Second Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2025
Johns Hopkins University
Co-ordinator: Philipp Koehn

6 November 2025

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Auditory System *15 points*

1. Define formants and explain their importance in vowel production. *(5 points)* *Formants are the resonant frequency bands of the vocal tract that amplify certain harmonics of the voice source. They appear as peaks in the spectrum of the vowel. The first formants (F1, F2) are key to distinguishing different vowels (e.g., /i/, /a/, /u/). They depend on the shape and length of the vocal tract during articulation, particularly the position of the tongue*

2. Describe how the outer, middle, and inner ear contribute to the process of hearing. *(5 points)* *Outer ear (pinna and ear canal): collects sound waves and directs them to the eardrum.*
   Middle ear (ossicles: malleus, incus, stapes): amplifies and transmits vibrations from the eardrum to the oval window of the cochlea.
   Inner ear (cochlea): converts mechanical vibrations into neural signals via movement of hair cells that trigger auditory nerve activity.

3. What physical property of a sound wave determines its perceived loudness? How is loudness measured? *(5 points)*
   Loudness is primarily determined by amplitude, which reflects the sound wave's energy or intensity (which is an estimate of rate of energy flow over are in units of W/m^2). Loudness is a relative measure of sound intensity against the reference intensity I_0 which captures the human threshold of hearing of a 1KHz tone ($I_0 = 10^{-12}(W/m^2)$). It is measured in decibels (dB). Larger amplitudes are perceived as louder sounds.

Q2. Speech Basics *15 points*

1. What is the difference between a phoneme and a phone? *(5 points)*
   A phoneme is a distinct speech element that could distinguish one word from another. For instance, between "cat" and "bat", the first phoneme changes. A phone is a distinct sound produced while speaking, a segment. It can correspond to a phoneme (most of the time it does) or a smaller part of a phoneme. For instance, to generate the sound /sh/ two people might make slightly different sounds. Those sounds will be phones. However, if due to their accent a person combines two distinct sounds to produce the /sh/ phoneme, that will count as two phones.

2. Describe what phonotactics are. Include one example of a possible sequence of phonemes (represented by graphemes) and an impossible sequence of phonemes according to English phonotactics. *(5 points)* *Phonotacics study the rules governing the possible phoneme sequences in a language. A possible English sequence of phonemes is "valid". An impossible one would be "gkkkrssbmmmnskka"*

3. According to the IPA table below, indicate which of the following English words include alveolar nasals AND labiodental fricatives:

   (a) Finally
   (b) Clearly
   (c) Best
   (d) Failing
   (e) Norfolk
   (f) Socket
   (g) Blurry

   (5 points) *Finally, Failing and Norfolk contain alveolar nasals (n) and labiodental fricatives (f). None of the other words do.*

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.

Q3. Speech Recognition: Fundamentals *15 points*

1. The Source-Channel Model of Speech Recognition *(10 points)*

$$\begin{aligned}
\widehat{\mathbf{W}} &= \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{A}) && (1)\\
&= \arg\max_{\mathbf{W}} \frac{P(\mathbf{A}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{A})} && (2)\\
&= \arg\max_{\mathbf{W}} P(\mathbf{A}|\mathbf{W})P(\mathbf{W}) && (3)\\
&\approx \arg\max_{\mathbf{W}} P_{\mathcal{A}}(\mathbf{A}|\mathbf{W})P_{\mathcal{L}}(\mathbf{W}) && (4)
\end{aligned}$$

   Explain, in one sentence each, why (2) is equal to (1), and why (3) is equal to (2).

   - *(2) equals (1) by Bayes' rule*
   - *argmax over W is insensitive to marginalization. [(3) = (2)] Since $P(\mathbf{A})$ does not depend on $\mathbf{W}$, it is constant under $\arg\max_{\mathbf{W}}$ and can be dropped.*

   Explain, in one sentence each, what the models $P_{\mathcal{A}}(\mathbf{A}|\mathbf{W})$ and $P_{\mathcal{L}}(\mathbf{W})$ represent and what data is needed to train them.

   - *Acoustic model, trained using large quantities of transcribed speech;*
   - *language model, trained using a large text corpus.*

2. What factors make the automatic transcription of a group meeting more challenging than a news broadcast? *(5 points)*

   Group meetings are usually spontaneous and conversational between everyday people, recorded in normal rooms, and recorded using distant microphones, while news broadcasts feature professional speakers who often read from a teleprompter in quite studios while wearing a lapel or head-mounted microphone. Finally, text documents are available in the same style, and sometimes even on the same topics, entities and events being talked about in a news broadcast.
   Spontaneous speech has co-articulation, mispronunciations and disfluencies; reverberation and background noises degrade audio recorded with distant microphones; multi-speaker conversations often overlapped speech; speech transcripts are not as abundant as news text for language modeling.

