

**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?

- ☐ A 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
- ☐ D 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
- ☐ C 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
- ☐ C The impulse response is infinitely long
- ☐ D The length of the impulse response depends on the sampling frequency

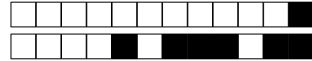
Question 5 Consider that the vector $x(n)$ contains a segment of a vowel sampled at 10^4 Hz. The autocorrelation function of the vector:

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

shows a peak in lag $m = 50$.

What is the fundamental frequency of the vowel?

- ☐ A 100 Hz
- ☐ B 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.
- ☐ B 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
- ☐ D 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
- ☐ B 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?

- ☐ A 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
- ☐ B 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
- ☐ B 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.
- ☐ C 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 4×4 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 $[8,4,4,4]$
- ☐ D $[8,1,1,1]$

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?

- ☐ A $(N, 2D)$
- ☐ 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?

- ☐ A 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
- ☐ C 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
- ☐ C 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
- ☐ B 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
- ☐ B 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
- ☐ C 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?

- ☐ A 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
- ☐ D 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?

- ☐ A 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
- ☐ D 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
- ☐ C 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
- ☐ C 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.
- ☐ D 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
- ☐ B 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:

- ☐ A 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.
- ☐ C 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 unit 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.



+1/6/55+

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: ☐ A ☐ B ☐ C ☐ D

QUESTION 2: ☐ A ☐ B ☐ C ☐ D

QUESTION 3: ☐ A ☐ B ☐ C ☐ D

QUESTION 4: ☐ A ☐ B ☐ C ☐ D

QUESTION 5: ☐ A ☐ B ☐ C ☐ D

QUESTION 6: ☐ A ☐ B ☐ C ☐ D

QUESTION 7: ☐ A ☐ B ☐ C ☐ D

QUESTION 8: ☐ A ☐ B ☐ C ☐ D

QUESTION 9: ☐ A ☐ B ☐ C ☐ D

QUESTION 10: ☐ A ☐ B ☐ C ☐ D

QUESTION 11: ☐ A ☐ B ☐ C ☐ D

QUESTION 12: ☐ A ☐ B ☐ C ☐ D

QUESTION 13: ☐ A ☐ B ☐ C ☐ D

QUESTION 14: ☐ A ☐ B ☐ C ☐ D

QUESTION 15: ☐ A ☐ B ☐ C ☐ D

QUESTION 16: ☐ A ☐ B ☐ C ☐ D

QUESTION 17: ☐ A ☐ B ☐ C ☐ D

QUESTION 18: ☐ A ☐ B ☐ C ☐ D

QUESTION 19: ☐ A ☐ B ☐ C ☐ D

QUESTION 20: ☐ A ☐ B ☐ C ☐ D

QUESTION 21: ☐ A ☐ B ☐ C ☐ D

QUESTION 22: ☐ A ☐ B ☐ C ☐ D

QUESTION 23: ☐ A ☐ B ☐ C ☐ D

QUESTION 24: ☐ A ☐ B ☐ C ☐ D

QUESTION 25: ☐ A ☐ B ☐ C ☐ D

QUESTION 26: ☐ A ☐ B ☐ C ☐ D

QUESTION 27: ☐ A ☐ B ☐ C ☐ D

QUESTION 28: ☐ A ☐ B ☐ C ☐ D

QUESTION 29: ☐ A ☐ B ☐ C ☐ D

QUESTION 30: ☐ A ☐ B ☐ C ☐ D

QUESTION 31:

0	1	2	3
---	---	---	---



+1/8/53+

QUESTION 32:

0 1 2 3

QUESTION 33:

0 1 2 3

**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 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 200 Hz ☐ B 100 Hz ☐ C 150 Hz ☐ D 50 Hz

Question 2 Which of the following characteristics are associated with the breathy voice?

