Spoken Language Processing 2022/23

Exam

Duration : 90 minutes.

Student number (6 digits):	First and last name:
... ...	....

Answers must be given exclusively on the answer sheet. Answers given on other sheets will be ignored.
All multiple-choice questions have exactly one correct answer.
For questions 1 to 30, each correct answer is worth 0.5 point. Very incorrect answers are worth -0.25 points.
Other incorrect answers, more than one answer and questions left unanswered are worth 0 points.
Open question 31 is worth 1.66 points each. Open questions 32 and 33 are worth 1.67 points.

Question 1 Consider two spectrograms, one with an FFT size of 128 samples and another with an FFT size of 1024. If the sampling rate of the speech signal is 16 kHz, which is the true statement?

■ the one with an FFT size of 128 is a wideband spectrogram
B the one with an FFT size of 128 is a narrowband spectrogram
C both spectrograms are wideband
D both spectrograms are narrowband

Question 2 What is the general case of a transfer function of a discrete-time LTI system defined by a difference equation?

A A polynomial function
B A discrete function
C An exponential function
■ A rational function

Question 3 What do rhythm, stress, and intonation have in common?

A They are all features of speech production
B They are all processed in the auditory system
■ They are all related to the timing and melody of speech
D They are all related to the classification of speech sounds

Question 4 What is the length of the impulse response of an ideal digital low-pass filter with a cutoff frequency of f_c?

A The impulse response has a finite length of N samples
B The impulse response has a length of $2\pi/f_c$ seconds
■ The impulse response is infinitely long
D The length of the impulse response depends on the sampling frequency

Question 5 Consider that the vector $x(n)$ contains a segment of a vowel sampled at 10^4 Hz.
The autocorrelation function of the vector:

$$R_{xx}(m) = \sum_{n=-\infty}^{+\infty} x(n)x(n+m)$$

shows a peak in lag $m = 50$.
What is the fundamental frequency of the vowel?

A 100 Hz ■ 200 Hz C 50 Hz D 150 Hz

Question 6 What is the difference between the Mel and linear frequency scales in speech processing?

- A The Mel scale is used to represent the duration of speech sounds, while the linear scale is used to represent the intensity of speech sounds.
- ■ The Mel scale is a non-linear scale that represents the perceived frequency of sound by the human ear, while the linear frequency scale represents the actual physical frequency of sound.
- C The Mel scale is used to represent the intensity of speech sounds, while the linear scale is used to represent the frequency of speech sounds.
- D The Mel scale is a linear scale that represents the perceived frequency of sound by the human ear, while the linear frequency scale represents the actual physical frequency of sound.

Question 7 Which of the following characteristics are associated with the breathy voice?

- A Loud and forceful
- B Rough and gravelly
- C Clear and resonant
- ■ Weak and whispery

Question 8 Consider a finite-duration speech signal with 102400 samples. If we wish to process the signal
without any zero padding using a window of 1024 samples with a hop length of 128 samples,
how many frames need to be processed?

A 100 ■ 793 C 512 D 800

Question 9 Consider a speech classification model based on a Transformer encoder, with a stack of 6 multi-head self-attention modules and 2 attention heads. How many attention weights are computed within the model when processing an input sequence of 10 elements?

■ 120 B 10 C 1200 D 3600

Question 10 In speaker recognition, the task of speaker verification can be defined as the task that:

- A Given a speech sample, decides to which of a set of fixed identities the sample belongs
- ■ Given a speech sample and a claimed identity decide if the sample belongs to the claimed identity
- C Given a speech sample and text transcription, verifies if the speech sample content corresponds to the text
- D Given a speech sample, a text transcription and a claimed identity, verifies if the speech sample content corresponds to the text and if the speech sample belongs to the claimed identity

Question 11 Regarding multilayer perceptrons (MLP), which of the following statements about MLPs **is false**:

- A MLPs can have more than one hidden layer
- ■ MLP is a generative model extensively used to model speech
- C MLP is a type of feed-forward neural network
- D MLPs are trained using backpropagation

Question 12 Consider the encoder-decoder Transformers that were introduced in the lectures for modeling speech. What are some of the main differences between the encoder and decoder parts of these architectures?

A Only the decoder features multi-head self-attention.

B Only the encoder features cross-attention operations.

■ The encoder uses bi-directional attention operations, while the decoder uses masked causal attention operations.

D The decoder uses multi-head self-attention, whereas the encoder uses regular self-attention operations.

Question 13 Consider the computations associated to the dot-product self-attention operation. Consider also an input sequence of four vectors [[8,0,0,0], [8,4,0,0], [8,0,4,0], [8,0,0,4]], and consider that queries, keys, and values are all computed through the projection matrix [[1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]] (i.e., diagonal 4x4 matrices with the same values). What would be the result of the dot-product self-attention operation for the first element in the sequence?
Recall that the softmax operation returns a uniform probability distribution when the input vectors have the same values in all dimensions.

