EE679: Speech Processing

A preview

Dept of Electrical Engineering
I.I.T. Bombay

Why do we need a special course for signal processing of speech?

"Signal processing" is concerned with the mathematical representation of the signal and the algorithmic operations carried out to modify the signal or to extract information from it.

The representation and the algorithms are application domain specific, i.e. there are no "generic" methods.

An understanding of the signal and of the application are crucial to the success of the signal processing methods

Human communication

- Vocal, visual, gestural
- Language is used for communication and is independent of the modality (writing, signing, speaking)
- Speech Communication is the transfer of information from one person to another via speech

The interdisciplinary nature… *

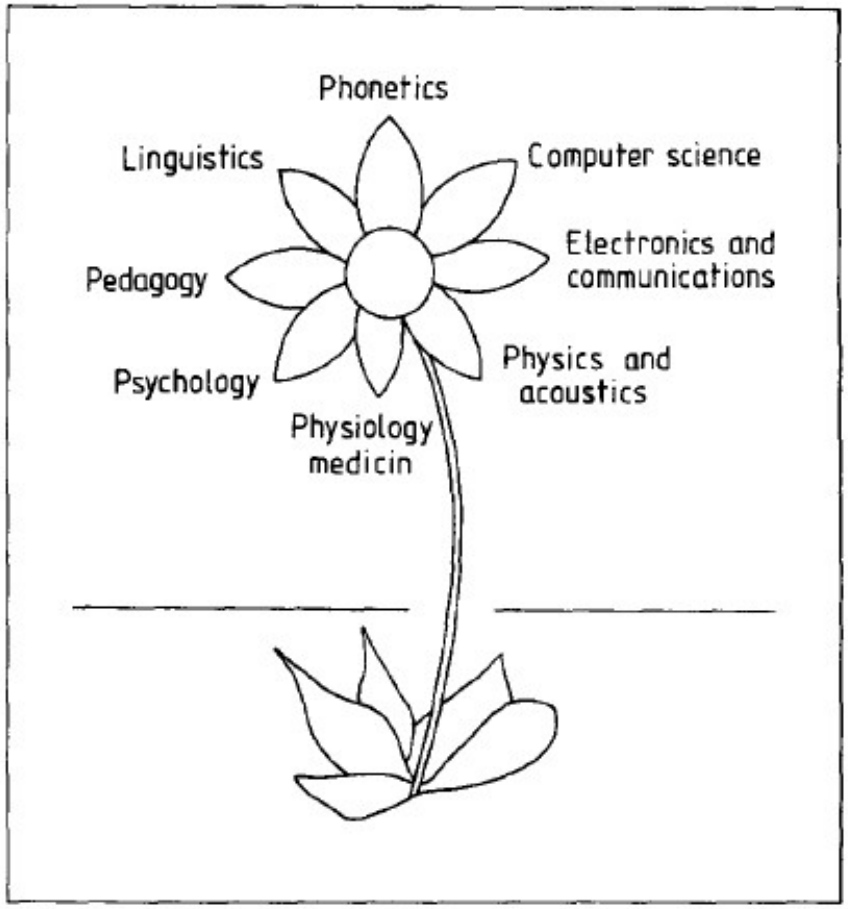
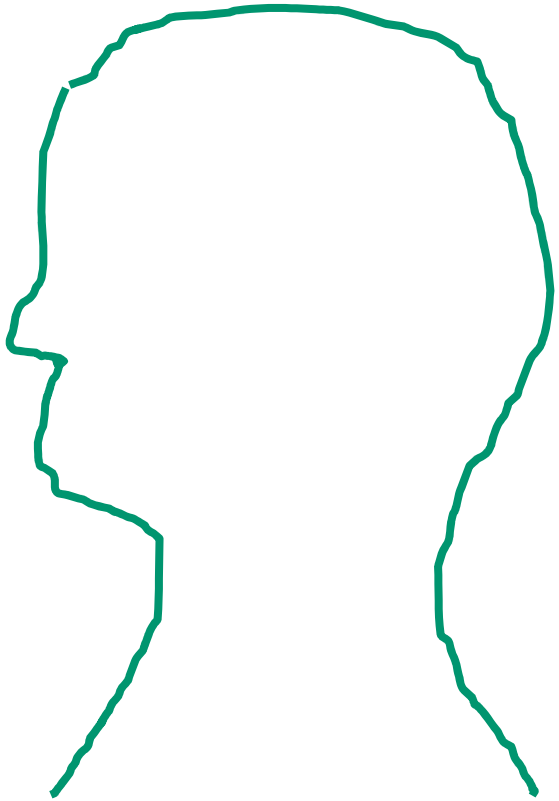


Fig. 4. The interdisciplinary nature of speech research.

