Spoken Language Processing 2022/23

Model 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 questions 31 and 32 are worth 1.5 points each. Open question 33 is worth 2 points.

Question 1 Which of the following best describes a "prosody unit" in speech?

- A The study of how speech sounds are produced and perceived.
- B The smallest unit of speech, typically consisting of a single sound or phoneme.
- ■ A unit of speech consisting of one or more words, typically marked by pauses or changes in pitch or intonation.
- D The process of combining individual sounds or phonemes into words and sentences.

Question 2 Which phonation type is characterized by the lowest F0?

A falsetto ■ vocal fry C whisper phonation D modal

Question 3 What is the difference between sound intensity and loudness?

- A Sound intensity and loudness are two terms that describe the same physical property of sound waves.
- ■ Sound intensity is a physical property of sound waves, while loudness is a subjective perception.
- C Loudness is a physical property of sound waves, while sound intensity is a subjective perception.
- D Sound intensity and loudness are two different types of sound waves that can be distinguished by their frequency.

Question 4 What is the main difference between the Discrete Fourier Series (DFS) and the Discrete-Time Fourier Transform (DTFT)?

- A The DFS is used to represent aperiodic signals while the DTFT is used to represent periodic signals
- B The DFS is used to represent periodic signals while the DTFT is used to represent aperiodic signals
- C The DFS and DTFT are essentially the same thing and can be used interchangeably
- ■ The DFS is computed at a finite set of frequencies while the DTFT is computed over all frequencies

Question 5 How many unique values has the magnitude of the discrete Fourier transform (DFT) of a sequence of real values of length N=1024?

A 512 ■ 513 C 1024 D 1023

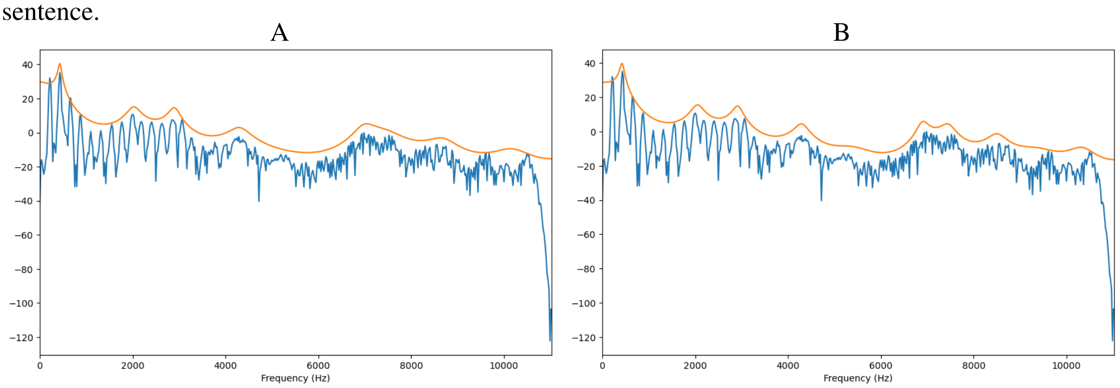
Question 6 In the source-filter model, the filter is used to model

- A the resonances in the voiced phonation
- B the glottal airflow velocity
- ■ the spectral envelope of speech sounds
- D the harmonic part of the spectrum

Question 7 Consider a finite-duration speech signal with 25600 samples. We wish to process the signal with zero padding using a window of 512 samples with a hop length of 64 samples. How many zero-padding samples are needed to have 400 frames?

A 64 B 512 ■ 448 D 0

Question 8 The following graphs show two spectral envelopes of the same vowel. Select the true sentence.



A Graph B was computed with a higher number of mel-frequency bins than Graph A

B Graph B was computed with a higher FFT order than Graph A

C Graph B was computed with a lower FFT order than Graph A

■ Graph B was computed with a higher LPC order than Graph A

Question 9 Imagine that you want to train an ASR system for European Portuguese. Which of the following feature sets should be the last one to consider?

A Log-mel-spectrograms

B Mel-frequency cepstral coefficients

C Linear predictive coding (LPC) features

■ Prosodic or pitch features

Question 10 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**?

■ In the fine-tuning stage, labeled data is used to fine-tune an existing general-purpose speech model to obtain a new task-specific speech model

B In the fine-tuning stage, unlabelled data is used to fine-tune an existing general-purpose speech model to obtain a new task-specific speech model

C In the fine-tuning stage, unlabelled data is used to fine-tune an existing task-specific speech model to obtain a new general-purpose speech model

D In the fine-tuning stage, labeled data is used to fine-tune an existing task-specific speech model to obtain a new general-purpose speech model

Question 11 Feature extraction methods for speech classification can be coarsely classified according to the region of analysis. Which of the following **does not** correspond to one of these categories?