A [64,0,0,0] B [16,4,4,4] C [8,4,4,4] ■ [8,1,1,1]

Question 14 Consider the numpy array formed by N feature rows each with dimension D, that is, with shape (N, D). What is the size of the array after concatenating the first-order delta features?

■ $(N, 2D)$ B $(2N, 2D)$ C (N, D) D $(2N, D)$

Question 15 Consider many-to-one Transformers, versus other Transformers that can address many-to-many tasks. Which of the following models, introduced in the classes, corresponds to many-to-one architectures?

■ Audio spectrogram Transformer

B Wav2vec 2.0

C VALL-E and VALL-E X

D SpeechT5 and Whisper

Question 16 Some recent trends for speech classification rely on self-supervised training followed by fine-tuning. Which of the following statements about self-supervised learning **is true**?

A In the self-supervised stage, partially labeled data is used to train a general-purpose speech model

B In the self-supervised stage, unlabelled data is used to train a task-specific speech model

■ In the self-supervised stage, unlabelled data is used to train a general-purpose speech model

D In the self-supervised stage, partially labeled data is used to train a task-specific speech model

Question 17 Imagine that you want to train an ASR system for European Portuguese. Which one of the following feature sets should be the first one to consider?

A x-vectors

B Voice-quality features

■ Mel-frequency cepstral coefficients

D Zero-crossing rate

Question 18 Gaussian mixture models (GMM) used to be an extremely popular model in speech classification tasks. GMMs can be classified as:

A a neural model

■ a generative model

C an autoregressive model

D a discriminant model

Question 19 Model-based or data-driven features are feature representations that are typically obtained from activations of a neural network. Which one of the following types of features **is not** a common model-based set used in speech classification?

- A Posterior features
- ■ Context-dependent features
- C Embedding features
- D Bottleneck features

Question 20 Which parts/components of the self-attention operation, as used in Transformer models, are calculated by passing inputs through a linear projection?

- A Word embeddings
- B Queries and keys
- ■ Queries, keys, and values
- D Values

Question 21 Consider a speech classification model based on a Transformer encoder, with a stack of 6 multi-head self-attention modules. The input embeddings of dimension 128 match the output shape of the self-attention layers. If we use multi-headed attention, with 4 heads, what dimensionality will the outputs of each head have?

■ 32 B 64 C 512 D 128

Question 22 Consider the OpenAI Whisper model, capable of addressing multiple speech tasks through a joint formulation. What are examples of tasks that the model CANNOT handle?

- A Speech translation, multilingual speech recognition
- B Voice activity detection and spoken language identification
- C Voice activity detection
- ■ Speech enhancement

Question 23 Transformer encoder models for speech (e.g., DiscreteBERT, variations on wav2vec, etc.) can be pre-trained with self-supervised objectives similar to masked language modeling, although the masked tokens/positions are frequently expanded into spans. What is the main reason for this?

- ■ Avoid exploring local smoothness
- B Facilitate the combination with contrastive pre-training objectives
- C Facilitate the processing of long input sequences
- D Speed-up model training

Question 24 The WaveNet autoregressive model uses a stack of convolution layers. To extend its range and capture more context it uses:

- A a large number of causal convolution layers
- B a small number of causal convolution layers
- C non-causal convolution layers
- ■ dilated convolution layers

Question 25 Complete the following sentence: When an ASR system is used to perform forced alignment...

- A The pronunciation model is not used
- B The decoder is not necessary
- ■ The language model is not used
- D The acoustic model is not used

Question 26 Conventional hierarchical large vocabulary continuous speech recognition systems (LVCSR) are composed of three sources of information or models. Which one of the following is not one of the three models?

- A Pronunciation model
- B Acoustic model
- ■ Semantic model
- D Language model

Question 27 Which of the following statements corresponds to the advantages of the use of large language models for developing task-oriented dialogue systems?

A Better avoid the generation of generic responses.

B Increase interpretability and controllability.

C Improve the integration with ASR and TTS components, and also the interaction with external databases.

■ Ease the unification of chit-chat and task-oriented dialogue.

Question 28 The addition of CTC loss to seq2seq attention architectures (a.k.a Hybrid CTC/Attention) helps to:

A Reduce the size of the model

■ Enforce monotonic attention alignments

C Perform online decoding

D Augment the vocabulary size

Question 29 The dictionary in a grapheme-to-phoneme conversion system needs to include the character-to-sound correspondences of:

■ the common words in the language whose pronunciation cannot be predicted

B the common words in the language whose pronunciation can be predicted

C the common words in the language

D the common words in the language that are not homographs

Question 30 Consider the BERTScore metric for evaluating language generation. Which of the following statements is correct:

A The metric includes a brevity penalty to avoid giving high scores to short-generated sentences.

B The metric corresponds to the use of a BERT model trained to estimate BLEU scores.

■ The metric depends on ground-truth references to evaluate language generation.

D The metric is based on n-gram overlaps.