Q4. Speech Recognition: Neural Models *20 points*

1. Draw a block diagram of an ASR system with an Encoder-Decoder-with-Attention architecture. Label the audio $\mathbf{A} = a_1, a_2, \ldots, a_T$, the transcript $\mathbf{W} = w_1, w_2, \ldots, w_N$. Draw arrows to illustrate (i) the system modules they go into or come out of, and (ii) the computational dependencies between the system modules. *(7 points)*

   Identify encoder, self- and cross-attention, and decoder blocks. Decoder has both masked

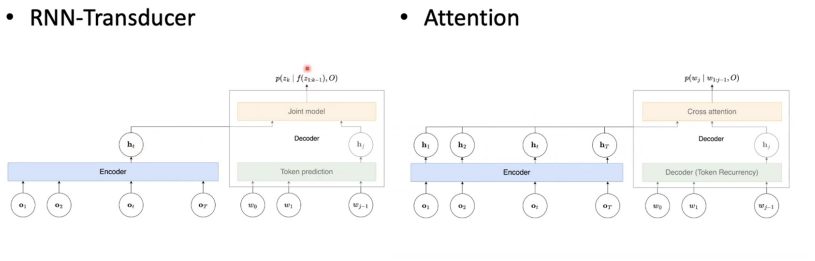


Figure 1: encoder-decoder ASR

   self-attention to its left and (masked or regular) cross-attention to the encoder outputs.

2. Draw a block diagram of an ASR system with a Recurrent Neural Network Transducer architecture. Let $\mathbf{A} = a_1, a_2, \ldots, a_T$ and $\mathbf{W} = w_1, w_2, \ldots, w_N$ as above. Draw arrows to illustrate (i) the system modules they go into or come out of, and (ii) the computational dependencies between the system modules. *(7 points)*

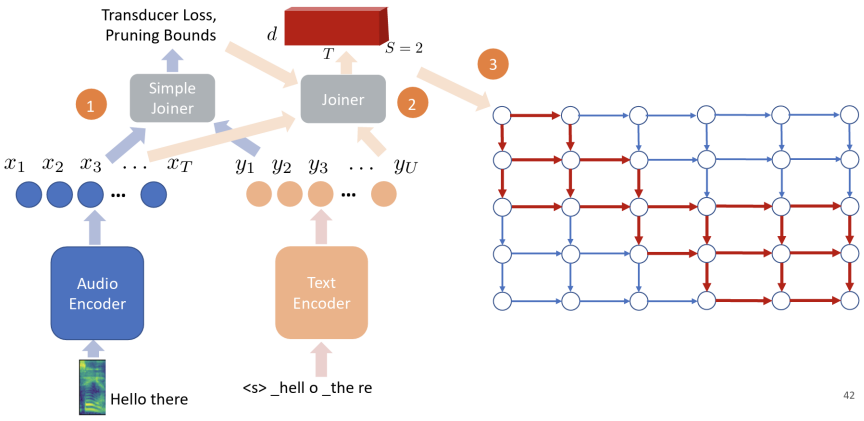


Figure 2: RNNT for ASR

   Identify the encoder, predictor and joiner blocks. Indicate (somehow) that the joiner advances through the encoder output in a time-synchronous manner, but the predictor output

advances in a symbol-synchronous manner. E.g. show that to produce an output token w_k at time frame t, the joiner joins the encoder frame t with the predictor output based on $w_1, w_2, \ldots, w_{k-1}$.

3. Which of the two architectures above requires a time-synchronous versus a label-synchronous decoder? Which architecture is typically preferred for real-time/streaming applications? Which one seems a more natural fit for incorporating an external language model (or LLM)? *(6 points) EDA is symbol synchronous, RNN-T is time-synchronous. RNN-T is preferred for real-time/streaming applications and is also "easier" for augmenting with an external LM.*

Q5. Self-Supervised Learning for Speech *15 points*

1. Draw a block-diagram to illustrate the architecture of a HuBERT model. *(3 points)* .

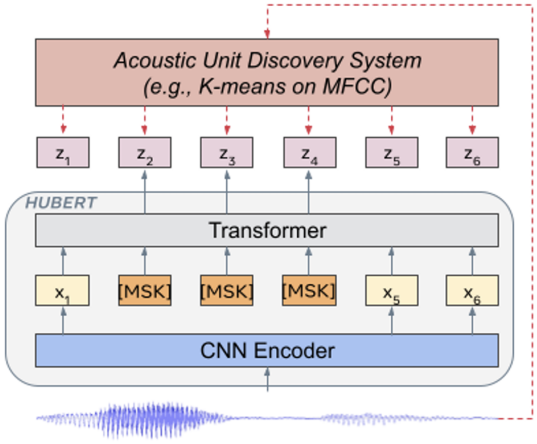


Fig. 1: The HuBERT approach predicts hidden cluster assignments of the masked frames (y_2, y_3, y_4 in the figure) generated by one or more iterations of k-means clustering.

