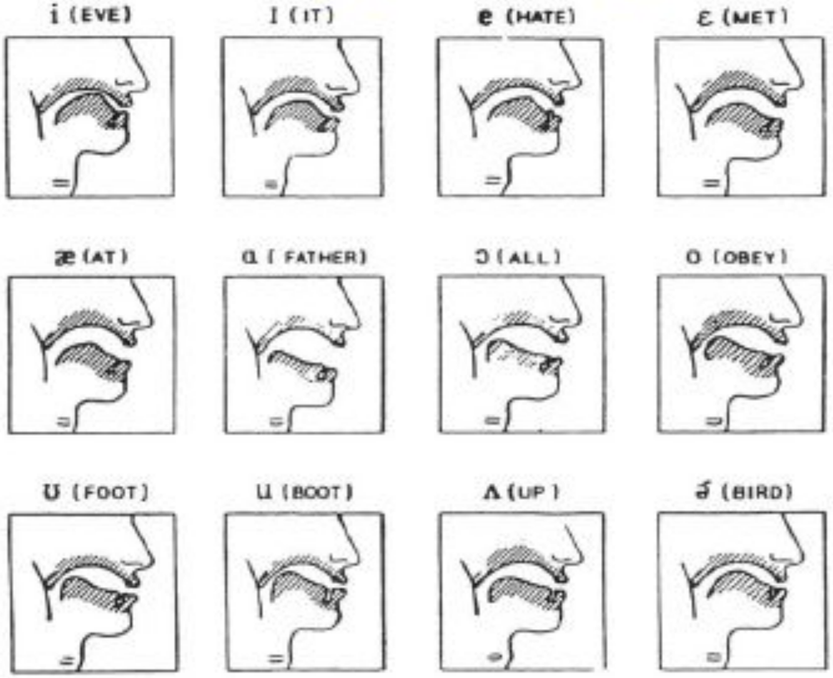


Lip rounding

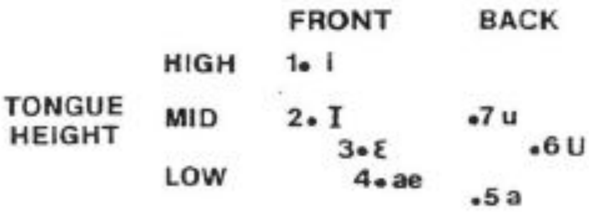


Vowel Articulatory Shapes





TONGUE POSITION

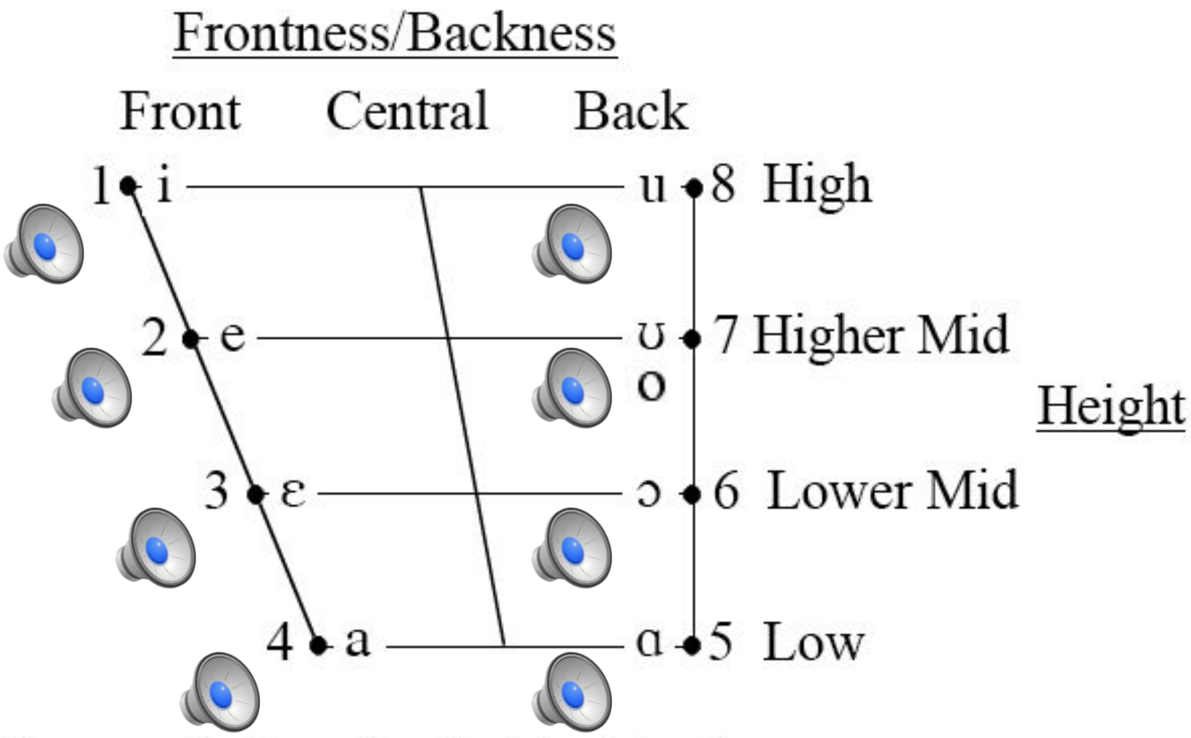


- tongue hump position (front, mid, back)
- tongue hump height (high, mid, low)
- /IY/, /IH/, /AE/, /EH/ => front => high resonances
- /AA/, /AH/, /AO/ => mid => energy balance
