Speech Processing

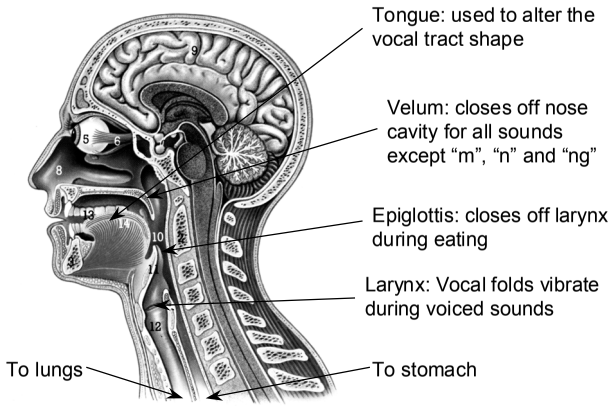
Speech is the way of choice for humans to communicate:

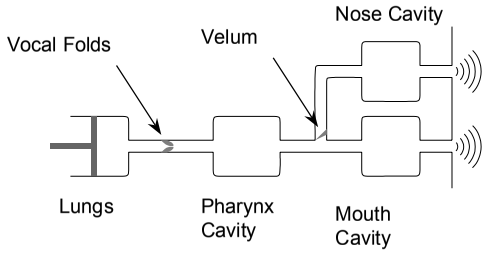
- no special equipment required
- no physical contact required
- no visibility required
- can communicate while doing something else

Speech Processing:

- Coding
- Synthesis
- Recognition
- Identity Verification
- Enhancement

The Vocal Tract





Sources of Sound Energy

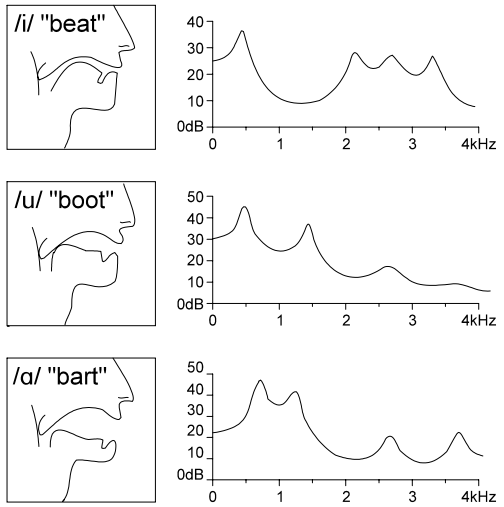
- **Turbulence**: air moving quickly through a small hole (e.g./s/ in “size”)
- **Explosion**: pressure built up behind a blockage is suddenly released (e.g. /p/ in “pop”)
- **Vocal Fold Vibration**: like the neck of a balloon (e.g. /a/ in “hard”)
  - airflow through vocal folds (vocal cords) reduces the pressure and they snap shut (Bernoulli effect)
  - muscle tension and air pressure buildup force the folds open again and the process repeats
  - frequency of vibration (fx) determined by tension in vocal folds and pressure from lungs
  - for normal breathing and voiceless sounds (e.g. /s/) the vocal folds are held wide open and don’t vibrate

Speech Sound Categories

- **Voiced**: speech sounds where the vocal folds vibrate.
- **Vowels**: no blockage of the vocal tract and no turbulence
- **Consonants**: non-vowels
- **Plosives**: consonants invovling an explosion

Vocal Tract Filter

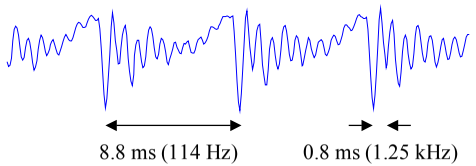
- The sound spectrum is modified by the shape of the vocal tract. This is determined by movements of the jaw, tongue and lips.
- The resonant frequencies of the vocal tract cause peaks in the spectrum called *formants*.
- The first two formant frequencies are roughly determined by the distances from the tongue hump to the larynx and to the lips respectively.



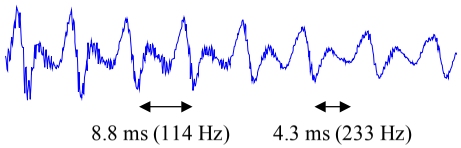
Speech Waveforms

Extracts from “my speech”

