Second Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2023
Johns Hopkins University
Co-ordinator: Philipp Koehn

9 November 2023

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Auditory System *15 points*

1. What are the 3 physical dimensions of sound? What does each one of these dimensions mean or represent? (5 pts)

   The 3 dimensions are:

   (a) *wavelength: It represents the distance traveled by one cycle. It is affected by the medium where sound travels (sound travels faster in non-porous solids and slower in gas (e.g. air)*

   (b) *Amplitude: It is the height of a cycle and is related to the perception of loudness*

   (c) *frequency: measured in cycles per seconds (or Hertz) and is related to the perception of pitch and spectral content of a sound.*

2. What is the McGurk effect and how does it illustrate the complexity of multimodal speech recognition? (5pts)

   (this is more text than is expected from a student answer) The McGurk effect is a perceptual phenomenon that occurs when speech sounds are combined with unmatched visual cues of a person producing a different sound, leading to a perceptual fusion of both auditory and visual information into a different sound than the original. This effect demonstrates the strong influence of visual cues, such as lip movements and facial expressions, on our perception of speech sounds. The McGurk effect highlights the multisensory nature of speech perception and how the brain integrates both auditory and visual information when processing spoken language. In practical applications, like automatic speech recognition (ASR) or human-computer interaction, understanding and accounting for the McGurk effect can lead to more accurate and context-aware speech recognition systems. However, it's essential to note that implementing effective multimodal speech recognition systems is a complex task that involves advanced technology for audio and visual processing, machine learning, and modeling of the human perception system.

3. How are sound waves converted into nerve signals? (5pts)

   Any level of details along these steps are acceptable.

   (a) *Sound waves from the environment are gathered by the outer ear and sent down the ear canal to the eardrum.*

   (b) *The sound waves cause the eardrum to vibrate, which sets three tiny bones in the middle ear into motion.*

   (c) *The motion of the bones causes the fluid in the inner ear (cochlea) to move.*

   (d) *The movement of the inner ear fluid causes hair cells in the cochlea to bend. The hair cells change the movement into electrical pluses.*

   (e) *These electrical impulses are transmitted to the rest of the brain, where they are interpreted as sound.*

Q2. Speech Basics *15 points*

1. Express in your own words why we need the speech signal to be predictable but not too predictable to communicate. (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 Figure 1) (5pt):
   a) Which is the main articulator involved in changing the frequency formants that define nasal sounds?
   The velum
   b) Vowels are sonorant sounds. Nasals are a type of sonorant consonants. But can a consonant, excluding nasal consonants, be sonorant? In other words, provide examples of consonant sounds in which the vocal folds vibrate while, not considering nasals.
   Trills can be sonorant. For instance /r/ in 'rat' is sonorant.
   c) According to the IPA table, can Fricatives be bilabial? If affirmative, please provide two examples of labiodental fricative sounds.
   Yes, they can. Examples: Φ and β.
   d) According to the IPA table, can plosives be labiodental? If affirmative, please, provide some examples of glottal nasal sounds.
   No, they cannot be glottal.

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.

Figure 1: IPA Table

3. What are the differences between phones and phonemes (2.5pt)
   A phoneme is a perceptually distinct speech element that could distinguish one word from another. A phone, is a segment that includes a distinct sound, that can correspond to a phoneme or not. A phone can be part of a phoneme or a phoneme.

4. Describe the differences between articulation place and manner classes. Describe at least two examples of each (2.5pt)
   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. Hidden Markov Models in ASR *20 points*

1. Why did the chicken cross the road? *To get to the other side.*

Q4. Speaker Recognition *16 points*

1. Name at least three applications of speaker recognition and name an example of each one of them. (5pt)
   (1.66pt/application)

   - *Law Enforcement and Forensics, e.g.:*
     - *Search for criminals in telephone conversations.*
     - *Detect known telephone fraudsters in black-lists.*
     - *Identify people making telephone threats.*
     - *Convict or discharge a defendant in court.*
   - *Identity authentication and access control, e.g.:*
     - *Access high security physical facilities.*
     - *Access computer networks and web services.*
     - *Telephonic banking.*
     - *Password reset.*
   - *Enriching meeting transcription, e.g.:*
     - *Knowing who spoke when, who said what.*
   - *Audio-Visual Media Indexing, e.g.:*
     - *Add metadata of who is speaking to Broadcast TV or youtube videos.*
     - *Improve indexing and searching documents.*
   - *Personalization, e.g.:*
     - *Voice assistants adapt to the user: play user's music, e-mails, parental control.*

2. Draw the block diagram of a speaker verification system, indicating the three different operation phases. Briefly explain what happens in each phase and the function of each block. (10pt)
   (2.5pt) for the figure.
   (2.5pt) In the training phase, we take a large database with speech of lots of people, we extract acoustic features from the audio and we train some statistical models on those features. Once we have done that; we can put the speaker verification system into production. In production, we have 2 phases: the enrollment phase and test phase.
   (2.5pt) In the enrollment phase, we register new speakers into our system. For example, we take Alice's voice, we compute acoustic features from it, and then we compute a speaker embedding for Alice. The speaker embedding is just a vector that compresses the speaker's identity information. This is typically calculated using a neural network. Finally, the speaker embedding is stored in the system's database.
   (2.5pt) Finally, in the test phase, I get a recording from somebody who says to be Alice, and