- /UH/, /UW/, /OW/ => back => low frequency resonances

The Vowel Space

Cardinal Vowels recorded by Jones in 1965 when he was 75.

(Audio clips from: http://www.let.uu.nl/~audiufon/)



Frontness/Backness

Front Central Back

1 i u 8 High

2 e ʊ 7 Higher Mid

o

Height

3 ɛ ɔ 6 Lower Mid

4 a ɑ 5 Low

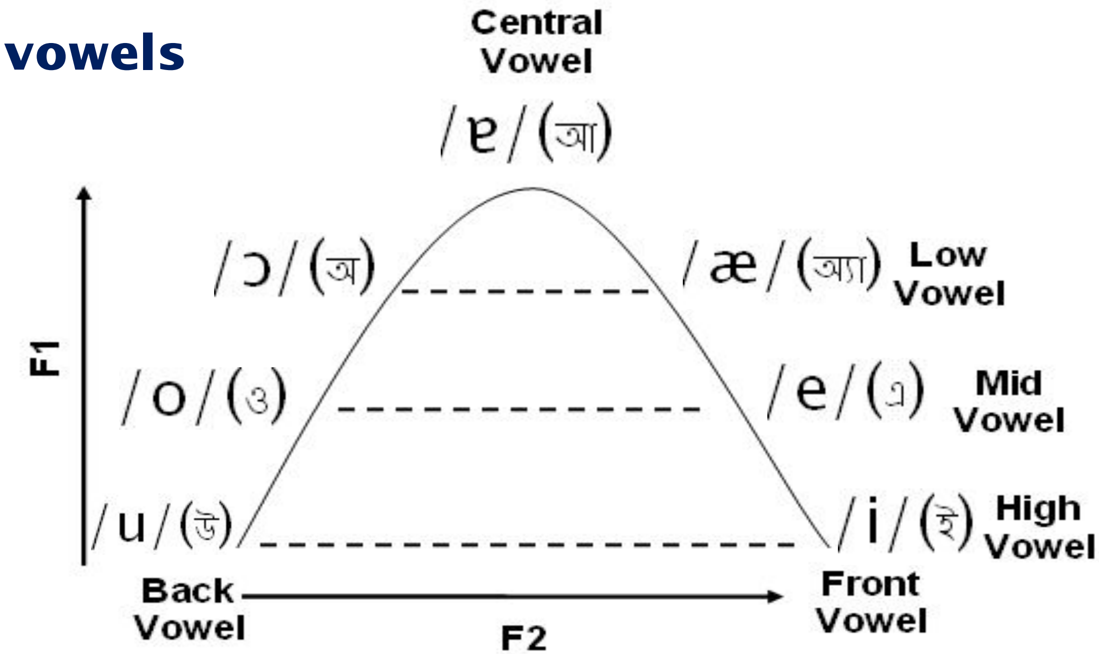
Primary cardinal vowels with rising intonation

