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 ?

   Audio: 50kb/s, message less than 50 bp/s. Important point is that student knows the signal has orders of magniturw higher bit rates than the actual message in speech. Plus if they know how these rates are computed.

2. Why do human speakers generate different speech sounds?

   To be able to create practically infinite number of speech messages by combining speech sounds (phones) to words to sentences and sentences to message.

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

   An equal loudness curve measures the sound pressure level (in decibels) for different frequencies to sound equally loud. The curves show that human perception of sound level is nonlinear (i.e. even if two frequencies have the same physical intensity, they are perceived as having different loudness levels).

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

   Any level of details along these steps are acceptable.
   1. Sound waves from the environment are gathered by the outer ear and sent down the ear canal to the eardrum.
   2. The sound waves cause the eardrum to vibrate, which sets three tiny bones in the middle ear into motion.
   3. The motion of the bones causes the fluid in the inner ear (cochlea) to move.
   4. The movement of the inner ear fluid causes hair cells in the cochlea to bend. The hair cells change the movement into electrical pluses.
   5. These electrical impulses are transmitted to the rest of the brain, where they are interpreted as sound.

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?

   A speech processor (with a microphone) can collect sounds, map signals into frequency channels, and stimulate different parts of the cochlea with the spectral content at each frequency band. Since the envelope of speech signals are more informative than the fine structure (in terms of speech intelligibility), it is more important to stimulate the cochlea at a lower sampling rate following the temporal envelope at different frequency bands.

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 points)

   Wide-band: $<$ 10 ms, narrow-band: $>$ 20 ms. Narrow-band used for ASR.

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 $x[n]$? (2 points)

   Between 100Hz (low) to 6-8 kHz (high). Half of the sampling rate is the Nyquist rate, which is the highest observed frequency.

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

   Dark bands on the spectrogram; resonances of the vocal tract. Voiced phonemes and vowels generate formants.

4. How does one identify fricatives sounds (e.g. an /sh/ or /zh/) in a spectrogram? (3 points)

   Energy above 2 or 3 kHz; center of gravity in vertical slice is in the upper half.

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? (4 points)

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

sampling, preemphasis, windowing, fourier, magnitude, mel-weighting, logarithm, DCT

State the psycho-acoustic motivation for (3 x 2 points)

- ignoring the phase of the Fourier transform,
  human ear is not sensitive to phase
- taking the logarithm, and
  human perception is non-linear
- the increasing width of the Mel-filters.
  human ability of frequency discrimination reduces with increasing frequency

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,

   Inference on the HMM

   $$P(A|W) = \sum_S \prod_n p(x_n|s_n)p(s_n|s_{n-1})$$

2. the Baum-Welch algorithm, and

   Updating HMM parameters

   $$p(x_n|s_n) \;\; \textit{and} \;\; p(s_n|s_{n-1})$$

3. the Viterbi algorithm.

   Most likely state sequence

   $$argmax_W P_A(A|W) = argmax_S P_A(S|A)$$

- 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.

  $O(ET) = O(S^2T)$ for a general HMM. $E = S$ for HMM used in ASR.

- 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.

  See above answer.

- 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?

  S is Thousand to ten thousand. E is of the same order.

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?

The role of LDA or PLDA is to project the speech representation in new space that maximize the discrimination between speaker and minimize within speaker variability such as channel, emotion states, languages...

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)

  CTC and HMM-based systems are based on conditional independence assumptions. Attention-based system is not explicitly based on conditional Independence assumptions.

  If there are such discussions, people can get 4 points.

- use of language and lexicon/pronunciation models

  Language model: *Attention-based system holds the language model like structure inside the decoder network and it would not require language models. CTC and HMM based systems require external language models, although it is not necessary for CTC.*

  If there are such discussions, people can get 2 points.

  lexicon/pronunciation model: *Attention-based and CTC systems do not require the lexicon/pronunciation model while the HMM-based system requires it.*

  If there are such discussions, people can get 2 points.

- implementation

  The HMM-based system requires a lot of implementation for each module. CTC is easy. We can just use a CTC objective function. Attention-based encoder-decoder is more difficult than CTC, but it is still based on a single neural network and it does not require the implementation of multiple modules, compared with the HMM-based system.

  If there are such discussions, people can get 4 points.

- speech recognition performance

  Basically all of them are comparable, but if the amount of training data is small, the HMM-based system is still powerful due to the modular optimization characteristics including the use of pronunciation/lexicon models.

If there are such discussions, people can get 4 points.

- others? (optional)

  If there are some discussions and they are reasonable, the student could get 2-4 points. Please let me know if you have some issues.