

- 1. 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)**

Second part of the question (30 points)

## 2. Linear prediction

- **What is linear prediction?** (5 points)
- **Why is linear prediction modeling well suited for speech signal processing?** (5 points)
- **What do the linear prediction coefficients model? How are they estimated (cost function)?** (10 points)
- **What is the residual signal or the linear prediction error signal model? How is the residual signal obtained?** (10 points)
- ..... (30 points)

- a. Given “only” two speech utterances recordings (no additional information whatsoever is available), (a) how to automatically determine if those two speech utterances represent the same lexical item, i.e., word and (b) how to automatically determine if those utterances were uttered by the same speaker? Clearly explain the steps, the features, and the matching algorithm. Discuss concisely the decision errors. (30 points)

Second part of the question (30 points)

**4. Sequence matching in speech processing**

- a) Describe concisely the principle of dynamic programming. (6 points)
- b) How is it applied in unit selection concatenative text-to-speech speech synthesis? (12 points)
- c) How is it applied in the hidden Markov model based automatic speech recognition? (12 points)
- d) .....
- e) .....

**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.**

*Missing parts of the question will cover 30 points.*

**5. 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)*
- .....
- .....

*Missing parts of the question will cover 30 points.*

**6. 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.**

- i. Clear block diagram of the processing steps, including inputs and outputs. (5 points)**
- ii. How are the lexical constraints being modeled and exploited? Type of model? Training (if any)? (5 points)**
- iii. How are the syntactic/grammatical constraints being modeled and exploited? Type of model? Training? (10 points)**
- iv. How is the acoustic information modeled? Type of model? Training? (10 points)**

*Missing parts of the question will cover 30 points.*

## 7. Speaker analysis

**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)**

*Missing parts of the questions will cover 30 points.*

**8. Evaluation of speech processing systems.**

- **How to evaluate text-to-speech systems? (15 points)**
- **How to evaluate automatic speaker verification systems? (15 points)**
- .....
- .....

**For each of these systems, describe concisely the types of errors, the performance measure, the method to measure performance and the resources needed to evaluate the systems.**

*Missing parts of the question will cover 30 points.*

**9. What is the main difference between statistical automatic speech recognition (ASR) and automatic speaker verification (ASV)?**

- **Definition of each problem (goal, input, output). (5 points)**
- **The “theoretical” decision criteria for each case? (5 points)**
- **Describe concisely how the different statistical quantities in the decision criteria for ASR and ASV are modeled? (20 points)**
- .....
- .....

*Missing parts of the question will cover 30 points.*

**10. Describe text-to-speech synthesis system**

- **What is input? What is output? (3 points)**
- **Contrast the three common text-to-speech synthesis approaches: (a) unit selection concatenative speech synthesis, (b) Statistical parametric synthesis (HMM-based) speech synthesis and (c) neural speech synthesis in terms of formulation, pros and cons. (27 points)**
- .....
- .....

*Missing parts of the questions will cover 30 points.*

### 11. Sequence modeling using Markov models

- Define concisely discrete Markov models (DMM)? What are the parameters of a DMM? How is DMM employed in statistical automatic speech recognition systems? How are the DMM parameters estimated? (15 points)
- Define concisely hidden Markov models (HMM)? What are the parameters of an HMM? How is HMM employed in statistical automatic speech recognition systems? How are the HMM parameters estimated? (15 points)

*Missing parts of the question will cover 30 points*

**12. Lexical constraints**

- a) Illustrate the following hidden Markov model (HMM) topologies
  - i. a K-state left-to-right HMM
  - ii. a K-state fully connected ergodic HMM

For illustration, choose a value of K of your choice. *(10 points)*

- b) In phone-based automatic speech recognition systems, what resource is needed to integrate lexical constraints and how is it obtained? How can the lexical constraints for new words (e.g., a new name, a new place) be obtained in an “automatic” manner? How can pronunciation variation be handled? Which of the HMM topology in part a) is best suited to integrate lexical constraints in ASR systems? Justify concisely. *(20 points)*

- c) ..... *(30 points)*

**13. Expectation-Maximization (EM) algorithm**

**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) ..... (30 points)**

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

*Missing parts of the question will cover 30 points*

#### 14. Paralinguistic speech processing

**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)**

**Suggested reading:** see EMIME bilingual corpus<sup>1</sup>

*Second part of the question will cover 30 points.*

---

<sup>1</sup> [https://www.cstr.inf.ed.ac.uk/downloads/publications/2010/wester\\_accent\\_2010.pdf](https://www.cstr.inf.ed.ac.uk/downloads/publications/2010/wester_accent_2010.pdf)

### 15. Statistical Speech Recognition

Let  $X = \{x_1, \dots, x_T\}$  be a sequence of acoustic features and  $W = \{w_1, \dots, w_K\}$  a sequence of words.

- How can  $P(X|W)$  be estimated? Type of model? What assumptions are required to efficiently estimate its parameters? What resources are necessary to build the  $P(X|W)$  estimator? Discuss concisely the answers. (10 points)
- How can  $P(W)$  be estimated? Type of model? What assumptions are required to efficiently estimate its parameters? What resources are necessary to build the  $P(W)$  estimator? Discuss concisely the answers. (10 points)
- How are estimates  $P(X|W)$  and  $P(W)$  are put together for speech recognition? What does it involve? (10 points)
- ..... (30 points)

## 16. Speech processing-based application

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)*

*Second part of the question will cover 30 points.*

**17. In the course, we have seen that speech signal carries different types of information.**

**What kind of speech information can be extracted/modeled from,**

- a. a 25 ms frame speech signal? (10 points)**
- b. a 250 ms segment of speech signal? (10 points)**
- c. a 7 second speech utterance? (10 points)**

**Provide a concise/short justification for each.**

*Second part of the question will cover 30 points.*

**18. Frame a question of your own based on what you have learned in the class, i.e., course content and answer it. A maximum of 20 points for the quality of the question framed (as different as possible from the questions that are already on the course Moodle). 10 points for the answer.**

*Second part of the question will cover 30 points.*