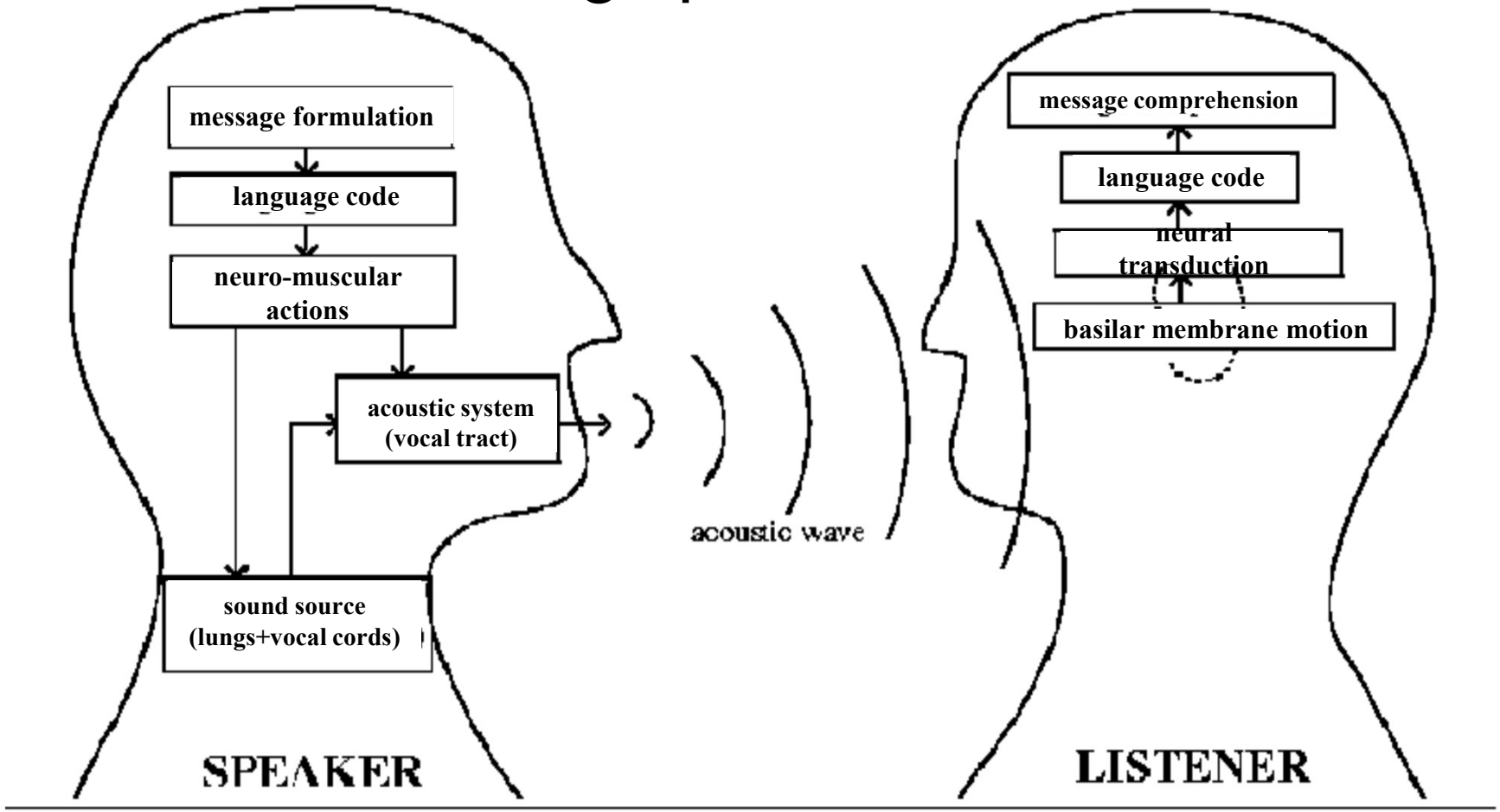
* Fant, G. (1990). Speech research in perspective. Speech Communication.