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

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

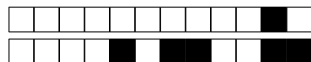
- ☐ A They are all related to the classification of speech sounds
☐ B They are all related to the timing and melody of speech
☐ C They are all processed in the auditory system
☐ D They are all features of speech production

Question 4 What is the general case of a transfer function of a discrete-time LTI system defined by a difference equation?

- ☐ A A rational function ☐ C A polynomial function
☐ B A discrete function ☐ D An exponential function

Question 5 What is the difference between the Mel and linear frequency scales in speech processing?

- ☐ A 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.
☐ B 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.
☐ C 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.
☐ D 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.



Question 6 What is the length of the impulse response of an ideal digital low-pass filter with a cutoff frequency of f_c ?

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

Question 7 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?

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

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 ☐ B 800 ☐ C 512 ☐ D 793

Question 9 Regarding multilayer perceptrons (MLP), which of the following statements about MLPs is false:

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

Question 10 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?

- ☐ A 120 ☐ B 10 ☐ C 3600 ☐ D 1200

Question 11 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?

- ☐ A 512 ☐ B 128 ☐ C 32 ☐ D 64

Question 12 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 Context-dependent features
- ☐ B Posterior features
- ☐ C Bottleneck features
- ☐ D Embedding features



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 4×4 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 $[8,1,1,1]$ ☐ B $[8,4,4,4]$ ☐ C $[16,4,4,4]$ ☐ D $[64,0,0,0]$

Question 14 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 The encoder uses bi-directional attention operations, while the decoder uses masked causal attention operations.
- ☐ B Only the encoder features cross-attention operations.
- ☐ C The decoder uses multi-head self-attention, whereas the encoder uses regular self-attention operations.
- ☐ D Only the decoder features multi-head self-attention.

Question 15 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, unlabelled data is used to train a general-purpose speech model
- ☐ B In the self-supervised stage, partially labeled data is used to train a general-purpose speech model
- ☐ C In the self-supervised stage, unlabelled data is used to train a task-specific speech model
- ☐ D In the self-supervised stage, partially labeled data is used to train a task-specific speech model

Question 16 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 Speech enhancement
- ☐ C Voice activity detection and spoken language identification
- ☐ D Voice activity detection

Question 17 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?

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

Question 18 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?

- ☐ A Wav2vec 2.0 ☐ C SpeechT5 and Whisper
- ☐ B VALL-E and VALL-E X ☐ D Audio spectrogram Transformer



Question 19 Gaussian mixture models (GMM) used to be an extremely popular model in speech classification tasks. GMMs can be classified as:

- ☐ A a neural model
- ☐ B a generative model
- ☐ C an autoregressive model
- ☐ D a discriminant model

Question 20 In speaker recognition, the task of speaker verification can be defined as the task that:

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

Question 21 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 Mel-frequency cepstral coefficients
- ☐ B Voice-quality features
- ☐ C Zero-crossing rate
- ☐ D x-vectors

Question 22 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?

- ☐ A (N, D)
- ☐ B $(N, 2D)$
- ☐ C $(2N, 2D)$
- ☐ D $(2N, D)$

Question 23 Which parts/components of the self-attention operation, as used in Transformer models, are calculated by passing inputs through a linear projection?

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

Question 24 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 Semantic model
- ☐ B Acoustic model
- ☐ C Language model
- ☐ D Pronunciation model

Question 25 The WaveNet autoregressive model uses a stack of convolution layers. To extend its range and capture more context it uses:

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

Question 26 Which of the following statements corresponds to the advantages of the use of large language models for developing task-oriented dialogue systems?

- ☐ A Ease the unification of chit-chat and task-oriented dialogue.
- ☐ B Improve the integration with ASR and TTS components, and also the interaction with external databases.
- ☐ C Increase interpretability and controllability.
- ☐ D Better avoid the generation of generic responses.