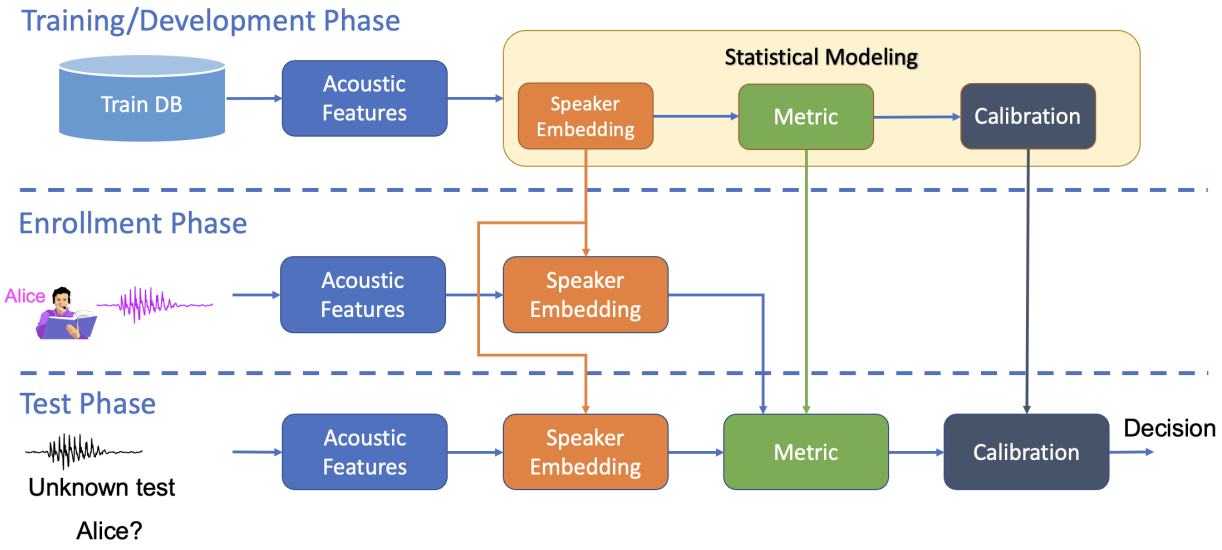

Figure 2: speaker verification pipeline

then I extract an embedding for that recording. Then, we put both Alice embedding and the test embedding into the metric block. This block compares both embeddings and produces a score. if the score is high, that means that the test speaker is Alice; if the score is low, it is not Alice. Sometimes, we have an extra block that calibrates the score into a well-calibrated probability. And finally, we put a threshold on the score to decide whether to accept or reject.

3. The speaker verification system produces a score proportional to the probability that the test audio belongs to the target speaker. We put a threshold to this score to decide if we accept or reject the trial. When doing so we can make errors. Name the types of errors that we can make, what are the error rates corresponding to them, and how do you calculate the equal error rate metric? (5pt)
(2.5pt) There are two types of errors:

   - *Misses or False rejection: when a target speaker is classified as an impostor, it is measured by the probability of miss.*
   - *False alarm: when an impostor is classified as the target speaker, it is measured by the probability of false alarm.*

   (2.5pt) Miss and false alarm probabilities depend on the decision threshold. EER obtained by finding the threshold at which the Miss rate is equal to the false alarm rate and them EER=$P_{miss} = P_{fa}$.

Q5. Neural Networks in ASR *20 points*

1. Name two of the modeling approaches, other than the transducer model, that are commonly used to build ASR systems (2 pts).

   Any of the following will work.

   - *(1 pt) Hybrid HMM/DNN*
   - *(1 pt) Connectionist Temporal Classification (CTC)*
   - *(1 pt) Encoder-Decoder with Attention*

   We will now describe a rudimentary CTC-based wake-word detection system in the following steps. A wake word detection can be thought of as a 1-phrase ASR system. When the wake-word is recognized, the device it is intended to activate, "wakes up".

