Speech Understanding

Richa Singh

Speech Processing

- Speech is the most natural form of human-human communications.
- Speech is related to language; linguistics is a branch of social science.
- Speech is related to human physiological capability; physiology is a branch of medical science.
- Speech is also related to sound and acoustics, a branch of physical science.
- Therefore, speech is one of the most intriguing signals that humans work with every day.

Purpose of speech processing

- To understand speech as a means of communication;
- To represent speech for transmission and reproduction;
- To analyze speech for automatic recognition and extraction of information
- To discover some physiological characteristics of the talker.

Requirements

- Human example suggests, plenty
  - What was said
  - Who said it
  - When they said it

  - What it meant
  - How to respond

Tasks that can be performed

- Coding
- Synthesis/generation
- Recognition
- Emotion recognition
- Text to speech

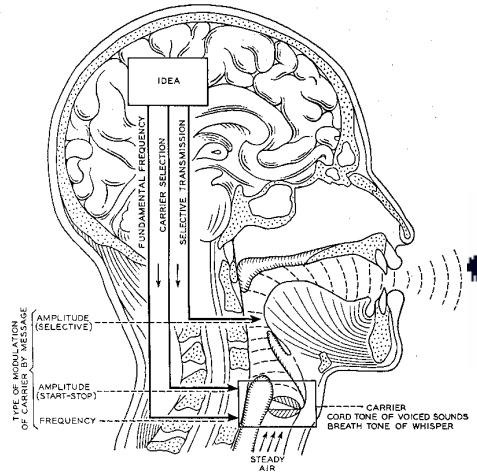
Why is it hard?

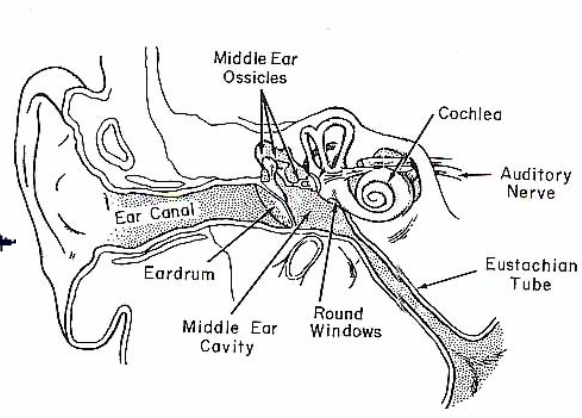
- Speaker variability (within and between)
- Noise, reverberation, channel
- Confusable vocabulary
- Meaning and tone

Automatic Speech Recognition

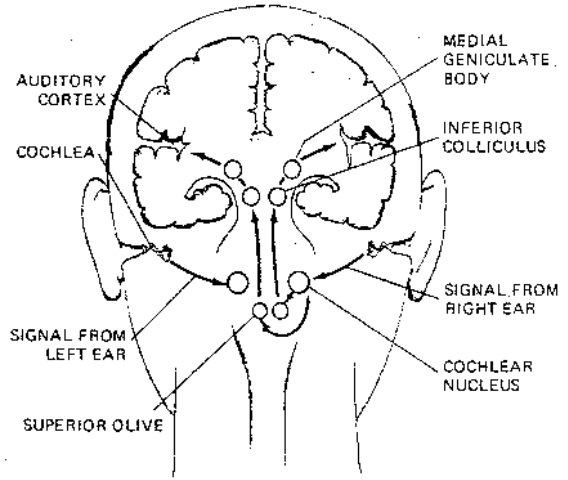
- What is the task?
- What are the main difficulties?
- How is it approached?
- How good is it?
- How much better could it be?

How do humans do it?





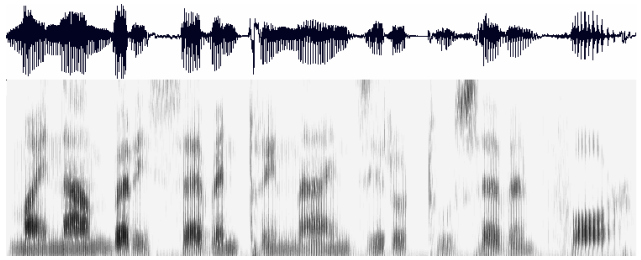
- Articulation produces sound waves which the ear conveys to the brain for processing

