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.

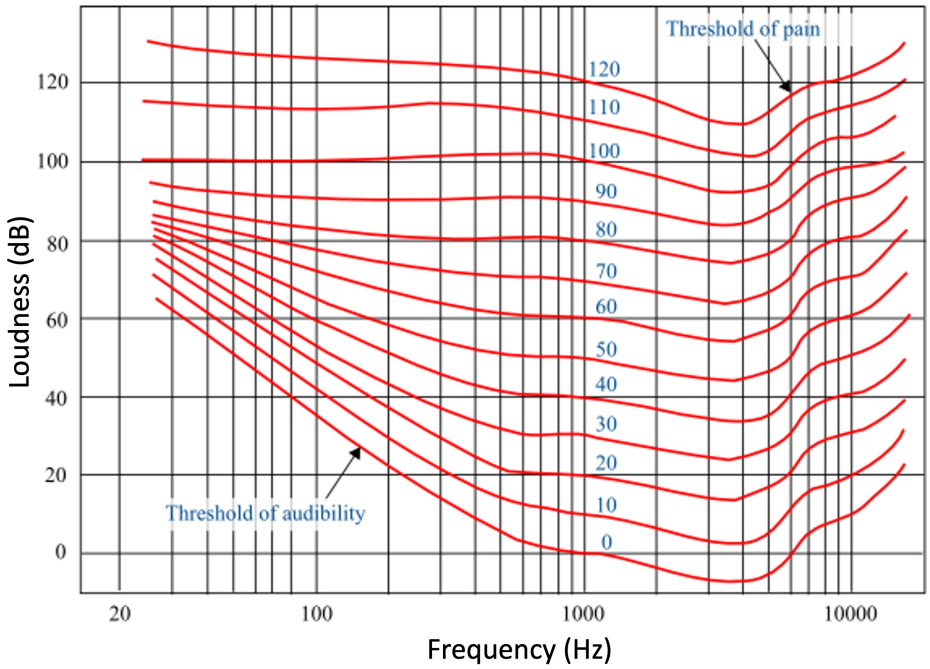
Use additional paper if needed.

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)

To find the minimum audible sound pressure levels according to the threshold of hearing (the lowest curve on the chart):

- *Locate the points on the lowest curve where it intersects 1 kHz and 10 kHz.*
- *At 2 kHz, the threshold of hearing is approximately 0 dB SPL, indicating that this is near the most sensitive range of human hearing.*
- *At 10 kHz, the threshold of hearing is typically around 10 dB SPL. Thus, a 10 kHz tone needs to be about 10 dB louder than a 1 kHz tone to be just audible.*

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

This difference in sensitivity is leveraged in MP3 and other perceptual coding formats. Since the human ear is more sensitive around 2 kHz, audio compression algorithms prioritize accuracy in this frequency range while reducing precision at frequencies like 10 kHz, where the ear is less sensitive. MP3 coding removes inaudible or less noticeable

sounds hence achieving a reduction of file size (compression) without a perceptible loss in quality.

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

   The vocal folds, also known as vocal cords, play a crucial role in speech production by controlling the airflow and sound generation in the larynx. They take on 3 general states:

   (a) *Breathing: where the fold muscles are relaxed, the glottis is wide open and air flows from and to lungs with no hindrance*

   (b) *Voiced: where the folds open and close in a quasi-periodic pattern that represents the pitch of the speaker. The folds are voiced during production of vowels and some voiced non-vowels.*

   (c) *Unvoiced: a similar state to breathing where there are no vocal fold vibrations but folds are closer together and more tense than in breathing state causing a hindrance to airflow. This is typical in production of a whisper sound like /h/*

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)
   Given that we do not all pronounce sounds in the same way and that there can be channel influences in the speech signals (background noise, excessive reverberation, etc), the fact that the speech signal is to some extent predictable due to grammar, phonotactics, language rules, context, allows us to "fill the gap" of sounds or even words we do not perceive well. But if the signal were 100% predictable all the time, that would mean that we do not need to speak as there is no new information to convey. Therefore, we need a balance between predictability and unpredictability.

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?
   Tongue, jaw, lips.
   b) Vowels are sonorant sounds. Nasals are consonants, but are they sonorant? Justify your response.
   Nasals are sonorant. Vocal folds are involved in producing nasals.
   c) According to the IPA table, can Flaps be bilabial? If affirmative, please provide two examples.
   No, they cannot
   d) According to the IPA table, can plosives be velar? If affirmative, please, provide some examples of plosive velars.
   Yes, they can. k and g.

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)
   A phoneme is a perceptually distinct speech element that could distinguish one word from another. A grapheme is a letter or combination of several letters that represent a sound (phoneme).

