Computational Linguistics 1

CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland

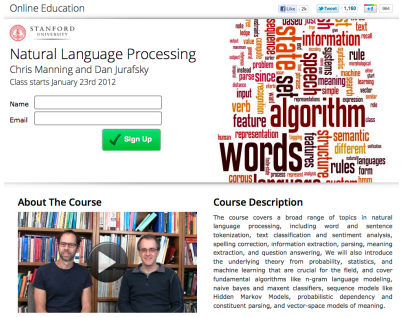
Lecture 23: 29 November 2011

Agenda

- HW5 grades
  - Online this evening
  - Pickup hard copies from Alex or next class
- HW7 "decision" due today
- HW6 due next Tuesday
  - Example using WordNet
- Course evals
- Online NLP course @ Stanford
- Questions, comments, concerns?
- Speech Recognition (ASR)
- Text-to-Speech (TTS)

Course Evaluations

Stanford NLP http://www.nlp-class.org/



Automatic Speech Recognition (ASR)

- IP notice: All following slides are from John-Paul Hosom, lectures 1 & 6 of ASR class at OHSU

Why is speech recognition difficult?

- Speech is:
  - Time-varying signal,
  - Well-structured communication process,
  - Depends on known physical movements,
  - Composed of known, distinct units (phonemes),
  - Modified when speaking to improve signal to noise ratio (SNR) (Lombard).

⇒ ***should be easy.***

Why is speech recognition difficult?

- However, speech:
  - Is different for every speaker,
  - May be fast, slow, or varying in speed,
  - May have high pitch, low pitch, or be whispered,
  - Has widely-varying types of environmental noise,
  - Can occur over any number of channels,
  - Changes depending on sequence of phonemes,
  - Changes depending on speaking style ("clear" vs. "conv.")
  - May not have distinct boundaries between units (phonemes),
  - Boundaries may be more or less distinct depending on speaker style and phoneme class,
  - Changes depending on the semantics of the utterance,
  - Has an unlimited number of words,
  - Has phonemes that can be modified, inserted, or deleted

Why is speech recognition difficult?

- To solve a problem requires in-depth understanding of the problem.
- A data-driven approach requires (a) knowing what data is relevant and what data is not relevant, (b) that the problem is easily addressed by machine-learning techniques, and (c) which machine-learning technique is best suited to the behavior that underlies the data.
- Nobody has sufficient understanding of human speech recognition to either build a working model or even know how to effectively integrate all relevant information.
- This lecture: present some of what is known about speech; motivate use of HMMs for Automatic Speech Recognition (ASR).

Speech Production

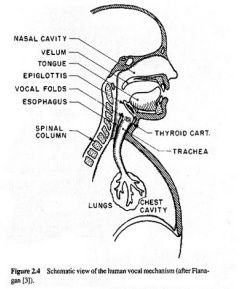


Figure 2.4 Schematic view of the human vocal mechanism (after Flanagan [3]).

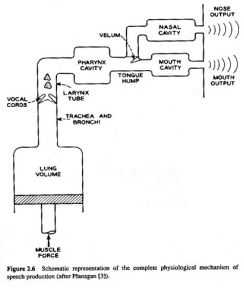


Figure 2.6 Schematic representation of the complete physiological mechanism of speech production (after Flanagan [3]).

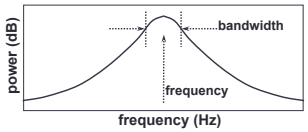
The Speech Production Process (from Rabiner and Juang, pp.16,17)

Speech Production

- Sources of Sound:
  - Vocal cord vibration
    - voiced speech (/aa/, /iy/, /m/, /oy/)
  - Narrow constriction in mouth
    - fricatives (/s/, /f/)
  - Airflow with no vocal-cord vibration, no constriction
    - aspiration (/h/)
  - Release of built-up pressure
    - plosives (/p/, /t/, /k/)
  - Combination of sources
    - voiced fricatives (/z/, /v/), affricates (/ch/, /jh/)

Speech Production

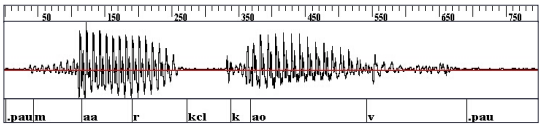
- Vocal tract creates resonances:



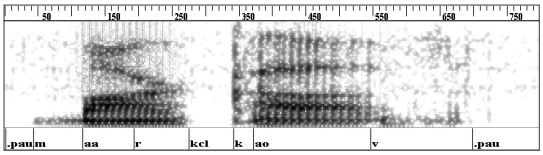
- Resonant energy based on shape of mouth cavity and location of constriction. Direct mapping from mouth shape to resonances.
- Frequency location of resonances determines identity of phoneme
- This implies that a key component of ASR is to create a mapping from observed resonances to phonemes. However, this is only one issue in ASR; another important issue is that ASR must solve both phoneme identity and phoneme duration simultaneously.
- Anti-resonances (zeros) also possible in nasals, fricatives