■ Invariant B Segmental C Global D Local

Question 12 In speaker recognition, consider an x-vector extractor that is trained using MFCCs of 40 dimensions as input features with a frame hop size of 10 ms. If we apply the x-vector extractor to the MFCCs computed for an audio signal of 10 seconds, the size of the output will be:

■ A single vector of dimension equal to the size of the embedding layer

B 1000 vectors each one of dimension equal to the size of the embedding layer

C A single vector of dimension 40

D 1000 vectors of dimension 40

Question 13 Support vector machines (SVM) are extremely popular machine learning models that have been extensively used for some speech classification tasks. SVMs can be classified as:

■ a discriminant model

B a generative model

C an autoregressive model

D a neural model

Question 14 OpenSMILE is a toolbox widely adopted by the speech community. It is specially dedicated to:

■ Extract features adequate for paralinguistic classification tasks

B Extract features adequate for linguistic classification tasks

C Train generative linguistic models

D Train discriminative paralinguistic models

Question 15 Consider a speech classification model based on a Transformer encoder, with a stack of 4 multi-head self-attention modules and 2 attention heads. How many query, plus key, plus value vectors are produced within the model when processing an input sequence of 20 elements?

A 80

■ 480

C 20

D 160

Question 16 Which parts/components of the self-attention operation, as used in Transformer models, are involved in the computation of attention weights?

A Values

■ Queries and keys

C Queries, keys, and values

D Queries, keys, values, word embeddings, and position embeddings

Question 17 Consider the use of Transformer models for speech processing tasks. Why do models typically add (or concatenate) position encodings to the inputs?

A To increase robustness to adversarial attacks in the embeddings for each element in the input sequence

■ Because the dot-product self-attention operation is agnostic to the ordering of the input sequence

C To help in the propagation of gradient information during model training

D To avoid biases associated with the processing of particular positions (e.g., the first and/or last elements) within the inputs

Question 18 Consider the SpeechT5 and OpenAI Whisper models that were introduced in the classes. In what way do these models differ?

- A Only one of the models uses representations based on log Mel spectrograms
- B Only one of the models can be used for speech recognition tasks
- ■ Only one of the models can produce multi-modal outputs
- D Only one of the models is based on a sequence-to-sequence Transformer

Question 19 Consider the computations associated to the dot-product self-attention operation. Consider also an input sequence of four vectors [[10,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]], 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). What would be the result of the dot-product self-attention operation for the first element in the sequence?
You can consider approximating the softmax operation (e.g., if differences between the values in the input vectors are high, softmax returns a peaked distribution that closely resembles a one-hot vector).

- A Approximately [100,0,0,0]
- B Approximately [10,1,1,1]
- ■ Approximately [10,0,0,0]
- D Approximately [2,0,0,0]

Question 20 Consider the wav2vec and wav2vec 2.0 models introduced in the lectures. What is the main idea behind the contrastive predictive coding task that is used for model pre-training?

- A Generate the text representation for a given speech input
- B Mask some of positions in the input sequence and reconstruct the masked inputs
- ■ For some input positions, distinguish the correct representation from a set with distractors sampled from other positions
- D Distinguish the correct text representation, for a given speech input, from a set with distractors

Question 21 N-gram statistical models have been traditionally used to model language in hierarchical ASR systems. Which of the following statements about n-grams is **false**?

- ■ n-grams produce syntactically correct sentences
- B n-grams are commonly defined considering a limited-size vocabulary
- C n-grams model the probability of the next word depending on the n-1 previous ones
- D large amounts of data are required to reliably estimate probabilities

Question 22 Hidden Markov Models have been traditionally used in ASR as a statistical tool for:

- ■ Acoustic modeling
- B Pronunciation modeling
- C Semantic modeling
- D Language modeling

Question 23 In Hybrid HMM/DNN systems for automatic speech recognition:

- A A single DNN model is used for both acoustic and language modeling
- B The transition probabilities of HMM models are computed by a DNN
- ■ DNN training requires frame-level alignments between the audio and the recognition units
- D The HMM is used as the language model and the DNN as the acoustic model

Question 24 Considering the following alignment between a text reference and the hypothesis generated by an ASR system:

REF: speech ** RECOGNITON is the task of transcribing audio into text
HYP: speech THE COGNITION ** ** task of transcribing audio into text

The word error rate (WER) for this sentence is:

■ 40.0% B 44.4% C 30.0% D 33.3%

Question 25 The main advantage of the HifiGAN over the WaveNet vocoder is:

■ the quality versus latency trade-off
B the use of softmax to directly produce discrete amplitude levels
C the use of dilated convolutions
D the use of an autoregressive model

Question 26 How many non-standard words has the sentence: “João paid 2 euros for the ice cream (a bargain!)”?