4. Describe the differences between articulation place and manner classes. Describe at least two examples of each (3pt)
   Manner classes are ways in which we articulate that produce consonant sounds. For instance, plosives are manner classes. In plosives we close completely the vocal tract and release the air stream to generate the sounds. Fricatives are also manner classes. In fricatives, we constrict the vocal tract partially, which creates turbulences that generate the fricative sounds. In contrast, articulation places are points or areas in the vocal tract where there is a constriction (with or without contact) which has the most relevance in the generated sound. Articulation points can be alveolar, glottal, labiodental, etc. In alveolar points, the tongue generates a constriction (total or partial) in the front of the hard palate. In bilabial articulation points, the lips generate the constriction.

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)

- *(1pt) $P(\mathbf{A}|\mathbf{W})$ is the "acoustic model"*
- *(1pt) The acoustic model models the the space of sounds that can manifest from a particular word seqeunce*
- *(1pt) $P(\mathbf{W})$ is the "language model"*
- *(1pt) The language model models how likely any given word sequence is, regardless of the observed audio*

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)
   (1pt) We have very few training examples of any particular word sequence, but many examples of all phonemes, which can be shared between word sequences.

   b. What is the main challenge resulting from breaking down word sequences into hundreds of sub-phonetic HMM states? (1 point)
   (1pt) There are heavy computational costs to evaluating all possible state sequences, requiring specialized algorithms.

   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)
   (1pt) Either of the following:
   - *The onset, stationary middle, and decay can be modeled separately*
   - *Self-loops aid in modeling phonemes like vowels which can vary in length*

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)
   X-Vector network has three parts:

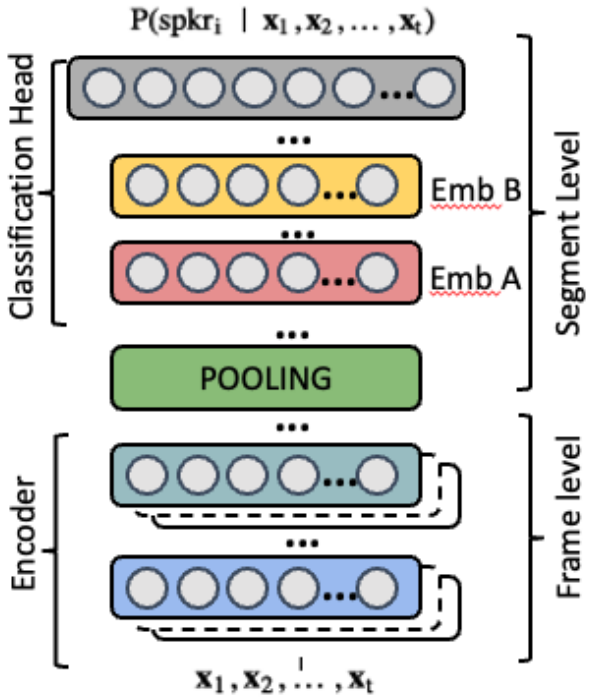


Figure 1: speaker embedding network (2 points)

- *Encoder: Gets log-Mel spectrogram features as input and produces speaker discriminant frame-level hidden representations as output. Different architectures in the literature include time-delay neural networks, 2d convolutional residual networks, and ecapa-tdnn networks (2 points).*
- *Pooling: Summarizes the frame-level hidden representations into a single vector per audio. Some methods are global mean pooling, global statistics pooling (mean+standard deviation), attentive statistics pooling, multi-head attention and channel-wise attentive statistics pooling (2 points).*
- *Classification Head: Predicts posterior probabilities for the training speakers given the pooling vector. At inference time, we extract the speaker embedding from a middle layer of the classification head (2 points).*

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

   - *P_{fa}: Is the probability for an impostor to be classified as a true target speaker. This is calculated by setting a decision threshold on the speaker verification scores, counting the number of times an impostor is accepted and dividing by the total number of impostor trials.*
   - *P_{fr}: Is the probability for true target speaker to be classified as an impostor. This is calculated by setting a decision threshold on the speaker verification scores, counting the number of times an true target is rejected and dividing by the total number of target trials.*

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

   - *DET curve plots the prob. of false alarm in the x-axis versus the prob. of false rejection or miss in the y-axis. To calculate it, we need to move the detection threshold from the minimum to the maximum score in our test scores. For each threshold value, we calculate the prob. of miss and prob. of false alarm. This gives us a point of the curve. The pairs of (P_{fa}, P_{fr}) for all thresholds give us the full curve*
   - *The EER is the point in the curve where the prob of false alarm is equal to prob. of rejection EER=$P_{fa} = P_{fr}$.*

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)