Representations of Speech

Time domain (waveform):

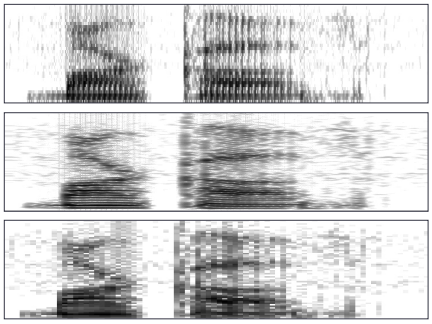


Frequency domain (spectrogram):



Representations of Speech

Spectrogram Displays:



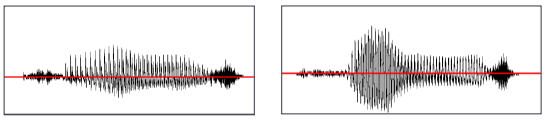
frame=0.5
win. = 7

frame=.5
win. = 34

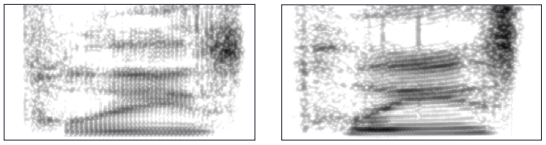
frame=10
win. = 16

Representations of Speech: Male vs Female

Time domain (waveform):



Frequency domain (spectrogram):

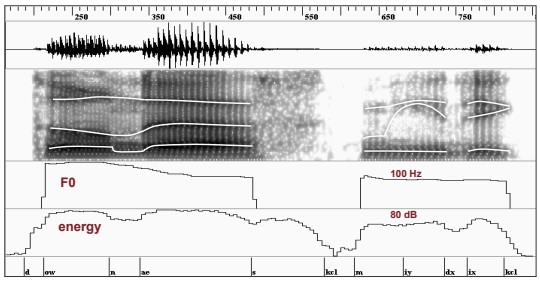


"please": male speaker

"please": female speaker

(from TIMIT sentence SX79.wav)

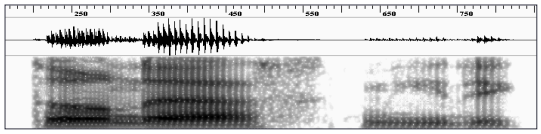
Representations of Speech: Pitch, Energy, Formants



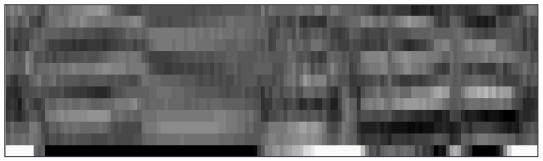
***F0* or Pitch:**
rate of vibration of vocal cords

Energy: $E = \frac{\sum_{i=1}^{N} x^2(i)}{N}$ or $\frac{\sum_{i=1}^{N} (x(i) \cdot h(i))^2}{N}, h(i) = 0.54 - 0.46\cos(\frac{2\pi i}{N-1})$

Representations of Speech: Cepstral Features

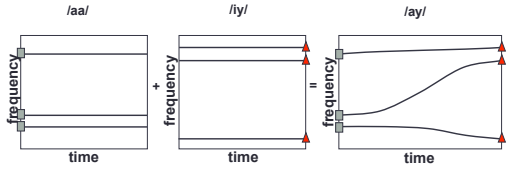


Cepstral domain (Perceptual Linear Prediction, Mel Frequency Cepstral Coefficients):



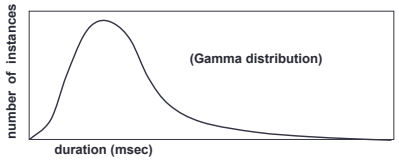
Time-Domain Aspects of Speech

- Coarticulation
  - Tongue moves gradually from one location to the next
  - Formant frequencies change smoothly over time
  - No distinct boundary between phonemes, especially vowels
  - Dynamics change as a function of speaking style
  - Dynamics as a function of duration not modeled well by linear stretching



Time-Domain Aspects of Speech

- Duration modeling
  - Rate of speech varies according to speaker, speaking style, etc.
  - Some phonetic distinctions based on duration (/s/, /z/)
  - Duration of each phoneme depends on rate of speech, intrinsic duration of that phoneme, identities of surrounding phonemes, syllabic stress, word emphasis, position in word, position in phrase, etc.