Question 31 In the source-filter model used in a vocoder, the source for unvoiced sounds is modeled with a zero-mean unite variance Gaussian white noise:

$$e_u(n) \sim \mathcal{N}(0, 1)$$

The root mean square (RMS) of this signal is:

$$\sqrt{E([e_u(n)]^2)} = 1$$

where $E()$ is the expected value function. In this model, the voiced sounds are modeled with a pulse train:

$$e_v(n) = A \sum_{k=-\infty}^{+\infty} \delta(n - kP)$$

where $\delta(n)$ is the unit pulse, and P the fundamental period in samples.
Find the value of A to ensure that the RMS of the source signal is always 1. Explain problems that may occur when the fundamental frequency is not constant and in voiced-unvoiced transitions

Question 32 Consider the x-vector model introduced in the lectures that was explored in the second lab of the course. Briefly describe the model, its architecture, main characteristics, how it is trained, advantages with respect to previous state-of-the-art models, and how it can be used for spoken language identification. You can draw a diagram to support your description.

Question 33 Consider the use of large language models for developing end-to-end task-oriented dialogue systems. Outline the general working of these systems, in comparison to traditional/modular task-oriented dialogue systems.

Student number (6 digits):

0 0 0 0 0 0
1 1 1 1 1 1
2 2 2 2 2 2
3 3 3 3 3 3
4 4 4 4 4 4
5 5 5 5 5 5
6 6 6 6 6 6
7 7 7 7 7 7
8 8 8 8 8 8
9 9 9 9 9 9

Answer Sheet

Answers must be given exclusively on this sheet. Answers given on other sheets will be ignored.

No corrections are allowed on this sheet.

Encode your student number (6 digits) by selecting the digits on the left, starting with 0 if it has just 5 digits, and write your name below.

First and last name:

....

Question 1: ■ B C D
Question 2: A B C ■
Question 3: A B ■ D
Question 4: A B ■ D
Question 5: A ■ C D
Question 6: A ■ C D
Question 7: A B C ■
Question 8: A ■ C D
Question 9: ■ B C D
Question 10: A ■ C D
Question 11: A ■ C D
Question 12: A B ■ D
Question 13: A B C ■
Question 14: ■ B C D
Question 15: ■ B C D
Question 16: A B ■ D
Question 17: A B ■ D
Question 18: A ■ C D
Question 19: A ■ C D
Question 20: A B ■ D
Question 21: ■ B C D
Question 22: A B C ■
Question 23: ■ B C D
Question 24: A B C ■
Question 25: A B ■ D
Question 26: A B ■ D
Question 27: A B C ■
Question 28: A ■ C D
Question 29: ■ B C D
Question 30: A B ■ D

Question 31: 0 1 2 ■

The pulse is a sequence of pulses of amplitude A followed by $P-1$ zeros. The RMS value of one period of the signal is:

$$\sqrt{\frac{1}{P}\sum_{n=0}^{P-1}[e_v(n)]^2} = \frac{A}{\sqrt{P}}$$

To ensure that the RMS of the source signal is 1, the amplitude of the pulses must be:

$$A = \sqrt{P}$$

When the fundamental frequency changes with time, the value of the fundamental period should only be updated at the end of the sequence of $P-1$ zeros.
When a voiced-unvoiced transition occurs in the middle of a period, the pulse amplitude must be adjusted to the square root of the distance to the transition in samples.

Question 32:

0 1 2 ■

An x-vector model is essentially a DNN that is typically trained for speaker identification using conventional backpropagation. The input to the network consists of frame-level features (such as log-mel spectrum features or MFCCs). At some point in the architecture, there is a pooling layer (statistical, attentive, etc.) that collapses the activations of all the frames of a segment into a single vector, making the remaining blocks work at the segment level.

The x-vector embedding corresponds to the activations of one of the feed-forward layers that operate at the segment level. These embeddings compress in a single vector speaker discriminant information, and it was proposed as a successor to previous speaker vector representations, such as GMM-MAP and i-vectors.

The main advantage is that it is based on discriminative deep learning and exploits all the advances in this field. In the lab for spoken language identification, we used an x-vector model that was pre-trained for language identification (107 languages), instead of speaker identification. The extracted embeddings could then be used to train a simple model using the provided training data of the six target languages, such as an SVM or k-NN.

Question 33:

0 1 2 ■

End-to-end task-oriented dialogue systems, based on the use of large language models, often rely on the use of text prompts to formulate/encode the typical sub-tasks of modular task-oriented dialogue (i.e., language understanding, dialogue management, and natural language generation).

Starting from the dialogue history, these systems mimic the inputs and outputs from the different modules as sequences of tokens to be processed/generated by the language model in an end-to-end method (e.g., first generating tokens corresponding to the domain and intent recognition, conditioned on the dialogue history, then generating a dialogue action conditioned on the previous outputs, and finally generating the responses).

Modular systems, as the name implies, instead use separate components (often also corresponding to language models) to perform language understanding, dialogue management, and natural language generation.