TTIC 31110
Speech Technologies

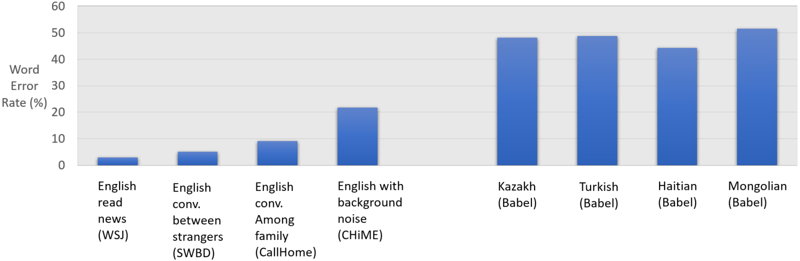
April 9, 2020

Announcements

- Please fill out survey
- Lecture 1 slides, Lecture 1 recording, “Homework 0”, Homework 1, readings are posted
- First tutorial (probability review, multivariate Gaussians): Monday 4/13 3:30-4:30pm
- HW1:
  - Due Friday 4/17 7pm
  - Should be able to do some of it now, some will be easier after Tuesday’s lecture

Questions?

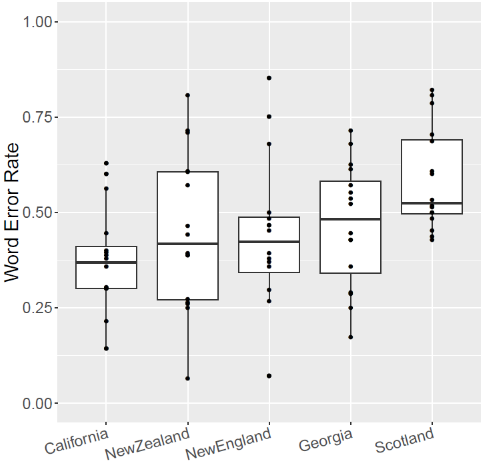
Recap: Some recent word error rates



https://github.com/syhw/wer_are_we

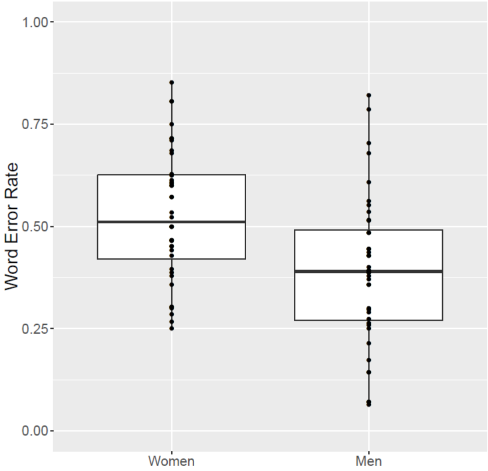
S. Dalmia *et al.*, "Sequence-based multi-lingual low resource speech recognition," ICASSP 2018.

Dependence on dialect



(Fig. from [Tatman 2017])

Dependence on gender



(Fig. from [Tatman 2017])

Questions?

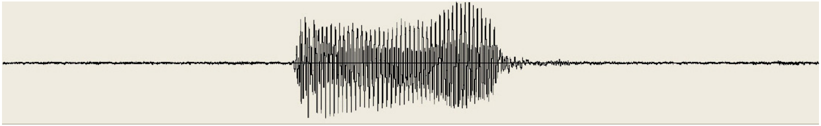
Outline

A “simple” speech recognition task

Speech production and perception

Single-digit classification

Given a 1-second speech waveform, determine which digit (0-9) was spoken



What are we looking at?

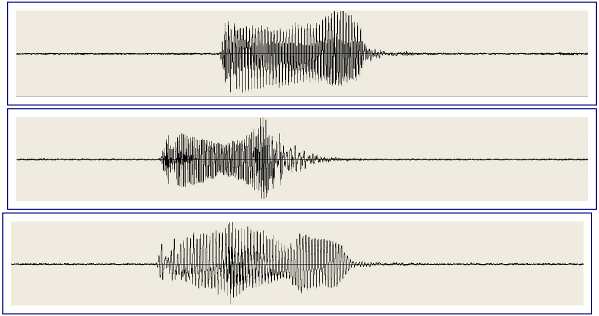
- Recording from a microphone: instantaneous air pressure vs. time
- Discretized in time (in this case, to 16,000 samples, i.e. sampling rate of 16kHz)
- Discretized in magnitude (in this case, to 16 bits per sample)
- Result: 16,000-dimensional vector, e.g. $a(t) = [3, 16, -1, 0, 427, 29, \ldots]$