Classification of vowels

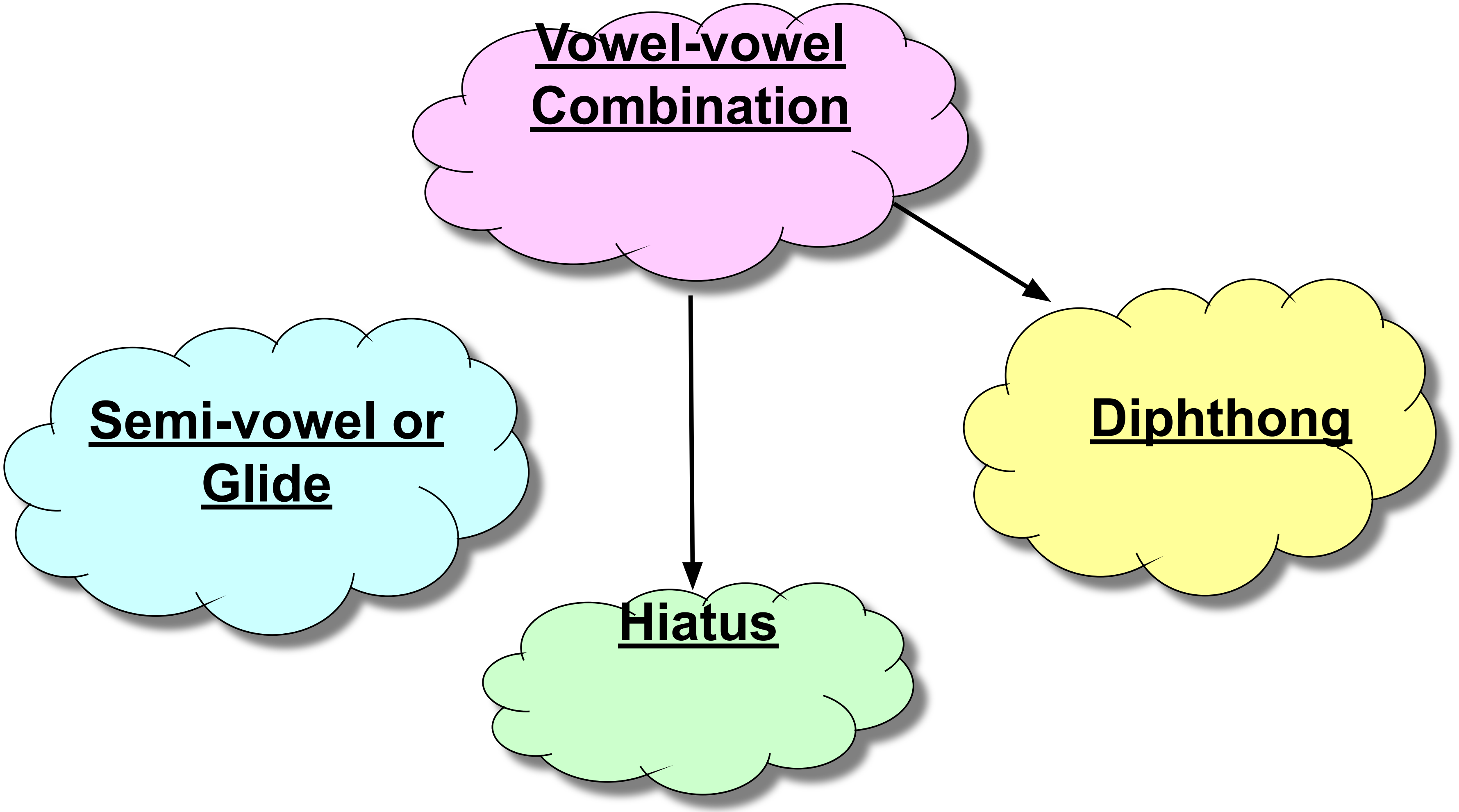
F1 & F2 are primarily determined by the position of tongue. F1 has a higher frequency when the tongue is lowered and F2 has a higher frequency when the tongue is forwarded.