Figure 3: HuBERT

What training objective is used to train a HuBERT model? *(2 points)*

BERT-like training with a masked language model objective

How are VQ labels generated to train a HuBERT model? *(2 points)*

Quantized MFCCs in the first iteration, quantized hidden layer in the second

2. Draw a block-diagram to illustrate the architecture of wav2vec 2.0 *(4 points)* .

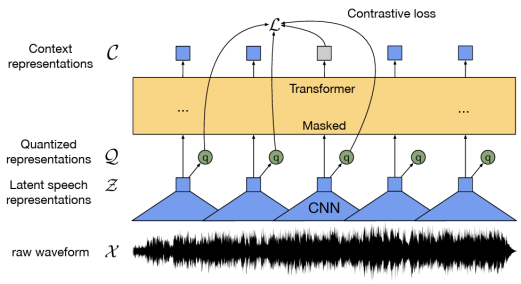


Figure 1: Illustration of our framework which jointly learns contextualized speech representations and an inventory of discretized speech units.

Figure 4: wav2vec

What training objective is used to train wav2vec 2.0? (*2 points)*

Noise contrastive estimation, plus diversity/entropy of codebook use

How are VQ centroids generated to train wav2vec 2.0? (*2 points)*

The VQ centroids (aka codebook) are learnt iteratively as part of the training.

Q6. 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:

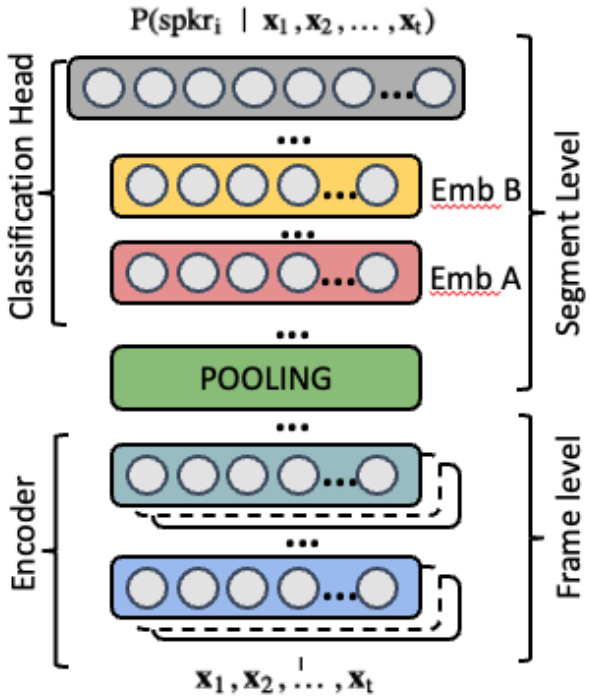


Figure 5: speaker embedding network (2 points)

- *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:*

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

Q7. Enhancement and Diarization *15 points*

1. Reverberation is a property of rooms, and its effect (given an acoustic source location and an observation point) is captured by the Room Impulse Response (RIR).

   a. What do the direct path, early reflections, and late tail components of the RIR capture? *(3 points)*

   - *(1pt) The direct path captures the delay and attenuation due to propagation from source to mic*
   - *(1pt) The early reflections capture the first few reflections off walls and surfaces*
   - *(1pt) The late tail is when the audio has reflected many times and resembles random noise*

   b. What is the effect of reverberation on the spectrogram and why does it make enhancement difficult? *(2 points)*

   - *(1pt) Reverberation smears the spectrogram*
   - *(1pt) It is difficult to discriminate between the speech and the reverberation*

   c. What is the most common approach to dereverberation and what is the idea behind the method? *(2 points)*

   - *(1pt) Weighted Prediction Error (WPE)*
   - *(1pt) Reverberation is essentially a linear function of the direct path observations and an inverse filter can be estimated*

2. What are the two primary diarization metrics, and give one example for each where that metric would be better for that conversational dynamic/application *(4 points)*

   - *The metrics are (1pt) Diarization Error Rate (DER) and (1pt) Jaccard Error Rate (JER)*
   - *(2pt) The basic difference is that DER is normalized by the amount of speech and JER is averaged over people in the recording. So, the examples should capture that JER is better if you care about individuals regardless of how much they speak, and DER is better if you care about overall performance.*

3. What is the foundational neural diarization system, and what was the "trick" required to make it work? Give one upside and one downside of this approach compared to conventional pipeline systems. *(4 points)*

   - *(1pt) The model is End-to-End Neural Diarization (EEND)*
   - *(1pt) Solving the problem that output order does not matter (i.e. the "permutation problem"). This was done using approaches from speech separation (Permutation Invariant Training (PIT) and Deep Clustering)*
   - *(1pt) Upsides include: joint optimization of the entire task and naturally handling overlapped speech*
   - *(1pt) Downsides include: requiring lots of conversational data and having problematic model memory requirements for long-term recordings*

Extra Space