A two B three C four ■ one

Question 27 Consider chit-chat versus task-oriented dialogue systems. Which of the following statements is more characteristic of task-oriented systems:

■ Systems often involve components for interacting with external databases and services
B Modern systems typically involve the use of large language models
C The capacity to achieve natural and human-like conversations is typically the most important evaluation aspect
D Existing systems usually consider a broad conversational domain

Question 28 Which of the following modules is **not** part of the typical architecture of a modular task-oriented dialogue system:

A Speech recognition and synthesis
■ Question answering
C Natural language generation
D Dialogue state tracking

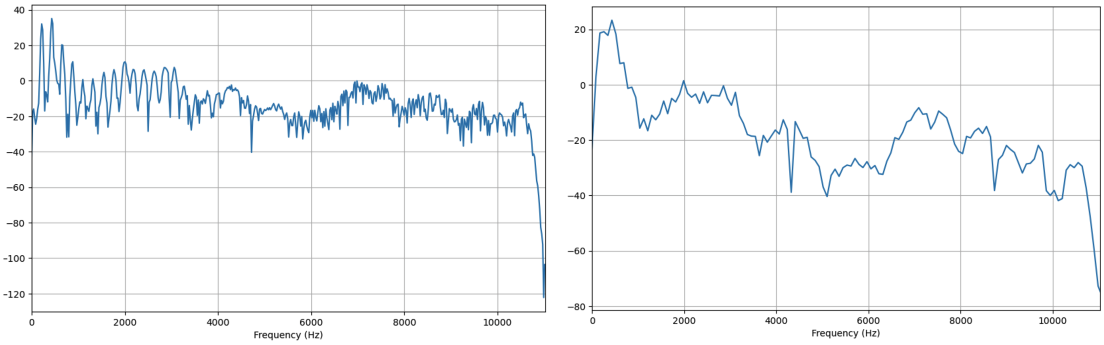
Question 29 Consider the use of learned metrics for dialogue response evaluation. Which of the following aspects is unlikely to improve relative to the use of standard metrics based on string overlaps:

A Simultaneously assess multiple quality aspects (e.g., fluency, conciseness, correctness, etc.)
B Avoid the use of ground-truth references for comparison
■ Improve correlation with human judgments of response quality
D Increase the computational performance (i.e., decrease the time to evaluate)

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

A Both metrics consider penalties for repetitions and for short-generated sentences
■ None of the metrics are directly learned from supervised examples
C Both metrics consider lexical overlaps between candidate generations and references
D None of the metrics consider semantic alignments between candidate generations and references

Question 31 The following graphs are plots of two magnitude spectrums in decibels of the same vowel sound at the same sampling frequency of 22050 Hz:



Explain the differences in the plots and if they are due to a different window type or a different window size.

Question 32 Consider the conventional processing pipeline for speech classification introduced in the lectures and, in particular, the baseline in the second lab of the course. Briefly describe this baseline, including the feature extraction, what kind of model it used, how it is trained, how it can be used for spoken language identification, and advantages/disadvantagesm with respect to other methods. You can draw a diagram to support your description.

Question 33 Consider the Sparrow system for conversational question-answering, introduced in the lectures. Briefly explain the architectural components in this system, and explain what are the similarities and main differences towards the conversational question-answering system developed in the context of Lab 3.

CORRECTED

Answer Sheet

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

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

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

QUESTION 31: 0 1 2 ■

In the left plot, the harmonic peaks and valleys are more visible than in the right one.
In the higher frequency region, the left plot has more detail, while the right keeps the same low resolution that is observed in the lower frequency region.
Finally, at the right end of the plots both of them, show a negative slope with a similar value.
From these observations, we can conclude that the first plot corresponds to a longer window that provides a higher frequency resolution. The different frequency resolutions do not allow correct identification of the width of the main lobe but the same final negative slope indicates windows of the same type given the same spectral leakage.

Question 32: 0 1 2 ■

The baseline consists of a conventional feature extraction (FE) stage based on mel-frequency cepstral coefficients (MFCCs) followed by Gaussian mixture modeling (GMMs) for each target language. After obtaining frame-level MFCCs, some additional steps can be applied, such as concatenating dynamic features (i.e., deltas, acceleration, or shifted delta cepstrum), silence frames removal and cepstral mean and variance normalization. Then, the training data of each target language is used to train a GMM for that language using a conventional EM algorithm.

For classification, the same FE pipeline is applied for a given new sentence, and the average log-likelihood given by each language model is computed. The language that obtains maximum likelihood is the identified one. Some of the limitations of this baseline for language identification include: i) the information that it handles is purely acoustic, ii) modeling is frame-based and decisions are taken based on very short segments of audio, and iii) models are generative.

Question 33: 0 1 2 ■

Sparrow is a conversational question-answering system, built by fine-tuning a large language model with reinforcement learning from human feedback.

The training of the system uses two separate reward models, trained from human feedback regarding (a) per-turn response preference, and (b) conformance with rules that correspond to desired model behavior. The system is also able to search the internet for evidence, sing text indicators within prompts that correspond to "Search Query" and "Search Result".

During training, the system learns to output the "Search Query" indicator followed by a textual search query, and then search results are obtained by retrieving and filtering a response from Google. Through the training procedure that considers the reward models to score sampled dialogue contexts from a buffer that is incrementally updated, the system is able to generate plausible and accurate answers, while avoiding the violation of pre-specified behavior rules.

The system developed in Lab 3 also corresponds to a conversational question-answering system, although much simpler. The system integrates ASR and TTS components (unlike Sparrow), but it used a prompting strategy with a much smaller language model that was not fine-tuned with reinforcement learning from human feedback with in-domain data, nor was it trained to retrieve context information from the Web.