TALKER



LISTENER

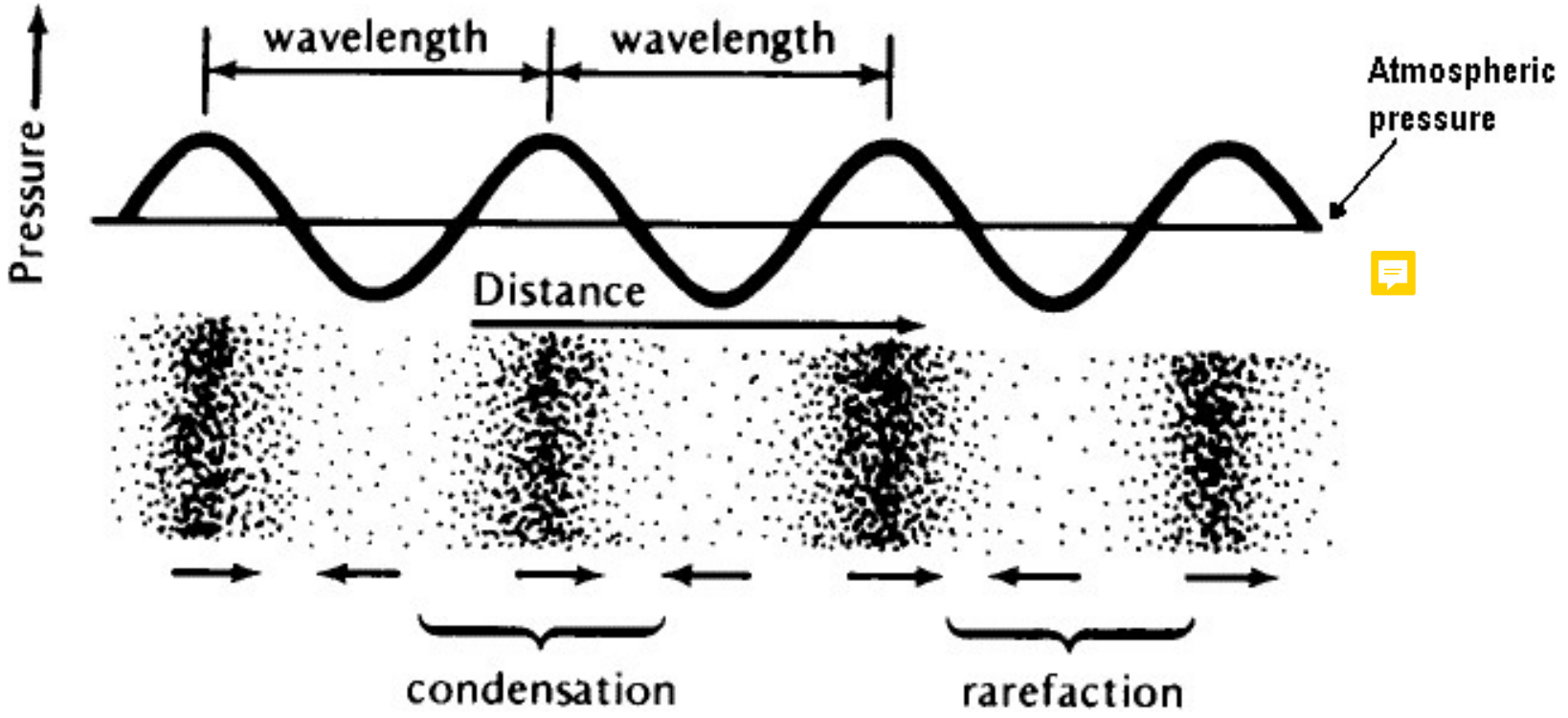
Understanding speech communication

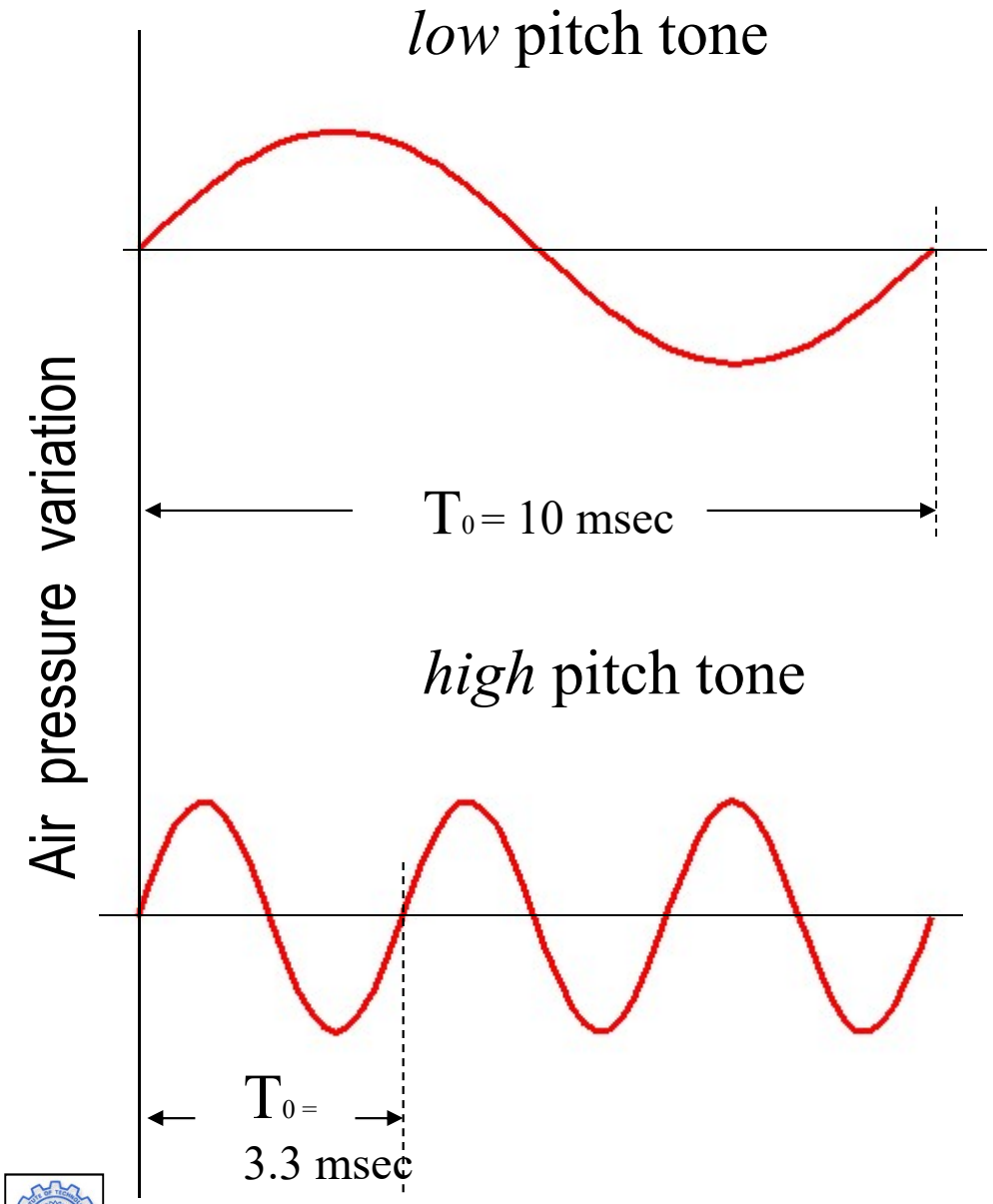


6

Acoustic waveforms

Speed = *wavelength* x *frequency*





Frequency (Fo) = 1/To
= 100 Hz

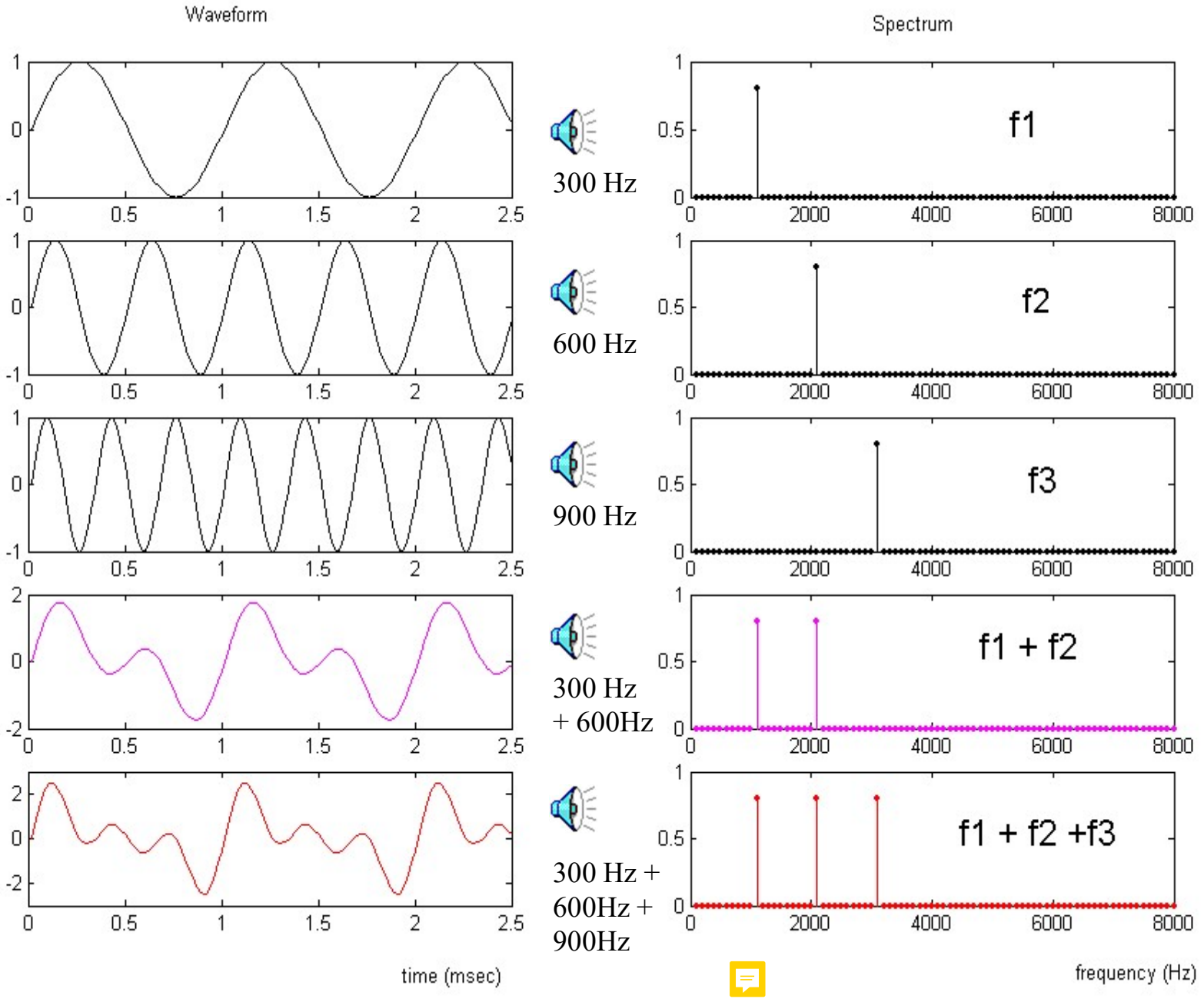
1 Hertz = 1 vibration/sec

Frequency = 300 Hz

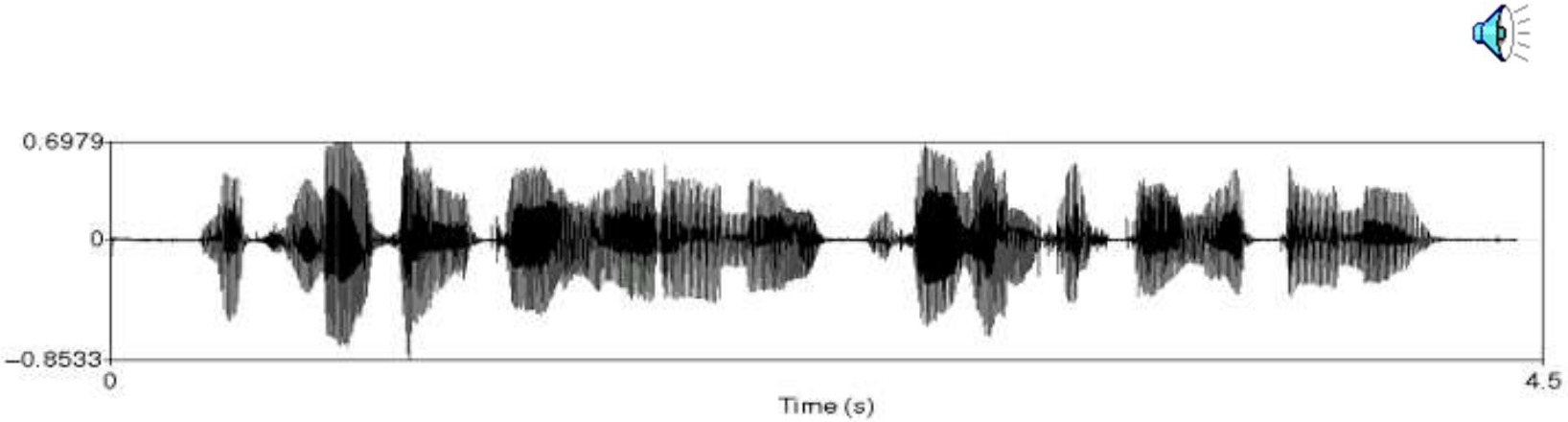
Components of sound

A sound is usually comprised of *several* frequency components.

Depending on the relationships of the frequency components, the sound can elicit a sensation of pitch.


Waveform
Spectrum
300 Hz
600 Hz
900 Hz
300 Hz
+ 600Hz
300 Hz +
600Hz +
900Hz
f1
f2
f3
f1 + f2
f1 + f2 +f3
time (msec)
