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)

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

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

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)

2. Respond to the following questions (you might need to use the figure below) (4pt):
a) Which is the main articulator involved in changing the frequency formants that define nasal sounds?

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.

c) According to the IPA table, can Fricatives be bilabial? If affirmative, please provide two examples of labiodental fricative sounds.

d) According to the IPA table, can plosives be labiodental? If affirmative, please, provide some examples of glottal nasal sounds.

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 phones and phonemes (3pt)

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

Q3. Speaker Recognition *20 points*

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

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)

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)

Q4. 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).

   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)

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)

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).

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

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

Q5. Enhancement and Diarization *14 points*

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

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)

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

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

Extra Space