Vowels are classified according to the height and position of the tongue inside the mouth.

Bangla vowels



Position of Bangla Vowels in Cardinal Vowel Diagram



Vowel-vowel Combination

Semi-vowel or Glide

Hiatus

Diphthong

Vowel-Vowel Combination

A. In continuous speech two vowels can come together in two different situations.

B. They may be in a single word.

C. They may be part of two adjacent words i.e., one word ends with a vowel and the next word starts with a vowel.

D. If the two vowels are within a single word, they may either be in two distinct syllables, or may merge into one syllable.

Examples : /b^{h}ulei/(ভুলেই) /pɐik/(পাইক)

Diphthong

- A **diphthong** is a monosyllabic vowel combination involving a quick but smooth movement from one vowel to another, often interpreted by listeners as a single vowel sound or phoneme.

- It is a sequence of two different or same vowels that are part of a single syllable. Usually one of the vowels is stronger than the other.

.

- Examples:

Bangla Word :

/tʃɐi/ (চই)

Bangla Word :

/b^hulei/ (ভুলেই)

Hiatus

□When two vowels coming together without any contraction or elision are pronounced separately as distinct from Diphthongs they are termed as **hiatus**.

□Hiatus may be of two types:

1) **Internal Hiatus**□ which occurs within a word.

Example: Bangla Word : /pɐik/(পাইক)

2) **External Hiatus** □ which refers to the break between two successive words. In this situation the first word ends with a vowel and the second word starts with a vowel.

Example:

Bangla Sentence : /ɐmi iliʃ k^hɐbo/(আমি ইলিশ খাব)

Semi-vowel or Glide

- **Semi-vowel** refers to a sound functioning as a consonant but lacking the PHONETIC characteristics normally associated with consonants.

- Its QUALITY is phonetically that of a vowel; though its DURATION is much less than that typical of vowel.

- Examples:

Bangla Word :

/ɔjon/(অয়ন)

Bangla Word :

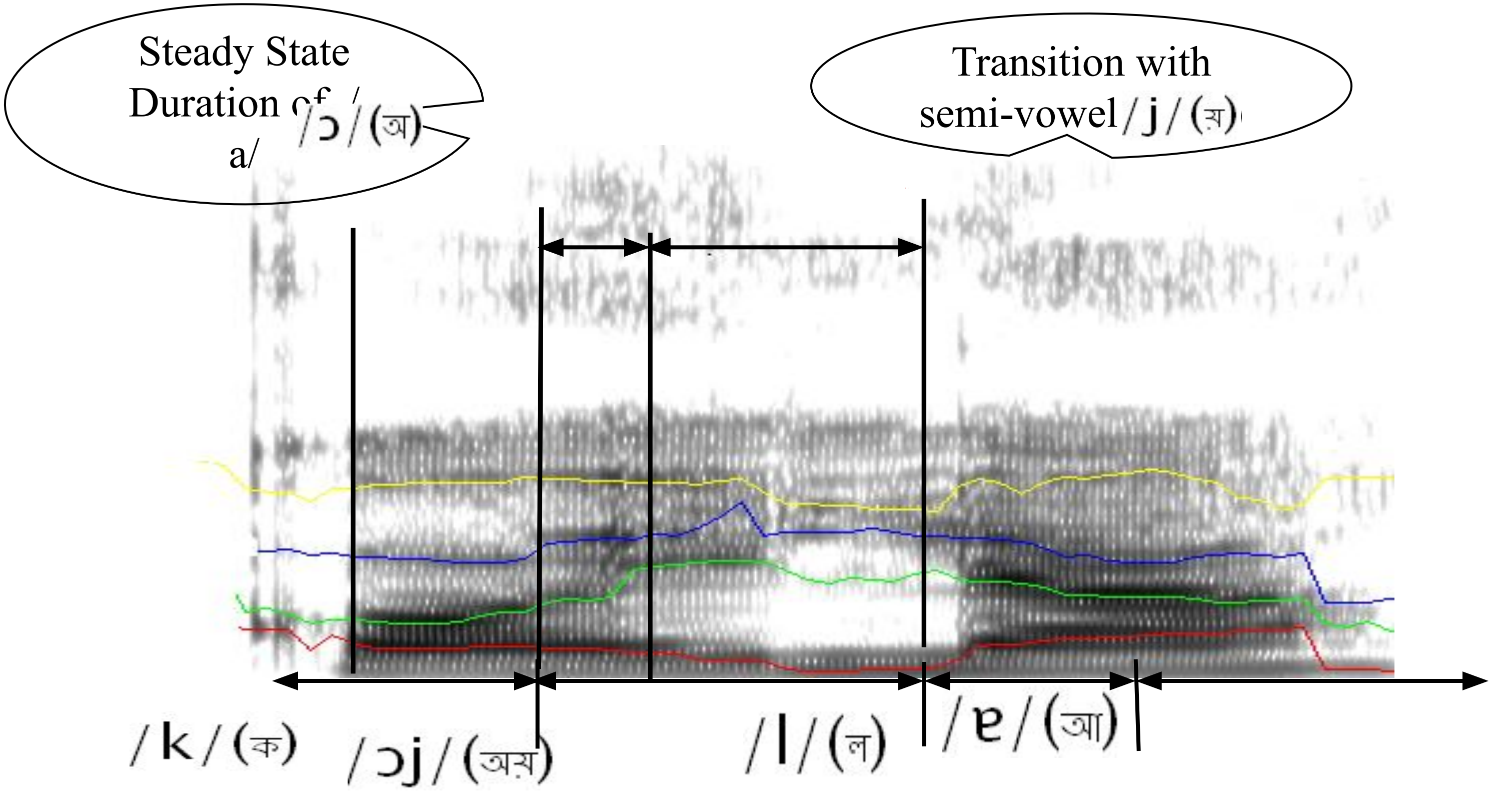
/meje/(মেয়ে)