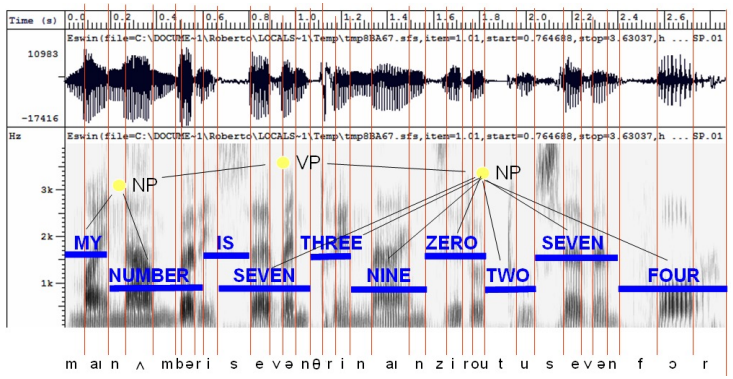


How might computers do it?

Acoustic waveform

Acoustic signal





Speech recognition

- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation

Phones

- In phonetics, (a branch of linguistics) a **phone** is any distinct speech sound or gesture, regardless of whether the exact sound is critical to the meanings of words.
- It is a speech segment that possesses distinct physical or perceptual properties and serves as the basic unit of phonetic speech analysis.

- A phoneme is a speech sound in a given language that, if swapped with another phoneme, could change one word to another. Phones are absolute and are not specific to any language, but phonemes can be discussed only in reference to specific languages.

https://en.wikipedia.org/wiki/Phone_(phonetics)

Phonetics

- ARPAbet
  - An alphabet for transcribing American English phonetic sounds.
- Articulatory Phonetics
  - How speech sounds are made by articulators (moving organs) in mouth.
- Acoustic Phonetics
  - Acoustic properties of speech sounds

Phonetics

- Modern systems are less reliant on encoding phonetic domain knowledge directly.
- Basic understanding helps with describing and debugging spoken language systems
  - E.g. how does an accent change the sound of pronunciations?
- Phonetic categories derived from how humans produce speech

- Competitive

Articulatory Phonetics

Sound is produced by the rapid movement of air. Most sounds in human spoken languages are produced by expelling air from the lungs through the windpipe (technically the trachea) and then out the mouth or nose through the larynx, commonly known as the Adam's apple or voicebox.

The larynx contains two small folds of muscle, the vocal folds (often referred to non-technically as the vocal cords) which can be moved together or apart. **The GLOTTIS space between these two folds is called the glottis.**