Idea 1

- Record an example ("template") of each digit, $a_c(t), c \in [0, \ldots, 9]$.
- For test waveform $a(t)$, compute Euclidean distance to each template, $\mathrm{dist}_c = \sqrt{\sum_{t=1}^{16000} (a(t) - a_c(t))^2}$.
- Pick digit c with minimum dist_c

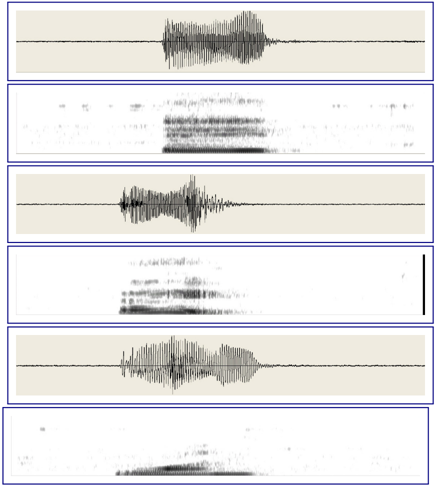
This is hard!

Which two are the same digit?



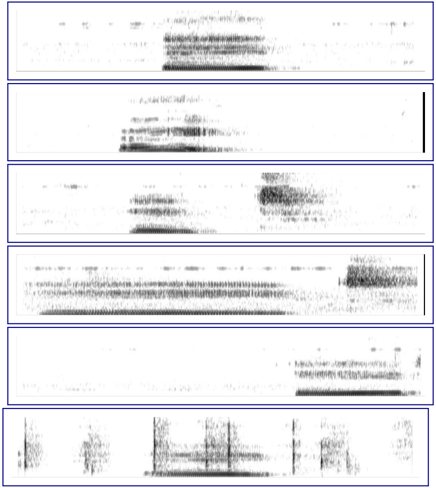
Idea 2: Go to frequency domain

Spectrogram: Fourier transform over short windows (e.g. 20ms)
$\longrightarrow$ plot of energy at each frequency over time $f_1(\omega), f_2(\omega), \ldots$



This is still hard!

Several examples of the digit "eight"



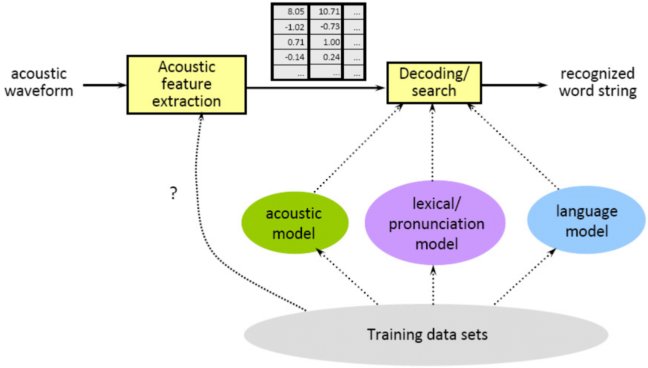
What is the problem?

- Finding a useful set of features $a(t) \to f_1, f_2, \ldots$ is difficult
- Many sources of variation:

acoustic:	channel, noise, vocal tract differences, pitch
phonetic:	*eight* $\longrightarrow$ [ey tcl] vs. [ey tcl t]
phonological:	*eight* before vowel $\longrightarrow$ [ey dx] *gas shortage*, *fish sandwich*
dialect:	*either* $\longrightarrow$ [iy dh er] vs. [ay dh er] *pin*, *pen*; *Mary*, *marry*, *merry*
coarticulation:	*she*, *shoe*
semantics:	*the baby cried* vs. *the Bay Bee cried*
situational context:	*it is easy to recognize speech* vs. *it is easy to wreck a nice beach*