Bangla Word :

/heʋe/(হাওয়া)

Semi-vowel after a vowel

Vowel-semivowel combination (V-j) consists of transitional duration with semivowel along with the preceding vowel’s steady state duration.

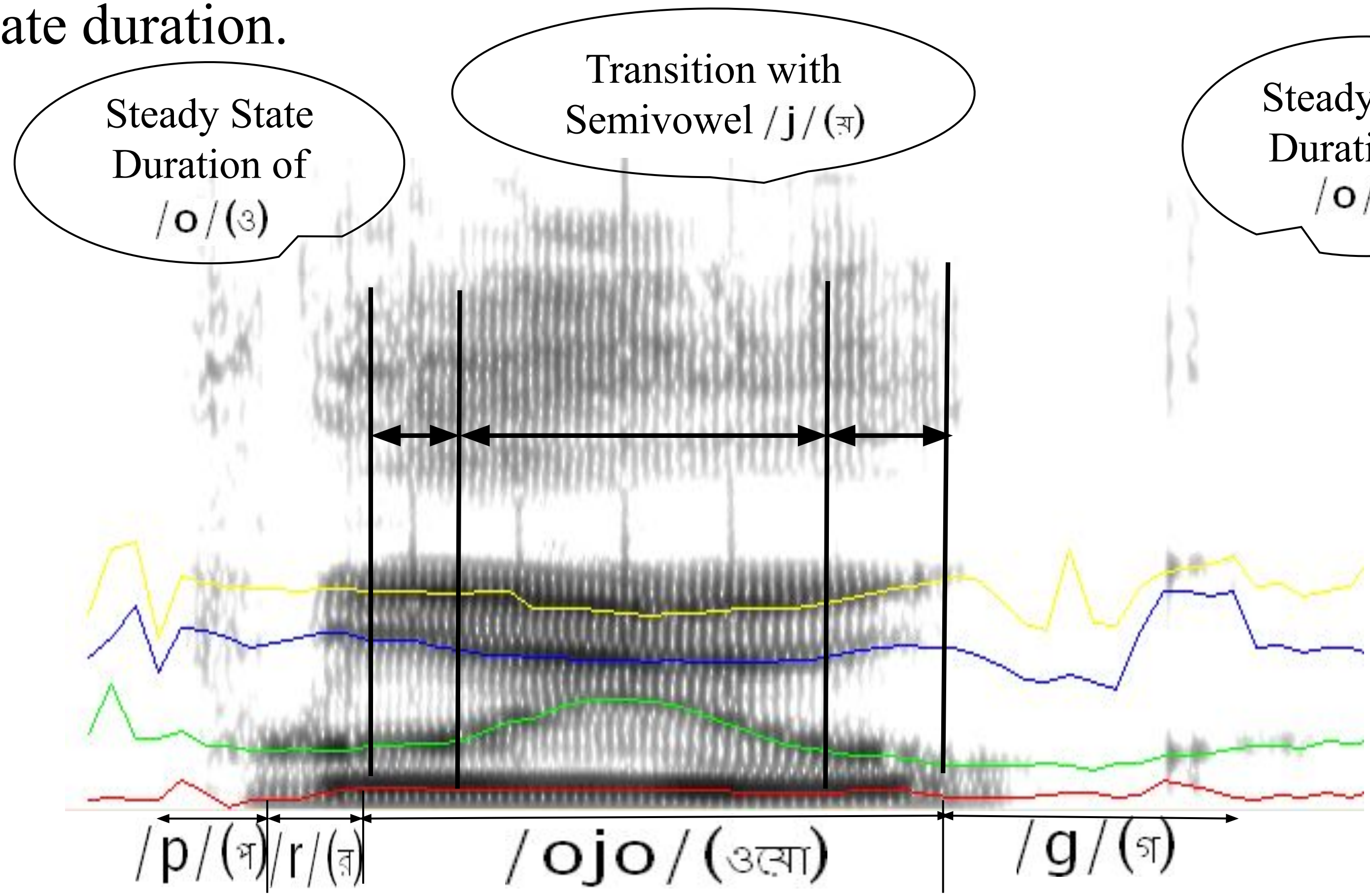


Spectrographic View of Bangla Word /kɔjlɐ/ (কয়লা) with V-j combination /ɔj/ (অয়)