frequency (Hz)

Speech “waveform”



“Information” in speech?

- Linguistic (message -> sentences -> words -> phonemes)

  The speech signal is characterised by an enormous range of elementary perceptually contrasting sounds!

- Paralinguistic:

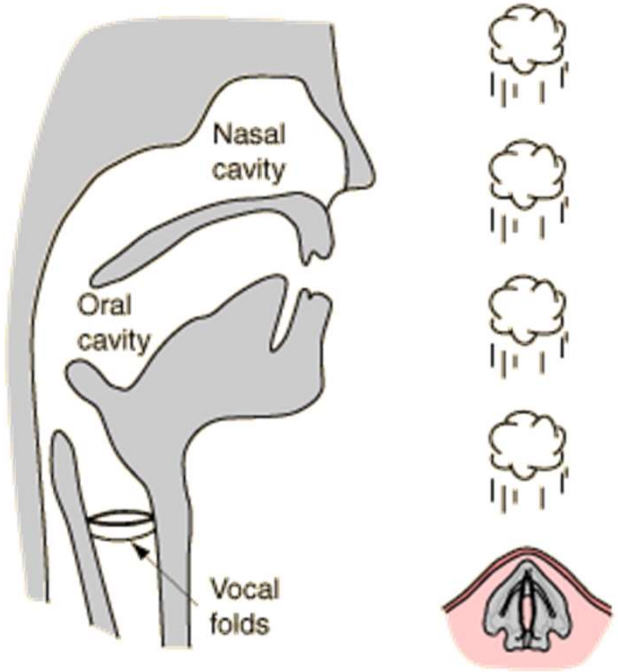
  --expressive (emotions, mood)

  --speaker-based (age, gender, accent and style)

“Everyday” speech technology

- Mobile telephony (speech compression)
- Human-computer interfaces (speech recognition/synthesis)
- Security (speaker identification in biometrics, forensics)
- Speech enhancement (improving intelligibility or quality)
- Behavioural analytics (monitoring well-being)