Architecture of traditional speech recognizer

Must take into account (and take advantage of!) sources of variation/constraint

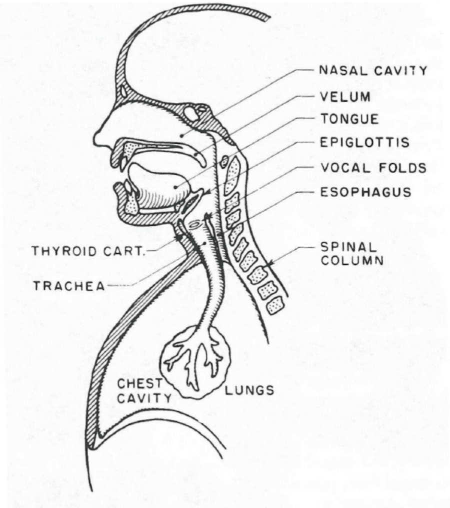


We'll start with feature extraction

- Traditionally, inspired by speech production/perception

Questions?

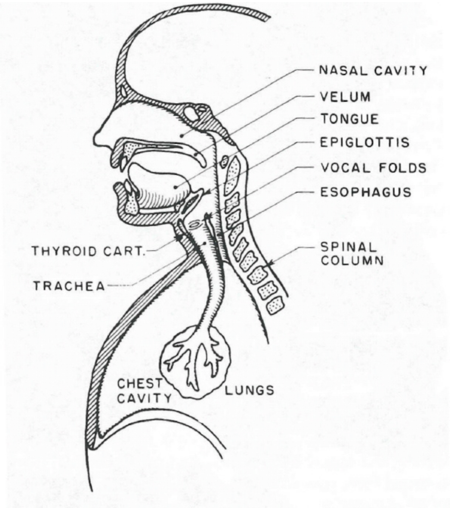
Speech production



(fig. from [Flanagan 1972])

- Air pushed up from lungs through vocal folds (glottis)
- Vocal folds: tensed for voiced sounds, spread for voiceless
- Tongue, lips, velum, nasal cavity form a “resonance chamber”

Speech production



(fig. from [Flanagan 1972])

- *Source-filter model*: Vocal tract acts as a filter, modulating the spectrum of the source signal
- Source is air pressure waveform either from glottis (e.g. vowels) or from another constriction (most consonants)

Demo!

Phonemes

- *Phonemes*: basic speech sounds that can be used to distinguish words
- *Allophones*: variants of phonemes in context, e.g.
  - aspirated vs. unaspirated [p]: *pin* vs. *spin*
  - nasalized vs. non-nasalized [ae]: *cat* vs. *can't*
  - No pair of words in English are distinguished by these pairs, so they are not phonemes (but in some languages they are!)
- We will use the term *phones* to refer to all basic sound units
- Always enclosed in square brackets

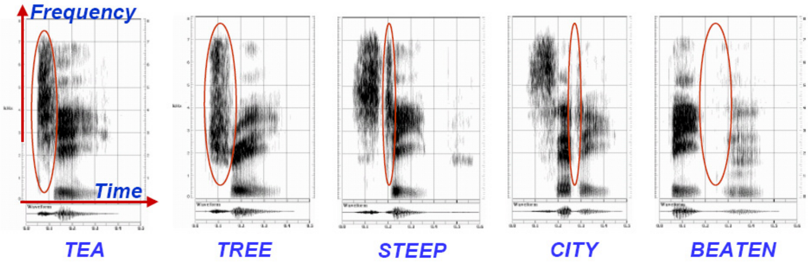
Phonemes of English

English has 40-50 "canonical" phones. Can be organized by *manner of articulation*:

vowels (~18)	[aa, ae, eh, iy, uw, ow, uh, ah, ...]
fricatives (8)	[f, v, th, dh, s, z, sh, zh]
stops (6)	[p, b, t, d, k, g]
nasals (3)	[m, n, ng]
semivowels (liquids, glides) (4)	[l, r, y, w]
affricates (2)	[ch, jh]
aspirant (1)	[h]

There are multiple "phonetic alphabets" in use. In this course you will see ARPAbet (this slide) and International Phonetic Alphabet (IPA). (See the cheat sheet in this week's readings!)

