Second Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2020
Johns Hopkins University
Co-ordinator: Philipp Koehn

10 November 2020

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Auditory System and Speech Basics *30 points*

1. What are the approximate estimates of bit rates of speech signal and of linguistic message in speech and how would you estimate it ?

2. Why do human speakers generate different speech sounds?

3. What is an equal loudness curve and what does it tell us about how the brain perceives sounds?

4. Explain how the brain converts sound waves into neural signals?

5. You were asked to develop a technology for a deaf person where you need to bypass damaged parts of their ear. You are focusing mostly on speech-based communication (less so on music and other sounds). Knowing what you know about both speech signals and the auditory system, and assuming no biological or computational/engineering constraints, what can you do to deliver a speech signal to help this person? Which parts of the signal are critical?

Q2. Spectrograms *10 points*

Recall that the spectrogram is a visual representation of $X[k, \omega]$ of a speech signal $x[n]$ in time-window $n \in [k - \Delta, k + \Delta]$ centered at k, with k represented along the x-axis of the spectrogram, ω along the y-axis, and the magnitude of the Fourier transform $|X[k, \omega]|$ plotted as the pixel intensity, darker for higher values.

1. What window sizes (in milliseconds) are associated with a *narrow-band* versus a *wide-band* spectrogram? What is commonly used for automatic speech recognition?

2. What is the typical range of frequencies (in Hz or kHz) on the y-axis in which speech signals have spectral energies? How does it depend on the sampling rate of $x[n]$?

3. What is a *formant*, and how it is visually manifested in a spectrogram? What kinds of phonemes typically generate formants?

4. How does one identify fricatives sounds (e.g. an /sh/ or /zh/) in a spectrogram?

Q3. Mel-Frequency Cepstral Coefficients *10 points*

Recall that a typical signal representation used for automatic speech recognition is a sequence of Mel-Frequency Cepstral Coefficients or MFCCs. In what order are the following signal processing operations carried out to compute MFCCs?

- ☐ Discrete cosine transform
- ☐ Discrete Fourier transform
- ☐ Logarithm computation
- ☐ Magnitude computation
- ☐ Mel-weighting (triangular filters of increasing width)
- ☐ Preemphasis
- ☐ Sampling the continuous-time signal
- ☐ Windowing

State the psycho-acoustic motivation for

- ignoring the phase of the Fourier transform,
- taking the logarithm, and
- the increasing width of the Mel-filters.

Q4. Hidden Markov Models *10 points*

A major reason for the popularity of hidden Markov models (HMMs) for acoustic modeling is their compositional nature: HMMs for phonemes can be strung together to create HMMs for words, which can be strung together to create HMMs for sentences, etc.

Using $\mathbf{x}$ to denote a sequence of observations $x_1, \ldots, x_T$ from an HMM, $\mathbf{s}$ the sequence $s_1, \ldots, s_T$ of unobserved states, and s_0 the (known) initial state, one may write

$$P(\mathbf{x}, \mathbf{s} \mid s_0) \;=\; \prod_{n=1}^{T} P(x_n|s_n)\, P(s_n|s_{n-1}).$$

State, both in words and in a mathematical precise expression, the computational problem associated with HMMs that is solved by

1. the forward-backward algorithm,

2. the Baum-Welch algorithm, and

3. the Viterbi algorithm.

- For each algorithm above, specify its computational (big-O) complexity in terms of the input length T, and the number of states S in the HMM.

- How are E and S related for a general HMM versus a left-to-right HMM? Refine your answer above in terms of the number of permitted transitions (edges) E in the HMM instead of the number of states S.

- What is the typical value of S in a large vocabulary speech recognition system?

  1. Ten to a hundred
  2. Hundred to a thousand
  3. Thousand to ten thousand
  4. Ten thousand to a million
  5. More than a million

  What is the typical value of E?

Q5. Speaker Recognition *20 points*

What is the role of linear discriminant analysis or probabilistic linear discriminant analysis applied to speech representation for speaker recognition?

Q6. Neural Speech Recognition *20 points*

Please discuss the difference between HMM-based, connectionist temporal classification (CTC), and attention-based speech recognition systems in terms of 1) conditional independence assumptions, 2) use of language and lexicon/pronunciation models, 3) implementation, 4) speech recognition performance. You can also include the other perspectives if you want (e.g., training cost and number of parameters, as discussed in our class).

- conditional independence assumptions (including Markov assumptions)

- use of language and lexicon/pronunciation models

- implementation

- speech recognition performance

- others? (optional)