Question 27 The dictionary in a grapheme-to-phoneme conversion system needs to include the character-to-sound correspondences of:

- ☐ A the common words in the language that are not homographs
- ☐ B the common words in the language
- ☐ C the common words in the language whose pronunciation cannot be predicted
- ☐ D the common words in the language whose pronunciation can be predicted

Question 28 Complete the following sentence: When an ASR system is used to perform forced alignment...

- | | |
|--|---|
| <input type="checkbox"/> A The language model is not used | <input type="checkbox"/> C The acoustic model is not used |
| <input type="checkbox"/> B The pronunciation model is not used | <input type="checkbox"/> D The decoder is not necessary |

Question 29 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.
- ☐ C The metric depends on ground-truth references to evaluate language generation.
- ☐ D The metric is based on n-gram overlaps.

Question 30 The addition of CTC loss to seq2seq attention architectures (a.k.a Hybrid CTC/Attention) helps to:

- | | |
|---|---|
| <input type="checkbox"/> A Enforce monotonic attention alignments | <input type="checkbox"/> C Reduce the size of the model |
| <input type="checkbox"/> B Augment the vocabulary size | <input type="checkbox"/> D Perform online decoding |

Question 31 In the source-filter model used in a vocoder, the source for unvoiced sounds is modeled with a zero-mean unit 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:

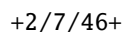
$$e_v(n) = A \sum_{k=-\infty}^{+\infty} \delta(n - kP)$$

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.



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

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

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.

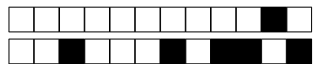
.....

- QUESTION 1: ☐ A ☐ B ☐ C ☐ D
- QUESTION 2: ☐ A ☐ B ☐ C ☐ D
- QUESTION 3: ☐ A ☐ B ☐ C ☐ D
- QUESTION 4: ☐ A ☐ B ☐ C ☐ D
- QUESTION 5: ☐ A ☐ B ☐ C ☐ D
- QUESTION 6: ☐ A ☐ B ☐ C ☐ D
- QUESTION 7: ☐ A ☐ B ☐ C ☐ D
- QUESTION 8: ☐ A ☐ B ☐ C ☐ D
- QUESTION 9: ☐ A ☐ B ☐ C ☐ D
- QUESTION 10: ☐ A ☐ B ☐ C ☐ D
- QUESTION 11: ☐ A ☐ B ☐ C ☐ D
- QUESTION 12: ☐ A ☐ B ☐ C ☐ D
- QUESTION 13: ☐ A ☐ B ☐ C ☐ D
- QUESTION 14: ☐ A ☐ B ☐ C ☐ D
- QUESTION 15: ☐ A ☐ B ☐ C ☐ D

- QUESTION 16: ☐ A ☐ B ☐ C ☐ D
- QUESTION 17: ☐ A ☐ B ☐ C ☐ D
- QUESTION 18: ☐ A ☐ B ☐ C ☐ D
- QUESTION 19: ☐ A ☐ B ☐ C ☐ D
- QUESTION 20: ☐ A ☐ B ☐ C ☐ D
- QUESTION 21: ☐ A ☐ B ☐ C ☐ D
- QUESTION 22: ☐ A ☐ B ☐ C ☐ D
- QUESTION 23: ☐ A ☐ B ☐ C ☐ D
- QUESTION 24: ☐ A ☐ B ☐ C ☐ D
- QUESTION 25: ☐ A ☐ B ☐ C ☐ D
- QUESTION 26: ☐ A ☐ B ☐ C ☐ D
- QUESTION 27: ☐ A ☐ B ☐ C ☐ D
- QUESTION 28: ☐ A ☐ B ☐ C ☐ D
- QUESTION 29: ☐ A ☐ B ☐ C ☐ D
- QUESTION 30: ☐ A ☐ B ☐ C ☐ D

0 1 2 3

[illegible]



+2/8/45+

QUESTION 32:

0 1 2 3

QUESTION 33:

0 1 2 3