Allophones of the phoneme [t]



(figs. from MIT 6.345 Spring '03, OpenCourseWare `http://ocw.mit.edu`)

Other dimensions of organization of phonemes

Place of articulation (for consonants):

- labial [b, p, m, w]
- labio-dental [f, v]
- interdental [th, dh]
- alveolar [t, d, s, z, n]
- palato-alveolar [sh, zh, ch, jh]
- palatal [y]
- velar [k, g, ng]

Voicing (for consonants):

- Refers to vibrating vs. non-vibrating vocal folds
- Voiced sounds are those that have a pitch
- voiced: [b, d, g, z, dh, v]
- voiceless/unvoiced: [p, t, k, s, th, f, h]

Other dimensions of organization of phonemes

Nasality (for consonants):

- Refers to airflow through nasal cavity vs. oral cavity
- nasal: [m, n, ng]
- non-nasal: all others

Vowel height:

- low [aa, ae]
- mid [eh, ah]
- high [iy, uw]

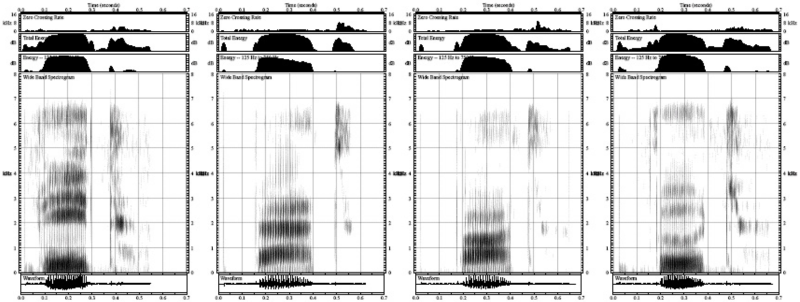
Vowel front/backness:

- front [iy, ae]
- mid [ah, ax]
- back [aa, uw]

Others...

Video: Ken Stevens' speech production

Spectrograms of the cardinal vowels

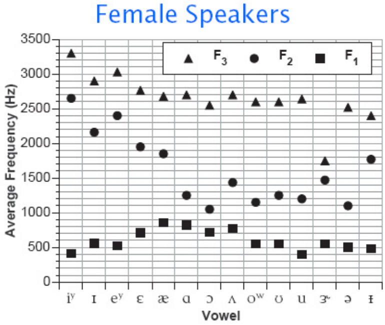


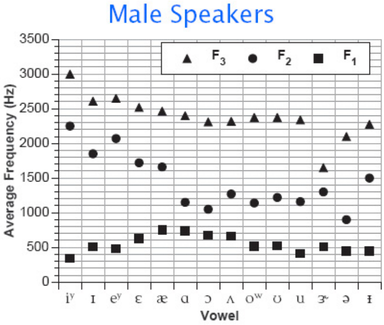
beet bat bott boot

(figs. from MIT 6.345 Spring '03, OpenCourseWare `http://ocw.mit.edu`)

Vowel formant frequencies

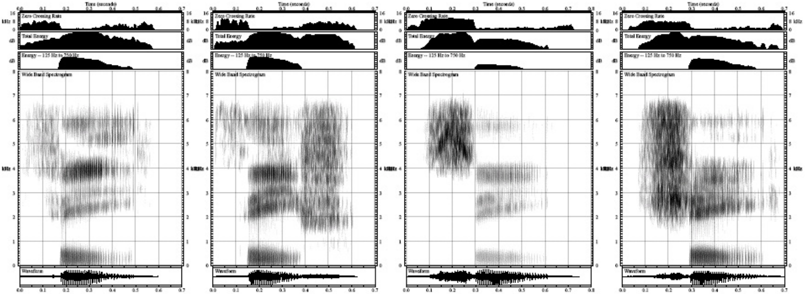
Vowels can be characterized by the first three *formants*: resonant frequencies of the vocal tract





(figs. from MIT 6.345 Spring '03, OpenCourseWare `http://ocw.mit.edu`)