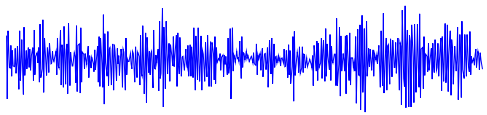
(a) start of “y” vowel



(b) “ee” vowel

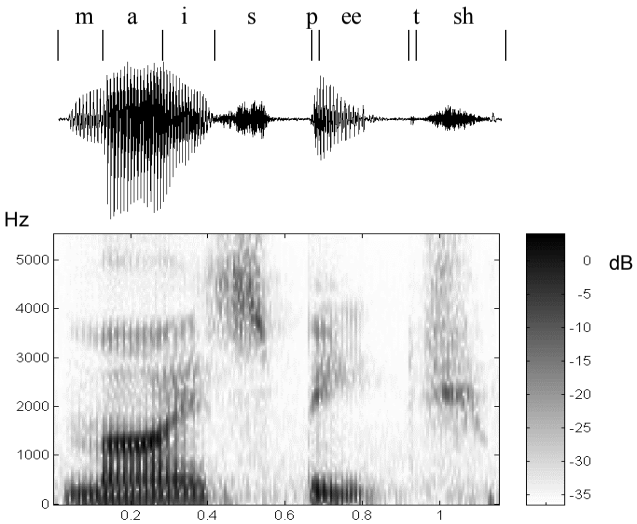


(c) “s” consonant



Speech Spectrogram

“my speech”



- Dark areas of spectrogram show high intensity
- Voiced segments are much louder than unvoiced
- Horizontal dark bands are the formant peaks
- “s” very high frequency of around 4.5 kHz (compare with telephone bandwidth: 0.5-3.4 kHz)
- “sh” is lower frequency because tongue is further back
- Vertical bands in “my” are individual larynx closures
- The “y” of “my” is a diphthong: two successive vowels

Phonemes

- Speakers and listeners divide words into component sounds called *phonemes*.
  - Native speakers agree on the phonemes that make up a particular word
  - There are about 42 phonemes in English

- The phonemes in a particular word may vary with dialect
  - High amplitude speech will mask noise at the same frequency
- The actual sound that corresponds to a particular phoneme depends on:
  - the adjacent phonemes in the word or sentence
  - the accent of the speaker
  - the talking speed
  - whether it is a formal or informal occasion

Speech Coding

What: To transmit/store a speech waveform using as few bits as possible while retaining high quality

Why: To save bandwidth in telecoms applications and to reduce memory storage requirements.

How:

- Correlation ⇒ Predictability ⇒ Redundancy
  - Predict waveform samples from previous samples and transmit only the prediction error
  - Autocorrelation is fourier transform of power spectrum: a peaky spectrum ⇒ strong short-term correlations (~ 0.5 ms)
  - Voiced speech is almost periodic ⇒ strong long-term correlations (~ 10 ms)
- Devote few bits to the aspects of speech where errors are least noticeable
  - High amplitude speech will mask noise at the same frequency
- Ignore aspects of the speech that are inaudible
  - Power spectrum is much more important tha precise waveform
  - For aperiodic sounds, the fine detail of the spectrum does not matter

Speech Synthesis

What: To convert a text string into a speech waveform

Why: For technology to communicate when a display would be inconvenient because:

(a) Too big, (b) Eyes busy, (c) Via phone,
(d) In the dark, (e) Moving around

Problems:

- The spelling of words doesn't match their sound
  - Pronounciation rules + an exceptions dictionary
- Some words have multiple meanings+sounds
  - Must guess which is the correct sound
- Simplistic speech models sound mechanical
  - Can use extracts from real speech
- Speech sounds are influenced by adjacent phonemes
  - Use phoneme pairs from real speech
- Important words must be slightly louder
  - Must try to understand the tex a nit
- Voice pitch and talking speed must vary smoothly throughout a sentence
  - Must be able to change pitch and speed without affecting formant frequencies

Speech Recognition

What: To convert a speech waveform into text

Why: To communicate and control technology when a keyboard would be inconvenient because:

(a) Too big, (b) Hands busy, (c) Via phone,
(d) In the dark, (e) Moving around

Problems:

- The spelling of words doesn't match their sound
  - Have a big phonetic dictionary
- The waveform of a word varies a lot between different speakers (or even the same speaker)
  - Extract *features* from the speech waveform that are more consistent than the waveform
- The extracted features won't be exactly repeatable
  - Characterise them with a probability distribution
- Speech sounds are influenced by adjacent phonemes
  - Use context-dependent probability distributions
- Speaking speed varies enormously
  - Try all possible speaking speeds
- No clear boundary between words or phonemes
  - Try all possible boundaries

Syllabus

Lectures

2	Modelling Speech Production Acoustics
3	Time/Frequency Representation
4	Properties of Digital Filters
5-7	Linear Predictive Modelling
8-10	Speech Coding
11	Phonetics
12-13	Speech Synthesis
14-19	Speech Recognition

Books

- "Discrete-Time Processing of Speech Signals", JR Deller, Jr, JG Proakis & JHL Hansen, Macmillan 1993, 0-02-328301-7
  Comprehensive and quite good but has a few errors.
- "Digital Processing of Speech Signals", LR Rabiner & RW Schafer, Prentice-Hall 1978, 0-13-213603-1
  Excellent treatment of linear prediction, too old for coding, synthesis and recognition.
- "Statistical Methods for Speech Recognition",F Jelinek, MIT Press 1998, 0-262-10066-5
  Excellent treatment of theory underlying recognition.

Website

- http://www.ee.ic.ac.uk/hp/stafff/dmb/dmb.html has numerous speech-related MATLAB routines + copies of these notes