Play

Semi-vowel in between two vowels

- Vowel-semivowel-vowel combination consists of transitional duration with semivowel along with the preceding and succeeding vowels’ steady state duration.

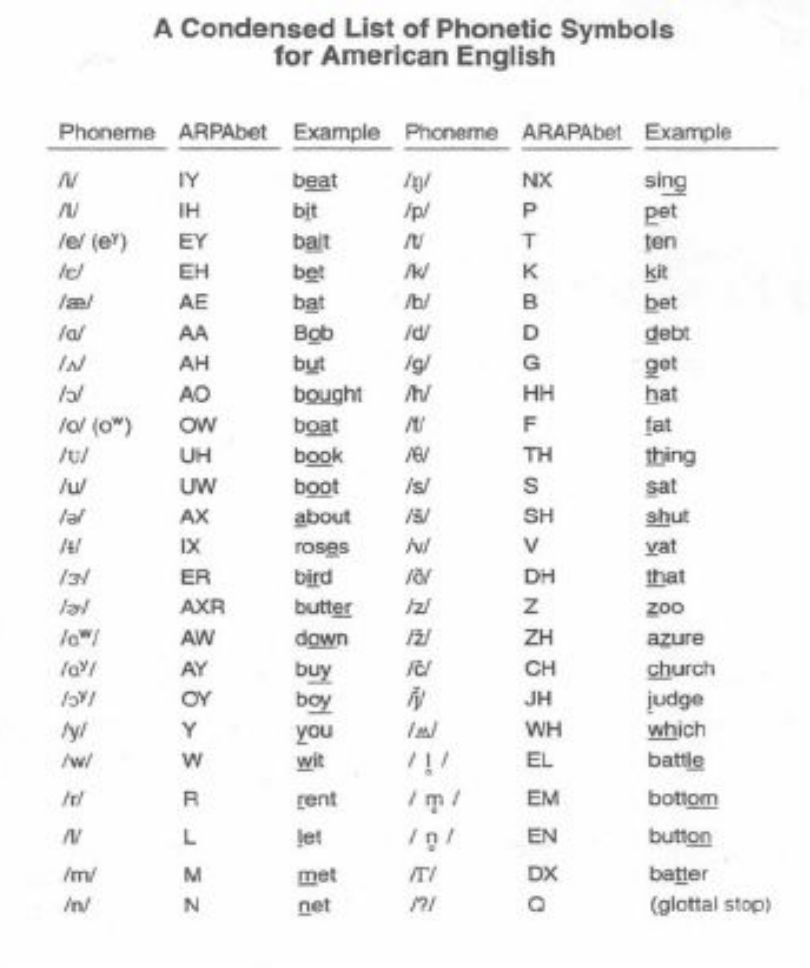


Play

/ojo/(ওয়ো)