   - *To accept a target trial, the posterior probability for target trial has to be $P(T|x_e, x_t) \leq 0.5$.*
   - *Write the posterior as a function of the LLR using Bayes Theorem:*

$$P(\text{Tar}|\mathbf{x}_1, \mathbf{x}_2) = \frac{P_{\text{Tar}} P(\mathbf{x}_1, \mathbf{x}_2|\text{Tar})}{P_{\text{Tar}} P(\mathbf{x}_1, \mathbf{x}_2|\text{Tar}) + (1 - P_{\text{Tar}}) P(\mathbf{x}_1, \mathbf{x}_2|\text{NonTar})} =$$

$$\frac{1}{1 + \exp\left(-\log \frac{P_{\text{Tar}} P(\mathbf{x}_1, \mathbf{x}_2|\text{Tar})}{(1-P_{\text{Tar}}) P(\mathbf{x}_1, \mathbf{x}_2|\text{NonTar})}\right)} = \text{sigmoid}\left(\log \frac{P_{\text{Tar}} P(\mathbf{x}_1, \mathbf{x}_2|\text{Tar})}{(1 - P_{\text{Tar}}) P(\mathbf{x}_1, \mathbf{x}_2|\text{NonTar})}\right) \geq 0.5$$

   - *Isolate the LLR from the eq. above:*

$$\log \frac{P(\mathbf{x}_1, \mathbf{x}_2|\text{Tar})}{P(\mathbf{x}_1, \mathbf{x}_2|\text{NonTar})} \geq -\log \frac{P_{\text{Tar}}}{(1 - P_{\text{Tar}})}$$

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)
   Audio for ASR is normally sampled at 16 kHz. Assuming a speaking rate of about 2 words per second, and an average sentence length of 15 words (any such similar numbers will do), the average audio input could easily be about 7.5 seconds, or 120,000 input samples. Even using MFCCs, which have a lower frame rate ($\sim 100Hz$) the average input length would be 700 samples. This is significantly longer than inputs used in Machine Translation. Because self-attention has a computational complexity of $\mathcal{O}(N^2)$ in the length, N, of the input, using such long sequences is prohibitively expensive.

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

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)
   CTC. It marginalizes *over all valid alignments, i.e., it is the sum of the scores of any valid alignment.*

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).

   (a) $a\ a\ l\ \oslash\ l$

   (b) $a\ l\ \oslash\ l\ l$

   (c) $a\ l\ l\ \oslash\ l$

   (d) $a\ \oslash\ l\ \oslash\ l$

   (e) $\oslash\ a\ l\ \oslash\ l$

   (f) $a\ l\ \oslash\ l\ \oslash$

   (g) $a\ l\ \oslash\ \oslash\ l$

Q6. Enhancement and Diarization *14 points*

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

   - *(1pt) Enhancement is the removal of interfering audio signals (e.g. noise or non-target speech) from a recording of desired speech.*
   - *(1pt) Enhancement is used to clean up audio for humans to listen to.*
   - *(1pt) Enhancement is used as pre-processing for downstream speech technology.*
   - *(1pt) Techniques developed for enhancement can be integrated into end-to-end systems for other tasks.*

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

   a. What is synthetic data in this context?
   (1pt) Synthetic data means artificially mixing clean speech and noise to produce noisy speech rather than collecting natural recordings of speech with noise.

   b. What is the purpose of using it?
   (1pt) It allows us to have a ground truth clean speech signal for a noisy recording.

   c. What are some implications for the development, testing, and deployment of systems developed using synthetic data?
   (2pt) Any two of the following:
      - *It allows supervised training (i.e. training to directly produce the ground truth)*
      - *It allows full-reference evaluation metrics*
      - *It excludes in-domain training for real evaluation conditions.*

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

   a. What is the permutation problem?
   (1pt) A system (generally neural network) necessarily is going to produce its outputs in a given order, but in some tasks all permutations of output order are equally valid, and so accordingly the order should not matter for the network's training.

   b. Where does the problem arise in enhancement/separation?
   (1pt) In speech separation, there is no reason to output the speech signals in any particular order. However, it does ***not*** *generally exist in enhancement, as the speech/noise distinction defines the outputs.*

   c. Where does the problem arise in diarization?
   (1pt) Diarization does not necessarily require true speaker labels, so the order of output speaker activities can be in any order.

   d. Give an example of an approach used to address this problem.
   (1pt) Either of the following:

   - *Permutation Invariant Training (computing the loss function on all permutations and only backpropagating the lowest loss)*
   - *Any sort of clustering-based method (e.g. Deep Clustering)*

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)
   (2pts) Any two of the following:

   - *Simply speaking, additional observations of the desired/interfering signals leads to more information to use.*
   - *Beamforming suppresses interfering signals (i.e. is a speech enhancement method) and can be used as a pre-processing step for diarization systems.*
   - *Multichannel audio collected by an array contains localizing information, and different people generally are observed in different locations.*

Extra Space