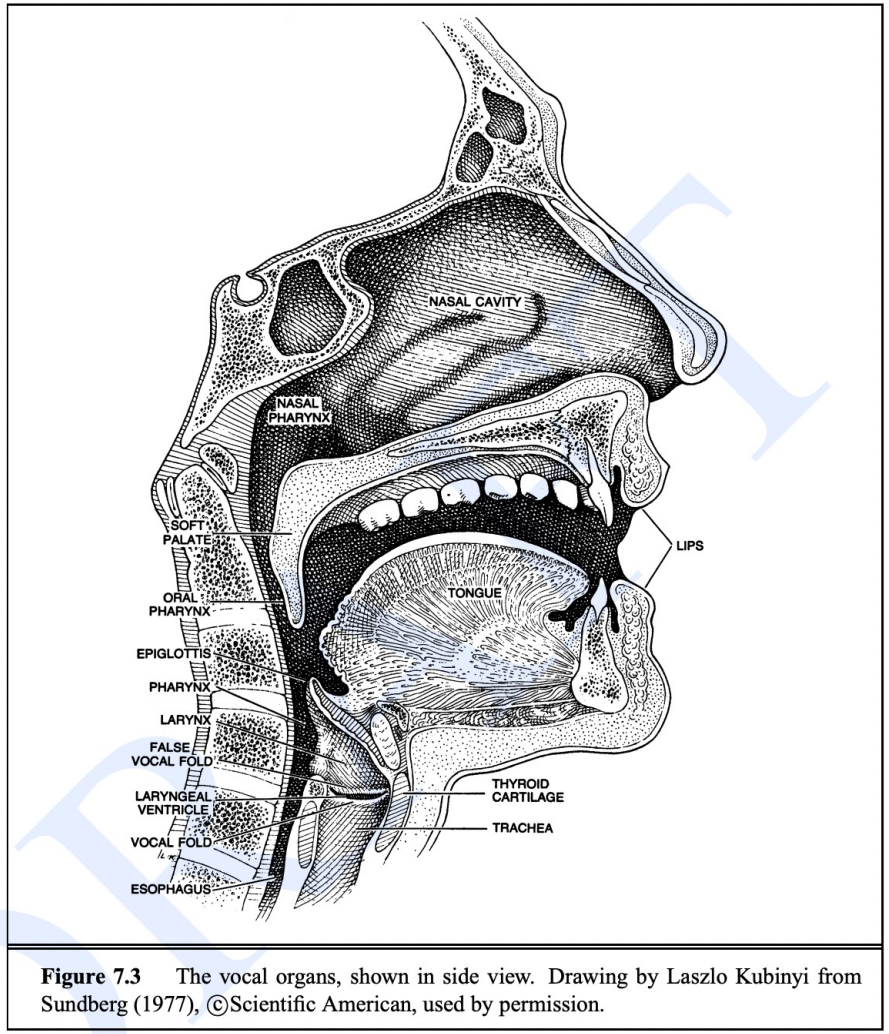


Figure 7.3 The vocal organs, shown in side view. Drawing by Laszlo Kubinyi from Sundberg (1977), ©Scientific American, used by permission.

https://pages.ucsd.edu/~bakovic/compphon/Jurafsky,%20Martin.-Speech%20and%20Language%20Processing_%20An%20Introduction%20to%20Natural%20Language%20Processing%20(2007).pdf

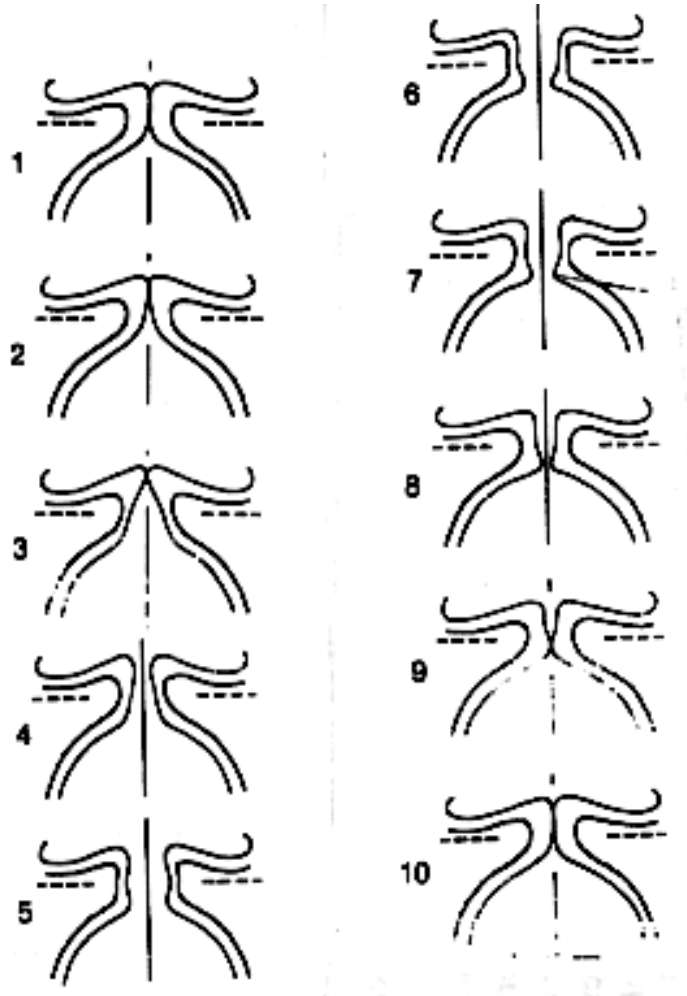
- Sounds made with the vocal folds together and vibrating are called voiced; sounds made without this vocal cord vibration are called unvoiced or voiceless.
- Voiced sounds include [b], [d], [g], [v], [z], and all the English vowels, among others.
- Unvoiced sounds include [p], [t], [k], [f], [s], and others
- Most sounds are made by air passing through the mouth.
- Sounds made by air passing through the nose are called nasal sounds, m, n, ng

Larynx and Vocal Folds

- The Larynx (voice box)
  - A structure made of cartilage and muscle
  - Located above the trachea (windpipe) and below the pharynx (throat)
  - Contains the vocal folds
  - (adjective for larynx: laryngeal)
- Vocal Folds (older term: vocal cords)
  - Two bands of muscle and tissue in the larynx
  - Can be set in motion to produce sound (voicing)

Text from slides by Sharon Rose UCSD LING 111 handout

Voicing:



- Air comes up from lungs
- Forces its way through vocal cords, pushing open (2,3,4)
- This causes air pressure in glottis to fall, since:
  - when gas runs through constricted passage, its velocity increases
  - this increase in velocity results in a drop in pressure
- Because of drop in pressure, vocal cords snap together again (6-10)
- Single cycle: ~1/100 of a second.