Models of Human Speech Recognition

- **The Motor Theory (Liberman et al.)**
  - Speech is perceived in terms of intended physical gestures
  - Special module in brain required to understand speech
  - Decoding module *may* work using "Analysis by Synthesis"
  - Decoding is "inherently complex"
- **Criticisms of the Motor Theory**
  - People able to read spectrograms
  - Complex non-speech sounds can also be recognized
  - Acoustically-similar sounds may have different gestures

Models of Human Speech Recognition

- **The Multiple-Cue Model (Cole and Scott)**
  - Speech is perceived in terms of
    a) context-independent invariant cues &
    b) context-dependent phonetic transition cues
  - Invariant cues sufficient for some phonemes (/s/, /ch/, etc)
  - Other phonemes require context-dependent cues
  - Computationally more practical than Motor Theory
- **Criticism of the Multiple-Cue Model**
  - Reliable extraction of cues not always possible

Models of Human Speech Recognition

- **The Fletcher-Allen Model**
  - Frequency bands processed independently
  - Classification results from each band "fused" to classify phonemes
  - Phonetic classification results used to classify syllables, syllable results used to classify words
  - Little feedback from higher levels to lower levels
  - $p(CVC) = p(c_1)\, p(V)\, p(c_2)$; implies phonemes perceived individually
- **Criticism of the Fletcher-Allen Model**
  - How to do frequency-band recognition? How to fuse results?

Models of Human Speech Recognition

- **Summary:**
  - Motor Theory has many criticisms; is inherently difficult to implement.
  - Multiple-Cue model requires accurate feature extraction.
  - Fletcher-Allen model provides good high-level description, but little detail for actual implementation.

⇒ No model provides both a good fit to all data AND a well-defined method of implementation.

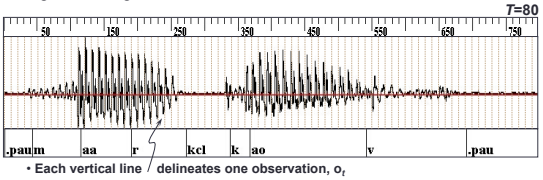
Why is speech recognition difficult?

- Nobody has sufficient understanding of human speech recognition to either build a working model or even know how to effectively integrate all relevant information.
- Lack of knowledge of human processing leads to the use of "whatever works" and data-driven approaches
- Current solution:
  - Data-driven training of phoneme-specific models
  - Simultaneously solve for duration and phoneme identity
  - Models are connected according to vocabulary constraints

  ⇒ Hidden Markov Model framework
- No relationship between theories of human speech processing (Motor Theory, Cue-Based, Fletcher-Allen) and HMMs.
- No proof that HMMs are the "best" solution to automatic speech recognition problem, but HMMs provide best performance so far

HMMs for Speech

- Speech is the output of an HMM; problem is to find most likely model for a given speech observation sequence.
- Speech is divided into sequence of 10-msec frames, one frame per state transition (faster processing). Assume speech can be recognized using 10-msec chunks.



- Each vertical line delineates one observation, o_t

HMMs for Speech

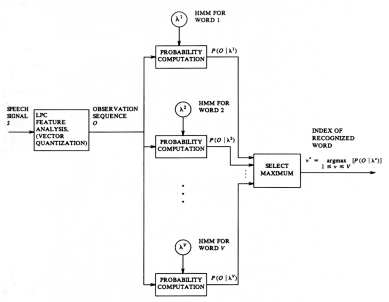


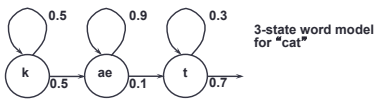
Figure 6.13 Block diagram of an isolated word HMM recognizer (after Rabiner [38]).

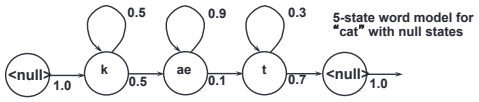
HMMs for Speech

- Each state can be associated with
  - sub-phoneme
  - phoneme
  - sub-word
- Usually, sub-phonemes or sub-words are used, to account for spectral dynamics (coarticulation).
- One HMM corresponds to one phoneme or word
- For each HMM, determine the probability of the best state sequence that results in the observed speech.
- Choose HMM with best match (probability) to observed speech.
- Given most likely HMM and state sequence, maybe determine the corresponding phoneme and word sequence.

HMMs for Speech

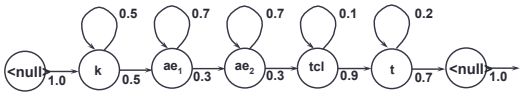
- Example of states for word model:





HMMs for Speech

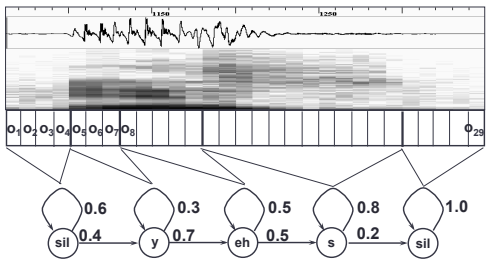
- Example of states for word model:



- 7-state word model for "cat" with null states
- Null states do not emit observations, and are entered and exited at the same time *t*.
  - Theoretically, they are unnecessary.
  - Practically, they can make implementation easier.
- States *don't have to* correspond directly to phonemes, but are commonly labeled using phonemes.