Generating speech*

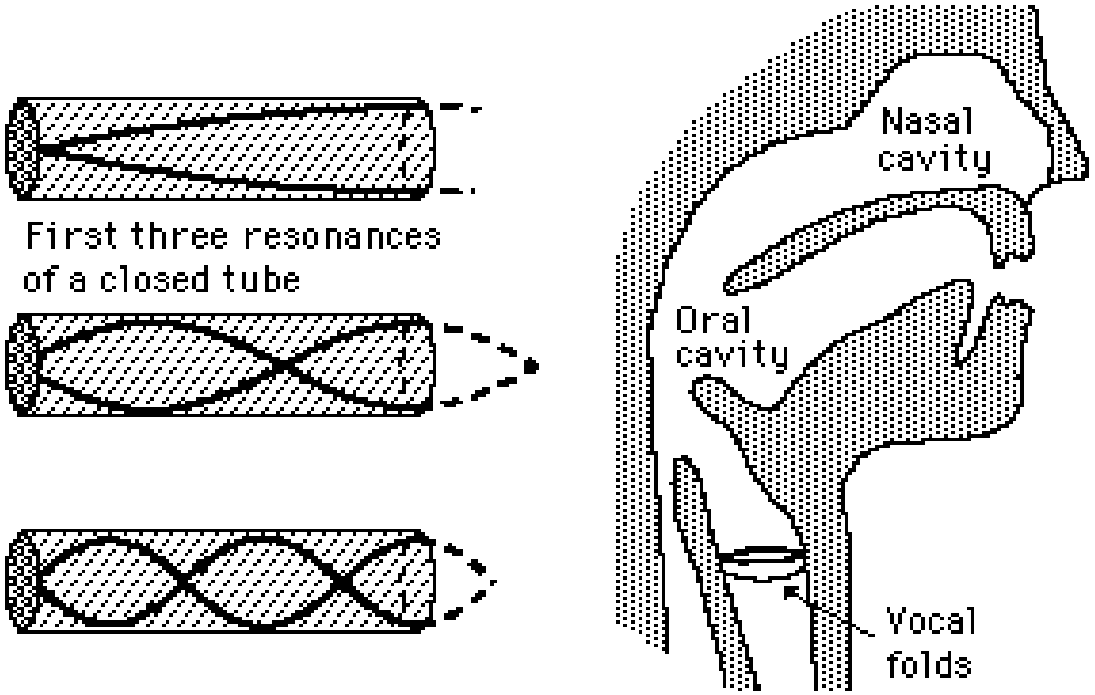


Respiration->**phonation** ->articulation

Vibrating vocal cords create puffs of air giving rise to *air pressure variations* which reach our ears.

**HyperPhysics, Sound and Hearing, Georgia State University*

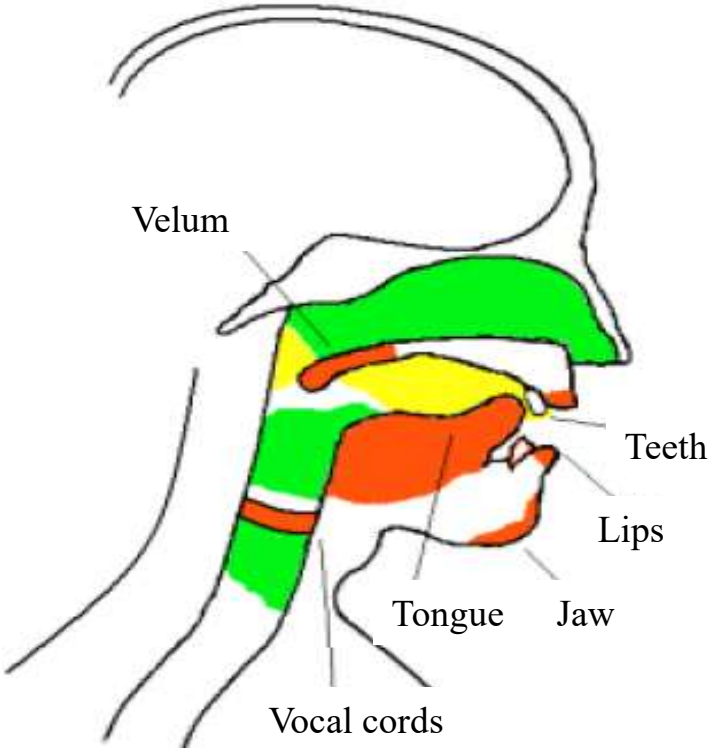
Vocal tract: Acoustic resonances*

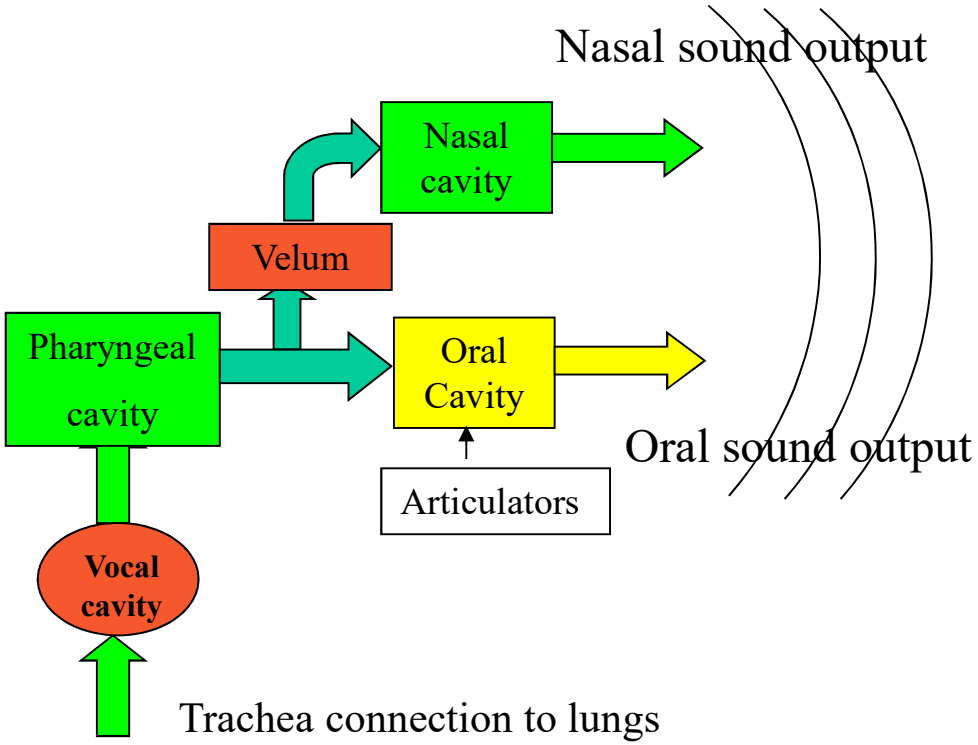


$$f_1 = \frac{c}{4L} \quad ; \quad f_2 = \frac{3c}{4L}; \quad f_3 = \frac{5c}{4L} \; ; \; \ldots\ldots$$