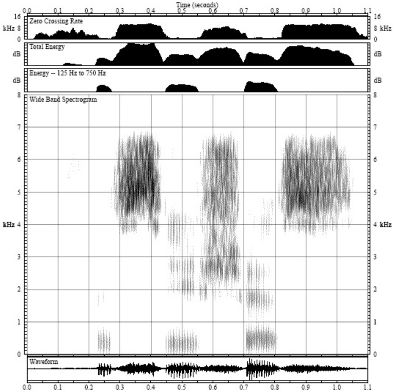
Spectrograms of English fricatives



fee thief see she

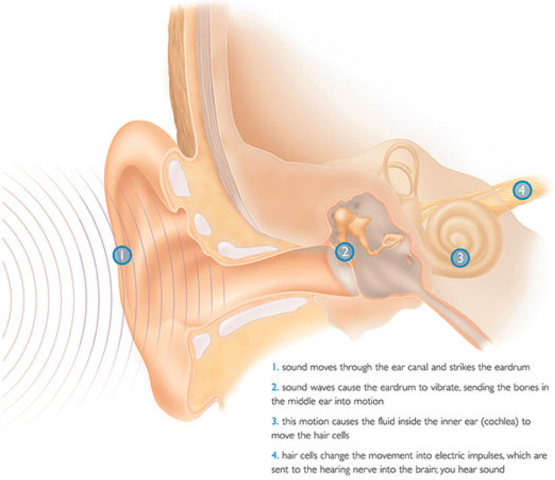
(figs. from MIT 6.345 Spring '03, OpenCourseWare `http://ocw.mit.edu`)

What is this word?

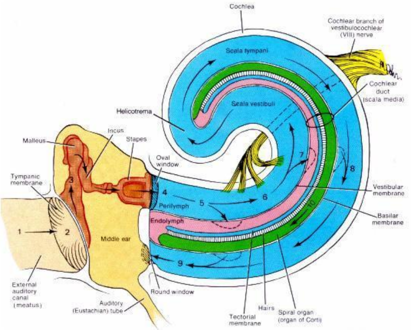


(fig. from MIT 6.345 Spring '03, OpenCourseWare `http://ocw.mit.edu`)

Physiology of hearing



Physiology of hearing (cont'd)



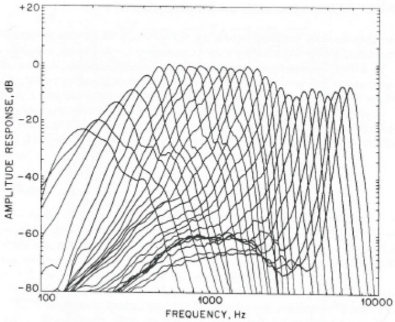


Figure 3.50 Frequency response curves of a cat's basilar membrane (after Ghitza [13]).

- Hairs on basilar membrane have different frequency responses

Perception

- Humans are less sensitive to differences in frequency at high frequencies than at low frequencies
- I.e., our internal "frequency axis" is not linear
- Stevens and Volkmann (1972) measured this warping with a set of perceptual experiments
- Result is the mel scale: $f_{\text{mel}} = 2595 \log_{10}(1 + f/700)$

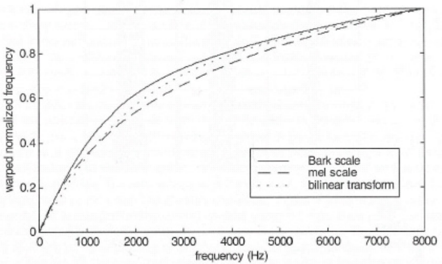
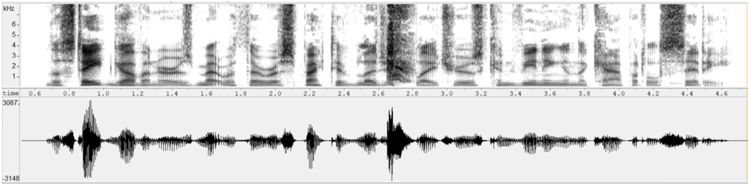


Figure 2.13 Frequency warping according to the Bark scale, ERB scale, mel-scale, and bilinear transform for $\alpha = 0.6$: linear frequency in the x-axis and normalized frequency in the y-axis.

Higher-level speech perception phenomena

Some less-obvious features of our speech perception facility...



The state governors met with their respective legislatures convening in the capital city.

- Did you hear the cough?
- Where was it?

Phonemic restoration effect:

Warren, Richard M., "Perceptual Restoration of Missing Speech Sounds." *Science* **167**(3917):392–3, 1970.

More phonemic restoration

- The *eel was on the orange.
- The *eel was on the axle.
- The *eel was on the shoe.