Spectrographic View of Bangla Word (/projog/(প্রয়োগ)) with V-j-V combination

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/ (e^y)	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/ (o^w)	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
/$ɑ^w$/	AW	down	/ž/	ZH	azure
/$ɑ^y$/	AY	buy	/č/	CH	church
/$ɔ^y$/	OY	boy	/ǰ/	JH	judge
/y/	Y	you	/ʍ/	WH	which
/w/	W	wit	/ l̩ /	EL	battle
/r/	R	rent	/ m̩ /	EM	bottom
/l/	L	let	/ n̩ /	EN	button
/m/	M	met	/ɾ/	DX	batter
/n/	N	net	/ʔ/	Q	(glottal stop)

ARPABET representation

- *48 sounds*
  - 18 vowels/diphthongs
  - 4 vowel-like consonants
  - 21 standard consonants
  - 4 syllabic sounds
  - 1 glottal stop

Consonents		Manner of Articulation				
S/N	Place of Articulation		Unvoiced		Voiced	
			Un-Aspirated	Aspirated	Un-Aspirated	Aspirated
1	Velar	Stop	/k/	/k^h/	/g/	/g^h/
2	Post-alveolar (Retroflex)		/ʈ/	/ʈ^h/	/ɖ/	/ɖ^h/
3	Dental		/t/	/t^h/	/d/	/d^h/
4	Bilabial		/p/	/p^h/	/b/	/b^h/
5	Alveolar -Post alveolar	Affricate	/ʧ/	/ʧ^h/	/ʤ/	/ʤ^h/
6	Alveolar	Fricative	/s/			
7	Post alveolar		/ʃ/			
8	Glottal		/h/		//	
9	Velar	Nasal Murmur			/ŋ/	
10	Palatal				/ɳ/	
11	Dental				/n/	
12	Bilabial				/m/	

S/N	Place of Articulation	Manner of Articulation				
			Unvoiced		Voiced	
			Un-Aspirated	Aspirated	Un-Aspirated	Aspirated
13	Dental	Lateral			/l/	
14	Alveolar	Trill			/r/	
15	Post alveolar	Retroflex Flap			/ɽ/	/ɽh/
16	Palatal	Approximant			/j/	
17	Bilabial				/w/	
Vowel						
1	Back vowel	Close, Rounded			/u/	
2	Back vowel	Close-mid, Rounded			/o/	
3	Back vowel	Open, Rounded			/ɔ/	
4	Front vowel	Open, Unrounded			/a/	
5	Front vowel	Open-mid, Unrounded			/æ/	
6	Front vowel	Close-mid, Unrounded			/e/	
7	Front vowel	Close, Unrounded			/i/	

TUTORIAL

1.Write the place and manner of articulation of the following phoneme

/k/, /g/,/u/,/g^{h}/, /ɽ/, /ʃ/

2. Write out the phonetic transcription for the following words:

/she/, /phonetic/, /marks/, /speech/,

How many syllable is present in each of the above word.

3. Draw Schematic representation of the physiological mechanism of speech production system and explain how the a voiced sound is produce.

4. A voiced operated lift operation is designee using the following words

a. stop, b. up, c. down d. floor e. first f. second g. third h. fourth and i. ground.

Figure 1 shows wideband spectrograms of one version of each of these words. Using your knowledge of acoustic phonetics, determine which wideband spectrogram corresponds to which word.

5. The following waveform is for the utterance /kolkata/ and the waveform samples are at a sampling rate of FS =22050 Hz. Segment the waveform into regions of "Voiced Speech (V)" and "Non-Voiced Speech (N)".

6. Which formant frequency is related to tongue height and which formant related to tougue position

7. Why the child speech has high F0 and formant compare to a adult