*HyperPhysics, Sound and Hearing, Georgia State University (http://hyperphysics.phy-astr.gsu.edu/hbase/sound/)

Articulation: producing the *various sounds* of speech*





***Securivox tutorial**

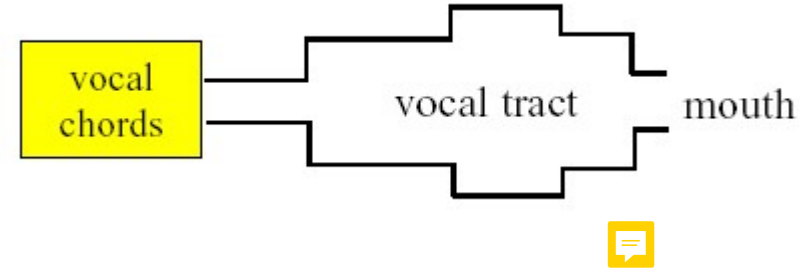
Moving muscles which alter the resonant cavities

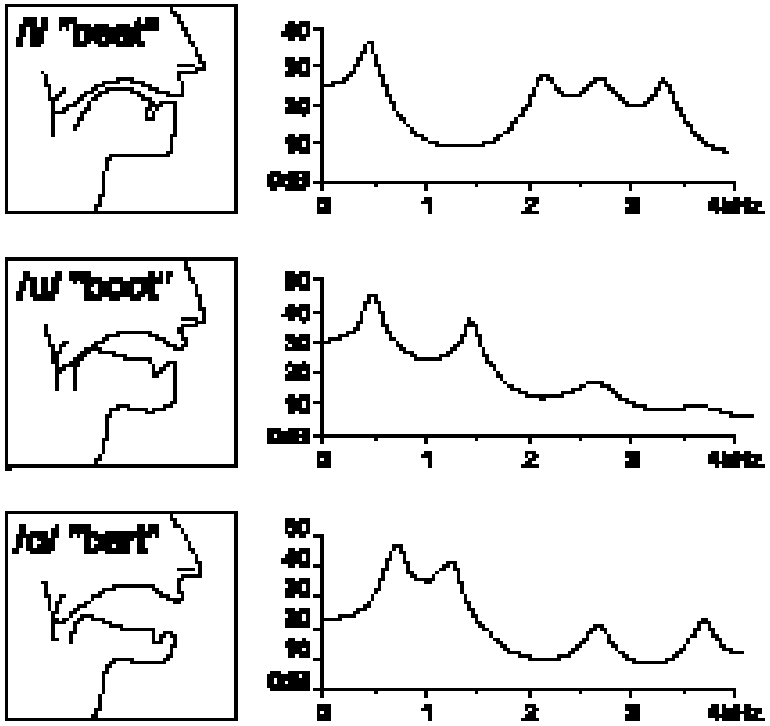
Dynamic cavity

Static cavity

Vocal tract “filter”*

- The sound spectrum is modified by the shape of the vocal tract.
- The resonant frequencies of the vocal tract cause peaks in the spectrum called *formants*.

