Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

Question 1(a): *Components of Speech [60 points]*

Describe and illustrate, what is (a) an analysis window: definition, typical length, for what etc., (b) a power spectrum, and (c) a spectrogram? On power spectrum as well as on spectrogram, what are the typical properties of the speech signal that can be observed? *(30 points)*

Illustrate and explain the differences in the auto-correlation function and the power spectrum for speech signals with pitch frequency (presuming production of same speech sound),
(i) F_0 = 100 Hz, and
(ii) F_0 = 400 Hz
What is your average pitch frequency? Suppose you are tasked to produce vowel /a/ in a sustained manner with a pitch frequency as close as possible to those pitch frequencies. Explain concisely what aspect you need to change and how during speech production to achieve that. *(20 points)*

What kind of spectrogram is best suited for observing formants? Justify concisely. *(5 points)*
What kind of spectrogram is best suited for observing pitch frequency? Justify concisely. *(5 points)*

Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

Question 1(b): *Speech Analysis-Synthesis [60 points]*

In the course, we have learned that the speech signal can be decomposed into source and system components, which can be put back together to get the speech signal.

a) **What are the two main methods to achieve that? Explain concisely.** *(10 points)*
b) **How is this understanding applied in speech coding to reduce the bit rate? Describe concisely what happens on the transmitter side and what happens on the receiver side. Illustrate with an example calculation how the bit rate is reduced when compared to sample-by-sample coding and transmission of the speech waveform.** *(20 points)*
c) **Illustrate and concisely describe how is this understanding applied to build statistical parametric, i.e., hidden Markov model-based (HMM) text-to-speech synthesis (TTS) system? Contrast the described HMM-based synthesis approach with respect to neural text-to-speech synthesis and concisely explain the differences between the two TTS approaches.** *(30 points)*

Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

Question 2(a): *Sequence Matching in Speech processing [60 points]*

a) **Describe concisely the principle of dynamic programming.** *(6 points)*
b) **How is it applied in instance-based speech recognition?** *(12 points)*
c) **How is it applied in the hidden Markov model based automatic speech recognition?** *(12 points)*
d) **How is it applied in concatenative text-to-speech synthesis?** *(10 points)*
e) **Compute the word error rate for the example given below.** *(algorithm: 5 points, illustration: 15 points)*

For each case, clearly describe the methodology, like what is being matched along with the typical equation, local score, local constraints, and the optimization criteria.

Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

car	8								
passing	7								
the	6								
at	5								
barked	4								
dog	3								
big	2								
the	1								
	0	1	2	3	4	5	6	7	8
		the	frog	barked	at	at	the	racing	car

Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

Question 2(b): Application of *Expectation-Maximization Algorithm* *[60 points]*

a) **Define concisely the general idea of EM algorithm.** *(5 points)*
b) **What are the parameters of a Gaussian mixture model? How is the EM algorithm employed to train parameters of Gaussian mixture model? How can we use Gaussian mixture modeling for speaker verification?** *(25 points)*
c) **What are the parameters of a hidden Markov model? How is the EM algorithm employed to train parameters of the hidden Markov model? In your answer, consider the two common ways to model emission distribution of HMM state.** *(30 points)*

For b) and c) clearly explain: what is the optimization criterion? What does E-step involve? and What does M-step involve?

Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

Question 3(a): *Speech Recognition and Speech Synthesis [60 points]*

Compare automatic speech recognition (ASR) and text-to-speech synthesis (TTS) systems.

- **What is the input and what is the output?** *(4 points)*
- **What are the major building blocks of each system?** *(14 points)*
- **What kind of resources are needed to build ASR and TTS systems? Which resources could be shared for development of the two systems?** *(12 points)*
- **How are speech synthesis systems evaluated? What kind of resources are needed for that? Explain concisely which aspect of speech synthesis system output can be objectively evaluated using automatic speech recognition system and how?** *(15 points)*
- **Today speech synthesis systems can generate highly natural speech. How can we detect if a given speech signal is synthetic speech or not? Concisely explain the methodology along with evaluation measures.** *(15 points)*

Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

Question 3(b): *Automatic Speech Recognition [60 points]*

Starting from the speech waveform, clearly describe the different processes (building blocks) involved in statistical continuous speech recognition and what type of prior information and/or models is being used.

a) **Clear block diagram of the processing steps, including inputs and outputs.** *(5 points)*
b) **How are the lexical constraints being modeled and exploited? Type of model? Training (if any)?** *(5 points)*
c) **How are the syntactic/grammatical constraints being modeled and exploited? Type of model? Training?** *(10 points)*
d) **How is the acoustic information modeled? Type of model? Training?** *(10 points)*
e) **Concisely explain how decoder puts together the different information together to output text. Optimization criteria?** *(10 points)*
f) **Suppose we want to use this speech recognition system framework to recognize phonetic sequences (instead of word sequences) in speech utterances (possibly for a new domain or language). Concisely describe what aspects change or do not change in a) - e) for this case. How will you evaluate such a system?** *(20 points)*

Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

Questions 4(a): *Automatic Speaker Analysis [60 points]*

In a home environment, suppose we want to make a robot/voice assistant be aware of with who in the family (consisting of N people) is the system interacting with.

a. **What kind of speech processing task is needed?** *(4 points)*
b. **What kind of speech features are relevant for this task? Justify concisely.** *(8 points)*
c. **What kind of machine learning method can be used to model the features? How the decision can be made (theoretical criterion)?** *(10 points)*
d. **How can we enable the robot/voice assistant detect "stranger's" voice (someone not part of the family)? Explain concisely the methodology.** *(8 points)*

Given an audio recording of a dyad conversation (conversation between two people, e.g. over telephone), how to detect speaker change time points in the audio? Concisely describe (a) the type of feature representations that can be used, (b) statistical modeling method and decision-making process, and (c) evaluation of the speaker change detection system. *(30 points)*

Automatic Speech Processing
Dr. Mathew Magimai Doss
2024/2025 Autumn semester exam
25 January 2025

EPFL

First name:
Family name:
Section:
SCIPER:

Question 4(b): *Paralinguistic speech processing [60 points]*

Define concisely the term paralinguistics. What do the notions "states" and traits refer to? Is accentedness a state or a trait? Suppose you are assigned the task of development of a system that assesses French speakers' speech in terms of degree of accentedness. How would you go about development of a data set for that purpose? Explain the key steps. *(30 points)*

Given the above-mentioned data set, how will you develop a degree of accentedness prediction system? Describe and justify concisely (a) choice of machine learning task, (b) choice of feature extraction and representation method, (c) choice of machine learning method, and (d) training and evaluation? *(30 points)*