Second Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2024
Johns Hopkins University
Co-ordinator: Philipp Koehn

7 November 2024

Complete all questions.

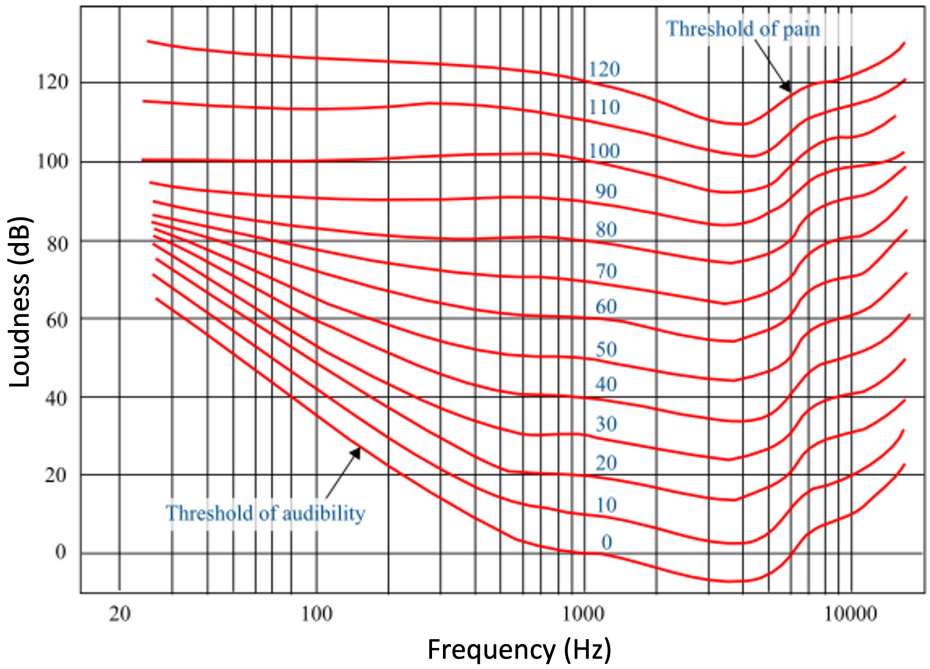
Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Auditory System *15 points*

1. You are provided with a chart of equal loudness curves.



a. Compare the minimum audible sound pressure level at 1KHz and 10KHz. Which frequency requires a higher dB SPL to be just audible, and by how much? (5 points)

b. Explain how this information relates to perceptual audio coding, like MP3 compression (5 points)

2. List the basic states of the vocal cords and describe their role in the generation of speech sounds. (5 points)

Q2. Speech Basics *15 points*

1. Express in your own words what would happen if human speech had very low entropy or very high entropy. Justify your response. (5pt)

2. Respond to the following questions (you might need to use the figure below) (4pt):
a) Name the three main articulators involved in the production of vowels?

b) Vowels are sonorant sounds. Nasals are consonants, but are they sonorant? Justify your response.

c) According to the IPA table, can Flaps be bilabial? If affirmative, please provide two examples.

d) According to the IPA table, can plosives be velar? If affirmative, please, provide some examples of plosive velars.

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

3. What are the differences between graphemes and phonemes (3pt)

4. Describe the differences between articulation place and manner classes. Describe at least two examples of each (3pt)

Q3. Classic Speech Recognition *7 points*

1. Speech recognition is typically formulated as identifying the most likely word sequence $\hat{\mathbf{W}}$ given an acoustic signal $\mathbf{A}$, i.e.:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{A})$$

This can be difficult to estimate, and is reformulated using Bayes' theorem:

$$= \arg\max_{\mathbf{W}} \frac{P(\mathbf{A}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{A})}$$
$$= \arg\max_{\mathbf{W}} P(\mathbf{A}|\mathbf{W})P(\mathbf{W})$$

What are $P(\mathbf{A}|\mathbf{W})$ and $P(\mathbf{W})$ referred to, and what do they model? (4 points)

2. Hidden Markov Models are state machines with both probabilistic transitions and outputs. Individual phonemes have been modeled with just a few HMM states, but can be strung together into larger models of words and then word sequences.

   a. What is the main benefit to modeling phonemes rather than words or word sequences directly? (1 point)

   b. What is the main challenge resulting from breaking down word sequences into hundreds of sub-phonetic HMM states? (1 point)

   c. Each phoneme is typically modeled with three states in a simple left-to-right topology with additional self-loop transitions. Give one reason this model is appropriate for phonemes. (1 point)

Q4. Speaker Recognition *20 points*

1. Draw the scheme of a generic speaker embedding network and explain the roles of the different parts of the network. Name different architecture choices for each one of the parts. (8 points)

2. Define the probability of false alarm and probability of false rejection and how to calculate them (4 points)

3. Explain how to plot the Detection Error Trade-off curve and how to calculate the EER (4 points)

4. Suppose you have an SV system that produces well-calibrated log-likelihood ratios, and you want to use it on an application where the prior probability of observing a target trial (true user) is P_T. Derive the formula for the decision threshold that we need to apply to the log-likelihood ratio. (4 points)

Q5. Neural Networks in ASR *15 points*

We discussed modifications of the Encoder-Decoder architectures used in Machine Translation and how they can be used for speech recognition. The first part of this question pertains to those modifications. In the next part of the question we will discuss encoder only alternatives to encoder-decoder models.

1. How does audio input cause problems for Encoder-Decoder models. Mention the computational complexity of self-attention. (4 points)

2. What operation is normally performed on speech inputs at the first few layers of the encoder to address this issue? (2 points)

3. Encoder only models are often used instead of encoder-decoder models for ASR. Name the commonly used objective function, discussed in class, used for sequence-level training of encoders in ASR and describe how it accounts for multiple possible valid alignments of the output sequence to the input speech. (3 points)

4. We discussed a particular way of aligning outputs to inputs in class by repeating output symbols and using a special blank symbol, $\oslash$. It is also the mechanism used in the sequence-level objective function described in the previous question. For an input speech utterance of length five, and using letters plus the blank symbol, $\oslash$, as the output units, enumerate all possible alignments of the word `all`, to the input speech. (6 points).

Q6. Enhancement and Diarization *14 points*

1. What is speech enhancement, and what are three motivations for developing speech enhancement technologies? (4 points)

2. "Synthetic" data is often used in speech enhancement research. (4 points)

    a. What is synthetic data in this context?

    b. What is the purpose of using it?

    c. What are some implications for the development, testing, and deployment of systems developed using synthetic data?

3. The "permutation" problem arises in both enhancement and diarization. (4 points)

   a. What is the permutation problem?

   b. Where does the problem arise in enhancement/separation?

   c. Where does the problem arise in diarization?

   d. Give an example of an approach used to address this problem.

4. Recording audio on an array of microphones (i.e. producing a multi-channel input signal and potentially performing beamforming) generally improves diarization performance. Give two examples of how this helps. (2 points)

Extra Space