HMMs for Speech

- Example of using HMM for word "yes" on an utterance:

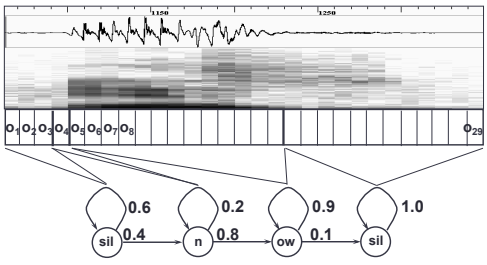


$b_{sil}(o_1)\cdot 0.6\cdot b_{sil}(o_2)\cdot 0.6\cdot b_{sil}(o_3)\cdot 0.6\cdot b_{sil}(o_4)\cdot 0.4\cdot b_{y}(o_5)\cdot 0.3\cdot b_{y}(o_6)\cdot 0.3\cdot b_{y}(o_7)\cdot 0.7$ ...

observation — state

HMMs for Speech

- Example of using HMM for word "no" on same utterance:



$b_{sil}(o_1)\cdot 0.6\cdot b_{sil}(o_2)\cdot 0.6\cdot b_{sil}(o_3)\cdot 0.4\cdot b_{n}(o_4)\cdot 0.8\cdot b_{ow}(o_5)\cdot 0.9\cdot b_{ow}(o_6)\cdot 0.9$ ...

HMMs for Speech

- Because of coarticulation, states are sometimes made dependent on preceding and/or following phonemes (context dependent).
  - **ae** (monophone model)
  - k-**ae**+t (triphone model)
  - k-**ae** (diphone model)
  - **ae**+t (diphone model)
- Constructing words requires matching the contexts:
  - "cat":

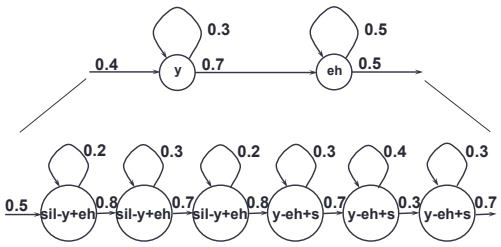
HMMs for Speech

- This permits several different models for each phoneme, depending on surrounding phonemes (context sensitive)
  - k-**ae**+t
  - p-**ae**+t
  - k-**ae**+p
- Probability of "illegal" state sequence is zero (never used)

- Much larger number of states to train on… (50 vs. 125,000 for a full set of phonemes, 39 vs. 59,319 for reduced set).

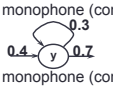
HMMs for Speech

- **Example of 3-state, triphone HMM (expand from previous):**



HMMs for Speech

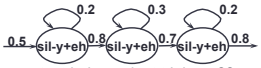
- 1-state monophone (context independent)



- 3-state monophone (context independent)

- 1-state triphone (context dependent)

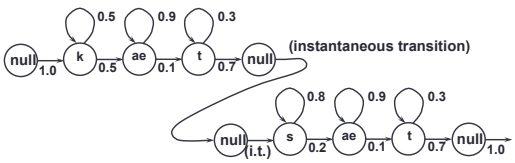
- 3-state triphone (context dependent)



- what about a context-independent triphone??

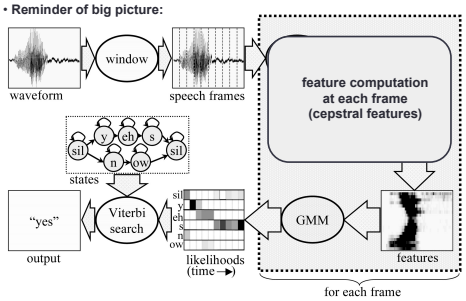
HMMs for Speech

- Typically, one HMM = one word or phoneme
- Join HMMs to form sequence of phonemes = word-level HMM
- Join words to form sentences = sentence-level HMM
- Use <null> states at ends of HMM to simplify implementation



HMMs for Speech

- **Reminder of big picture:**



(from Encyclopedia of Information Systems, 2002)

Agenda

- HW5 grades
- HW7 "decision" due today
- HW6 due next Tuesday
- Course evals
- Online NLP course @ Stanford
- Questions, comments, concerns?
- Speech Recognition (ASR)
- Text-to-Speech (TTS)
  - Next time