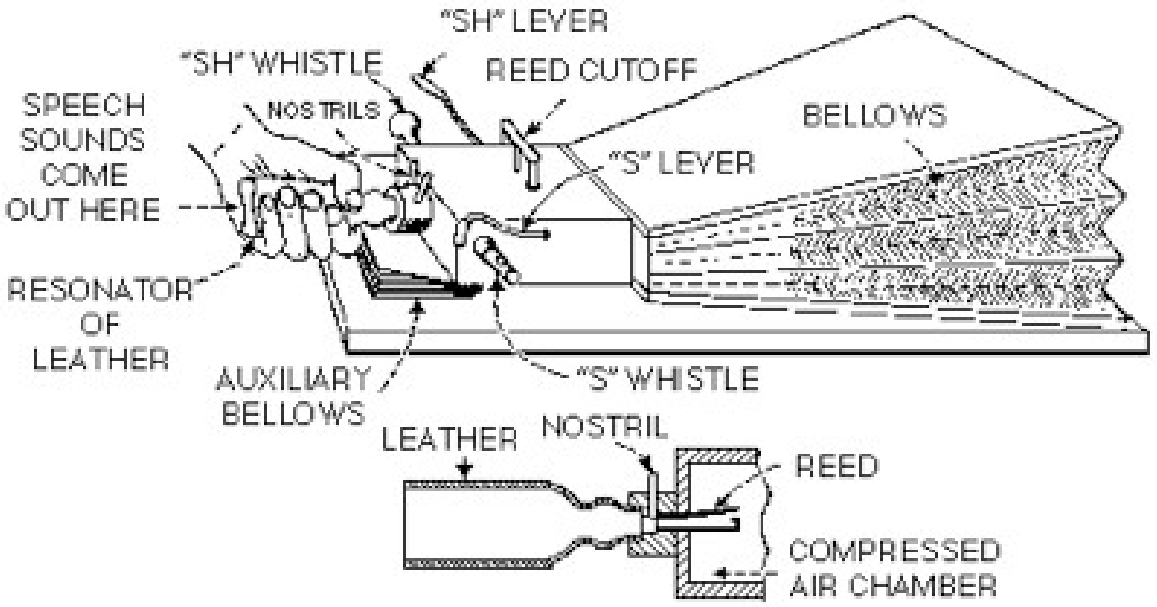




*Childers, Speech Overview

Von Kempelen's talking machine

1791



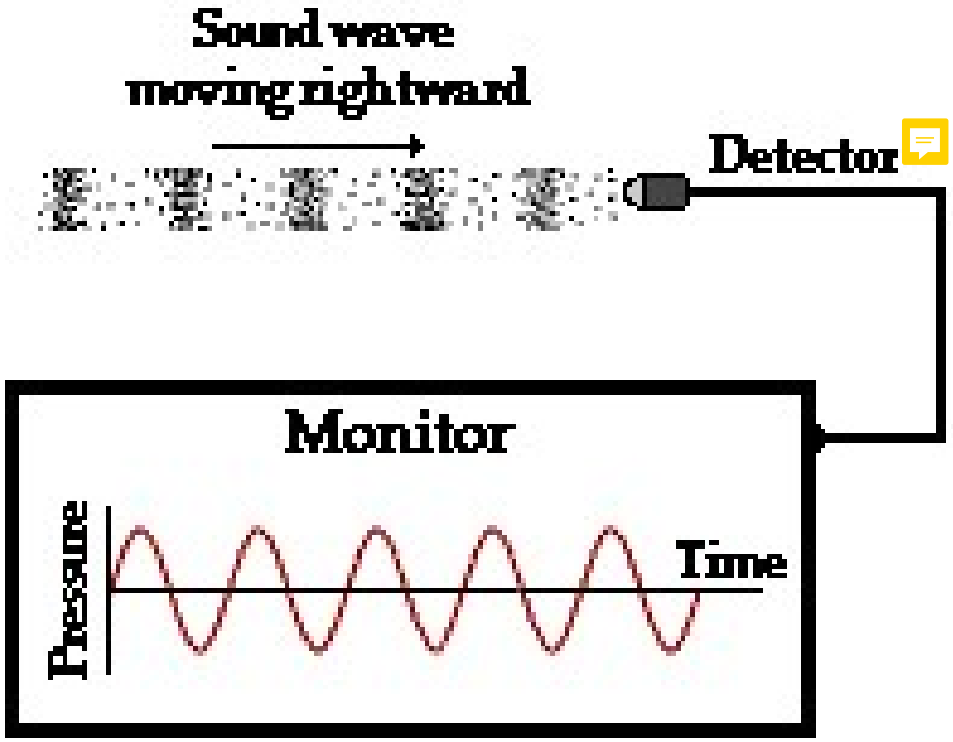
1875

- Alexander Bell invents the method of, and apparatus for, "transmitting vocal or other sounds telegraphically ... by causing electrical undulations, similar in form to the vibrations of the air accompanying the said vocal or other sound".

=> Major impetus to modern speech processing.

- 1930s: Electrical synthesis of speech by Dudley's vocoder

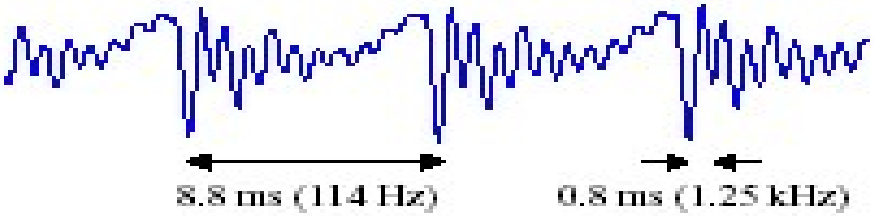
Sound -> electrical form*



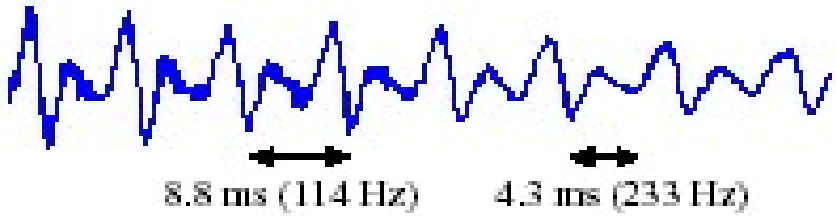
*The Physics Classroom:http://www.glenbrook.k12.il.us/gbssci/phys/Class/sound/u11l2a.html

Speech Waveforms from "my speech"

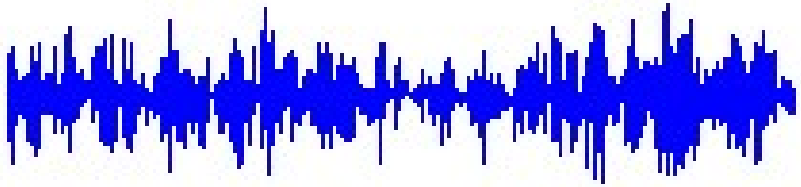
(a) start of "y" vowel



(b) "ee" vowel



(c) "s" consonant



Basic sounds of speech: Phones

- The speech signal can be divided into sound segments with *fixed articulation and acoustics over short intervals*.

  i.e. articulatory configuration <=> acoustic properties

  Smallest meaningful sound unit: "***phone***"

  (i.e. set of distinctive sounds of a language)

  In Indian written scripts, one symbol represents one phone.

