Second Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2021
Johns Hopkins University
Co-ordinator: Philipp Koehn

9 November 2021

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: ______________________

Q1. Auditory System *15 points*

1. What is the role of the larynx in speech production?

2. How is a sound frequency represented in the auditory system?

3. What is auditory masking? How is it relevant to audio compression (e.g. MP3 files)?

Q2. Speech Basics *15 points*

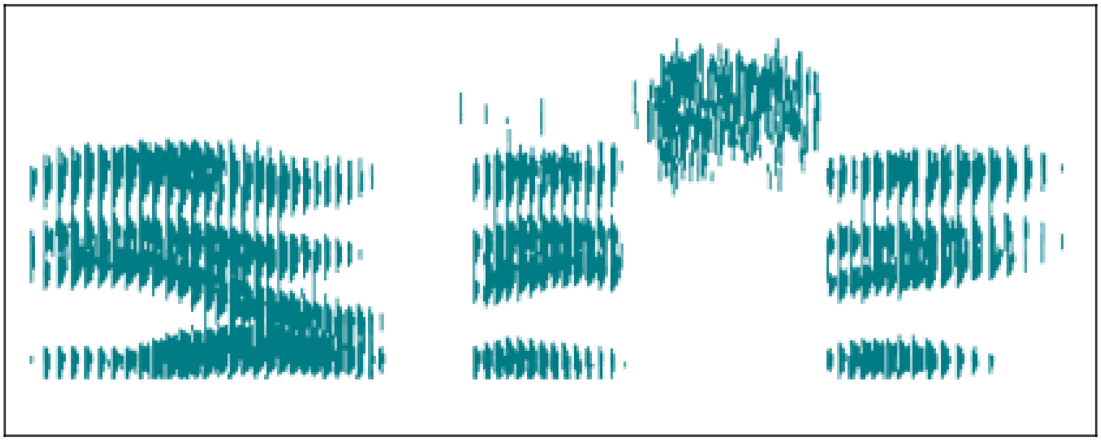
1. Speech and hearing are closely related. Could you speculate how they influence each other?
   How would you support your arguments?

2. According to source-filter theory of speech production, speech is represented by a sequence of short-time spectra, produced by filtering short segments of signals from the voice source by filter representing filtering function of stationary vocal tract. What is the main carrier of message in speech, the voice source or the vocal tract shape?

3. Do you know about any problems with short-time spectra of speech as carriers of linguistic messages ? Is there any alternative view of speech?

Q3. Spectrograms *15 points*

The logo of the Linguistic Data Consortium (LDC) is a spectrogram that represents the voice of its founding director, the late Jack Godfrey, saying the letters L-D-C.



Assuming canonical pronunciations in terms of `phonemes` as given below,

L → `É L` D → `D Ì` C → `S Ì`

answer the following w.r.t. this spectrogram. Read them all first, before answering.

1. Circle the region corresponding to the two instances of the vowel `Ì`.

2. Circle the location of the consonant `S`. Explain your placement in one sentence.

3. Circle the location of the consonant `D`. Explain your placement in one sentence.

4. What are the dark bands called? What do they represent?

5. Why is it harder to mark the boundary between the `É` and the `L`?

Q4. Hidden Markov Models in ASR *15 points*

Imagine that you are going to build and use a hybrid automatic speech recognition (ASR) system, whose acoustic model $P(\boldsymbol{A} \,|\, \boldsymbol{W})$ utilizes a different 3-state left-to-right hidden Markov model (HMM) to represent each phoneme, and whose language model $P(\boldsymbol{W})$ is a bigram. Revisit the LDC utterance from the previous problem in this setting.

1. How many *distinct* HMMs are needed to construct $P(\boldsymbol{A} \,|\, \boldsymbol{W})$ for $\boldsymbol{W} =$ "L D C"?

2. Draw the composite HMM for $P(\boldsymbol{A} \,|\, \text{"L D C"})$. How many states does it have?

3. If this were a *training* utterance for your ASR system, which algorithm would you use to estimate the HMM parameters? What is its computational complexity in terms of the length T of the utterance and the number of states S of the HMM?

4. If you want to automatically find the time-alignment of each of the phonemes, as you did *visually* in the previous problem, which algorithm would you use? What is its computational complexity?

5. For the purpose of language modeling, would you represent LDC as one word, or a sequence of three “words”? Why? What about other abbreviations and acronyms like CDC and LDA? Or like NATO and UNESCO? Or . . .

Q5. Speaker Recognition *20 points*

1. Why do you need a mixture of Gaussians to model the ceptral features instead of a single Gaussian?

2. What is the role of the Universal Background Model (UBM) to train the Gaussian Mixture Model Target model?

3. Why it is better to use Maximum A Posteriori (MAP) estimation to train the GMM Target model instead of Maximum likelihood?

Q6. Neural Networks in ASR *20 points*

1. Give an example of one of the earliest uses of neural networks for acoustic modeling in speech recognition: cite a paper (authors, year) and describe the network's inputs, outputs and approximate topology.

2. Give an example of one of the earliest uses of neural networks for language modeling in speech recognition: cite a paper (authors, year) and describe the network's inputs, outputs and approximate topology.

3. Write down the mathematical expressions for the HMM and CTC training objectives, $\mathcal{L}_{\text{HMM}}(\boldsymbol{\vartheta})$ and $\mathcal{L}_{\text{CTC}}(\boldsymbol{\theta})$, discussed in class.

4. What conditional independence assumptions are made by HMMs vs CTC? Hint: see which joint probability is written as a product of probabilities in each expression.

5. Examine the slide (from class) showing the *Encoder-Decoder with Attention*. The slide illustrates the computation of attention weights for one step of decoding; two steps if animated. What is the computational complexity of the attention layer in terms of its input length n and output length k?

6. In ASR, the length of the output (transcript) naturally/typically grows linearly with the length of the input (speech). What does this say about the effective computational complexity of an attention layer in terms of input length n?