2. Chose a modeling unit other than graphemes. Describe approximately how many units you expect there to be for a normal ASR system, and benefits of that choice. Describe any additional special symbols used in the CTC objective and describe their purpose(s). (3 pts)

   (1 pt for unit, number and reason, 2 pts for description of Blank symbol purposes)

   (a) *Phones / Triphones (40-10000. Depending on the language, there can be very few phones, or the annotation accuracy can be poor so the annotation essentially corresponds to phonemes. With enough data, up to around 10,000 tied triphones states might be used. They model sounds of the language.)*

   (b) *Phonemes (about 40. They model sounds of the language along with variation due to context-dependent effects that do not change the meaning of the word.)*

   (c) *Byte-Pairs (BPE) (500-10000 range is acceptable. They can model context dependent units and are derived from text, which is readily available.)*

   An additional symbol `<not-wake-word>` *for speech that corresponds to non-wake word instances could be included. In all cases a special blank symbol,* β*, must be included. It could be used instead of a dedicated symbol to model the non-wake words. In general* β *has two purposes:*

   (a) *It models silence between and inside of words.*

   (b) *It disambiguates between repeated symbols and symbols of long duration*

3. Our wake-word system will ingest sequences of up to length 6 and recognize its name, anna. Show all the possible length-6 recognized strings that correspond to the wake-word, anna. Use graphemes (the letters) as the modeling units produced for this question. Do not treat uppercase and lower case units separately. (3 pts)

   - β *a n* β *n a*
   - *a* β *n* β *n a*
   - *a n* β β *n a*
   - *a n* β *n* β *a*
   - *a n* β *n a* β
   - *a a n* β *n a*
   - *a n n* β *n a*
   - *a n* β *n n a*
   - *a n* β *n a a*

4. (a) Write the expression for the probability of the wake-work according to the CTC training objective for the sequence anna. Use $\mathbf{x}$ to denote the input speech, $\mathcal{S}(\texttt{anna})$ for the set of valid strings corresponding to the wake-word. A neural network, $f(\mathbf{x})$, is used to produce a sequence of scores over symbols (modeling units), s, for each time step, t. The Softmax function produces probabilities for these scores (4 pts).

$$p(\textit{anna}|\mathbf{x}) = \sum_{s \in \mathcal{S}(\textit{anna})} \prod_{t=0}^{5} \frac{e^{f^t_{s_t}(\mathbf{x})}}{\sum_{s'_t} e^{f^t_{s'_t}(\mathbf{x})}}$$
$$= \sum_{s \in \mathcal{S}(\textit{anna})} \prod_{t=0}^{5} p(s_t|\mathbf{x})$$

   (b) Assume the letters a and n are recognized with $p(a|\mathbf{x}) = 0.4$, $p(n|\mathbf{x}) = 0.1$, and $p(\beta|\mathbf{x}) = 1.0$. What is the CTC probability for the wake-word anna? (2 pts)

5 paths have 2 a*, 2* n*,* 2 β $\implies$ $5(0.16)(0.01) = 0.0016$
2 paths have 3 a*, 2* n*,* 1 β $\implies$ $2(0.064)(0.01) = 0.00064$
2 paths have 2 a*, 3* n*,* 1 β $\implies$ $2(0.16)(0.001) = 0.00016$
$5(.0016) + 2(0.00064) + 2(0.00016) = 0.0096$

5. What algorithm efficiently performs the calculation from question 4.b and what is its computational complexity? (2 pt)

The forward algorithm. For each step of the forward algorithm, and for each modeled unit, the incoming paths from all previous possible units must be considered. If there are k units, and the sequence is of length T, the the complexity is $\mathcal{O}(k^2T)$

6. What are the three components of the transducer model? At a high level, describe the role of each component. (4 pts)

- *Encoder - It is responsible for modeling the acoustics of the input audio.*
- *Predictor (Text Encoder, Also sometimes called Decoder) - It acts as a language model*
- *Joiner - It is responsible for merging the predictor and encoder outputs.*

Q6. Enhancement and Diarization *14 points*

1. How is mask-based enhancement performed and why is it useful for speech enhancement? (4pts)

    - *(1pt) The input signal is transformed into a spectral domain (i.e. has time and frequency resolution)*
    - *(1pt) The transform is invertible (waveform $\rightarrow$ spectrum $\rightarrow$ waveform)*
    - *(1pt) Each time-frequency bin is multiplied by a value in $[0, 1]$ based on the presence of the source/interference in that bin*
    - *(1pt) Signals have differing and identifiable structure in time-frequency domain*

2. What information is required to perform delay-and-sum beamforming in multi-microphone recordings and how is it used to enhance a target signal? (3pts)

    - *(1pt) Time Difference of Arrival (TDOA), the time delay between when the target source reaches each microphone*
    - *(1pt) The target source is aligned for constructive interference in the sum and is accordingly attenuated*
    - *(1pt) The interfering signals are out-of-alignment and are suppressed via destructive interference*
    - *(alternate 1pt) Knowing the direction of the target source*

3. In order, what are the four primary components of traditional diarization systems? (4pts)

    - *(1pt) Initial segmentation or speech activity detection*
    - *(1pt) Speaker representation extraction*
    - *(1pt) Clustering*
    - *(1pt) Resegmentation*

4. What are the two main diarization metrics and what is the primary difference in what/how they measure? (3pts)

    - *(1pt) Diarization Error Rate (DER)*
    - *(1pt) Jaccard Error Rate (JER)*
    - *(1pt) DER is normalized according to the total amount of speech while JER is normalized on a per-speaker basis*

Extra Space