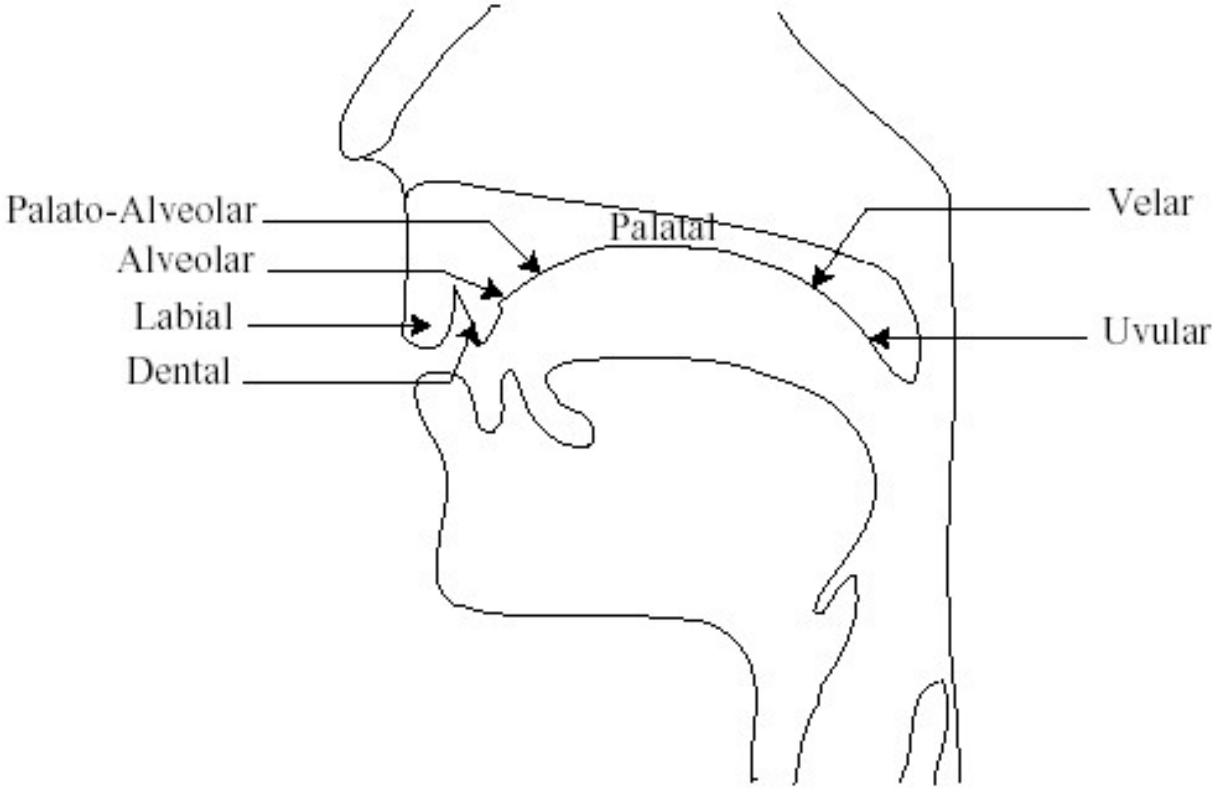
Classification of speech sounds

Vowels and Consonants

- Vowels: steady sounds specified by position of the articulators (typically, tongue)
- Consonants: are (dynamic) sounds classified by **place** and **manner** of articulation

Place of articulation

(*constriction of vocal tract*)



Pulmonic consonants:

	bilabial	labiodental	dental	alveolar	post-alveolar	retroflex	palatal	velar	uvular	pharyngeal	glottal
plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
lateral fricative				ɬ ɮ							
trill	ʙ			r					ʀ		
tap/flap		ⱱ		ɾ		ɽ					
central approximant		ʋ		ɹ		ɻ	j	ɰ			
lateral approximant				l		ɭ	ʎ	ʟ			

Physiology (articulator motion)

↓

Sound with specific acoustic characteristics (seen in waveform and spectrum)

↓

Perception of certain sound qualities

Speech production basics

- **Vocal cords** (larynx) modulate the airflow from the lungs by rapid opening-closing; the *rate of vibration* is determined by their mass and tension.

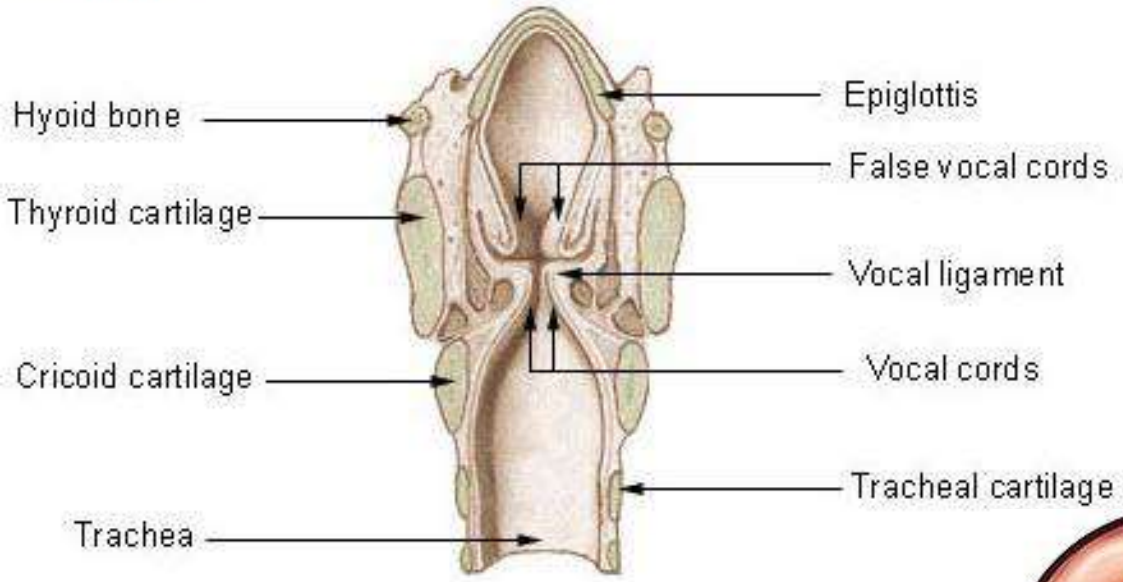
  Pitch frequency ranges:

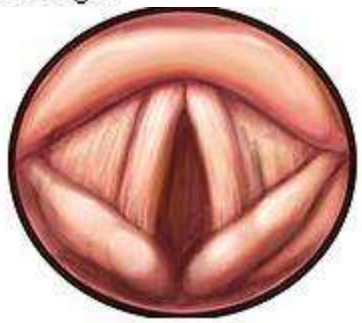
  male: 80-160 Hz; female:160-320 Hz;

  singers: over 2 octaves.

- **Vocal tract** shapes the vocal cord vibrations into the intricate sounds of speech via *changes in shape* to produce various *acoustic resonances*.

Larynx





Vocal cords open during breathing to allow air into lungs.

Vocal cords close when speaking so air from the lungs presses between them to cause the vibrations that produce sound.

Outline

- Speech production (physiology)
- Classification of sounds: articulatory, acoustic
- Speech signal representations and analyses (signal processing methods for information extraction)
- Hearing, and speech perception
- Speech technology (compression, ASR,TTS,…)
- Audio/music technology

Text / References

- Douglas O'Shaughnessy, Speech Communications: Human and Machine, Universities Press (India) Ltd., 2001
- Rabiner and Schafer, Digital Processing of Speech Signals
- IITB Moodle for all course-related hand-outs

Speech technology out there…

- **Speech recognition**. Systems for the conversion of speech to text, for spoken dialogue with computers or for executing spoken commands.
- **Speech synthesis**. Systems for converting text to speech or (together with natural language generation) concept to speech.
- **Speaker recognition**. Systems for identifying individuals or language groups by the way they speak.
- **Forensic speaker comparison**. Study of recordings of the speech of perpetrators of crimes to provide evidence for or against the guilt of a suspect.
- **Language pronunciation teaching**. Systems for the teaching and assessment of pronunciation, used in second language learning.
- **Assessment and therapy for disorders of speech and hearing**. Technologies for the assessment of communication disorders, for the provision of therapeutic procedures, or for communication aids.
- **Monitoring of well-being and mood**. Technologies for using changes in the voice to monitor physical and mental health.