Figure & text from John Coleman's web site

Consonants and vowels

- Consonants are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced. Example: [p], [b], [t], [d], [k], [g], [f], [v], [s]
- Vowels have less obstruction, are usually voiced, and are generally louder and longer-lasting than consonants. Example: [aa], [ae], [ao], [ih], [aw], [ow], [uw],
- Semivowels (such as [y] and [w]) have some of the properties of both; they are voiced like vowels, but they are short and less syllabic like consonants.

Place of Articulation

- Consonants are classified according to the location where the airflow is most constricted.
- This is called place of articulation
- Three major kinds of place articulation:
  - Labial (with lips)
  - Coronal (using tip or blade of tongue)
  - Dorsal (using back of tongue)

Consonants: Place of Articulation

Labial: Consonants whose main restriction is formed by the two lips coming together have a bilabial place of articulation. In English these include [p] as in possum, [b] as in bear, and [m] as in marmot.

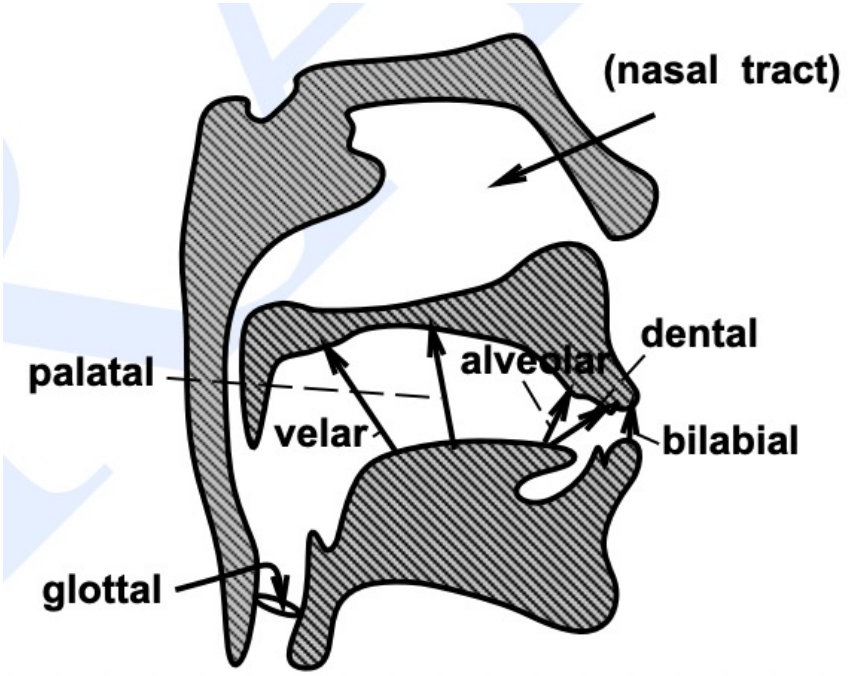
Dental: Sounds that are made by placing the tongue against the teeth are dentals. The main dentals in English are the [th] of thing or the [dh] of though,

Alveolar: The alveolar ridge is the portion of the roof of the mouth just behind the upper teeth. Most speakers of American English make the phones [s], [z], [t], and [d] by placing the tip of the tongue against the alveolar ridge.

Palatal…

Velar….

Glottal ….



Manner of Articulation

- Stop: complete closure of articulators, so no air escapes through mouth
- Oral stop: palate is raised, no air escapes through nose. Air pressure builds up behind closure, explodes when released
  - p, t, k, b, d, g
- Nasal stop: oral closure, but palate is lowered, air escapes through nose.
  - m, n, ng

Articulatory parameters for English consonants (in ARPAbet)

MANNER OF ARTICULATION	PLACE OF ARTICULATION													
	bilabial		labio-dental		inter-dental		alveolar		palatal		velar		glottal	
stop	p	b					t	d			k	g	q	
fric.			f	v	th	dh	s	z	sh	zh			h	
affric.									ch	jh				
nasal		m						n				ng		
approx		w						l/r		y				
flap							dx							

VOICING: voiceless | voiced

Table from Jennifer Venditti