Spoken Language Processing 2022/23

Second 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 What is the distinguishing feature of fricative consonants?

- [A] They are produced with complete closure of the vocal tract.
- ■ They are produced with turbulent airflow through a narrow constriction in the vocal tract.
- [C] They are produced with rapid vibration of the vocal folds.
- [D] They are produced with a burst of air released from a complete stop.

Question 2 What are the differences between vowel and consonant sounds in speech?

- ■ Vowel sounds are produced with a relatively open vocal tract, while consonant sounds involve some degree of constriction in the vocal tract.
- [B] Consonant sounds are longer in duration than vowel sounds.
- [C] Vowel sounds are louder than consonant sounds.
- [D] Vowel sounds are produced with vocal cord vibration while consonant sounds are not.

Question 3 Which of the following organs are included in the phonatory system?

- [A] Tongue, lips, and teeth.
- [B] Trachea, bronchi, and lungs.
- [C] Nasal cavity, pharynx, and epiglottis.
- ■ Larynx, vocal folds, and glottis.

Question 4 What is the transfer function of an LTI system?

- [A] The Fourier transform of the output signal divided by the Fourier transform of the input signal.
- ■ The z-transform of the output signal divided by the z-transform of the input signal.
- [C] The inverse z-transform of the output signal divided by the inverse z-transform of the input signal.
- [D] The inverse Fourier of the output signal divided by the inverse Fourier transform of the input signal.

Question 5 If $H(z)$ is a transfer function of an LTI system with input $X(z)$ and output $Y(z)$ and $H(z) = Y(z)/X(z) = P(z)/Q(z)$. What are the roots of the polynomial $Q(z)$ called?

- [A] Zeros of the transfer function $H(z)$.
- [B] Zeros of the output signal $Y(z)$.
- ■ Poles of the transfer function $H(z)$.
- [D] Poles of the input signal $X(z)$.

Question 6 A signal processing function produced 101 signal frames that need to be recombined using the overlap-add algorithm that was used in Lab1. If the window length is 512 samples and the hop length is 256 samples, what is the length of the reconstructed signal?

[A] 25600 [B] 25856 ■ 26112 [D] 26368

Question 7 In the source-filter model, the pulse train is used to model

- [A] the spectral envelope of speech sounds.
- ■ the filter excitation for voiced phonation.
- [C] the filter excitation for unvoiced fricatives.
- [D] the resonances in the voiced phonation.

Question 8 Which of the following is a common application of linear prediction in speech signal processing?

- [A] Extracting the fundamental frequency from a speech signal.
- [B] Removing background noise from a speech signal.
- ■ Estimating the spectral envelope of a speech signal.
- [D] Removing background noise from a speech signal.

Question 9 Two words that have the same spelling but sound differently are:

- [A] non-homographs and homophones.
- [B] homographs and homophones.
- ■ homographs and non-homophones.
- [D] non-homographs and non-homophones.

Question 10 The Griffin-Lim algorithm is used to:

- [A] estimate the original magnitude information based on the phase of the spectrum.
- [B] estimate the original speech waveform based on the complex spectrum.
- ■ estimate the original phase information based on the magnitude of the spectrum.
- [D] estimate the original speech waveform based on the complex cepstrum.

Question 11 Consider a Gaussian mixture model (GMM) of 16 mixtures with diagonal covariance matrices. The model has been trained to fit MFCC speech feature vectors of dimension 5. What is the model's total number of different parameters, including weights, means, and covariances?

■ 176 [B] 240 [C] 96 [D] 80

Question 12 Consider a 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 adding/concatenating the first and second-order delta features, a.k.a, double delta or acceleration:

■ (N, 3D) [B] (3N, D) [C] (N, 2D) [D] (2N, D)

Question 13 Consider a feature extraction frame-based method that analyses speech using windows of 40 msec with a hop size of 20 msec. If this method is applied to a speech signal of length 4 seconds, what is the resulting number of feature vectors?

- [A] It will produce 100 feature vectors.
- [B] It depends on the dimensionality of the feature extraction method.
- ■ It will produce 200 feature vectors.
- [D] It depends on the sampling rate of the speech signal.

Question 14 The last step of the conventional MFCC extraction pipeline is a discrete cosine transformation (DCT). One of the objectives of the DCT is:

- ■ to separate low/fast-varying contributions of the spectrum envelope.
- [B] to remove convolution noise.
- [C] to remove type-varying channel effects.
- [D] to convert convolutive signal contributions into additive ones.

Question 15 Feature extraction methods for speech classification can be coarsely classified according to the type of information extracted. Which one of the following **does not** correspond to one of these categories?

[A] Spectral ■ Global [C] High-level [D] Prosodic

Question 16 The GMM-UBM approach for speaker verification uses a universal background model to adapt it to the characteristics of each target speaker, in contrast to previous strategies that train a model from scratch for each speaker. This approach introduced some advantages. Which of the following **is not** an advantage of the GMM-UBM approach?

[A] It permits using larger GMM models.
[B] It permits using less enrolment/adaptation data.
■ It keeps mean parameters unchanged.
[D] It keeps correspondence among means for different speakers.

Question 17 In Hybrid HMM/DNN systems for automatic speech recognition:

[A] The language model is always a DNN-based model trained on manual transcriptions.
[B] The observation probabilities of the HMM models are provided by GMM.
[C] The decoding is performed by a DNN encoder-decoder architecture.
■ The transition probabilities of the HMM models are provided by a conventional previously trained HMM/GMM system.

Question 18 In conventional hierarchical large vocabulary continuous speech recognition systems (LVCSR), which of the following components **is not** a common module?

[A] Decoder
[B] Pronunciation model
[C] Language model
■ Speaker model

Question 19 Considering the following alignment between a text reference and the hypothesis generated by an ASR system:

```
REF: speech RECOGNITON is known as THE task of transcribing audio INTO ** text
HYP: speech IGNITION   is known as **  task of transcribing audio IN   TO text
```

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

[A] 23.1% [B] 30.8% ■ 33.3% [D] 25.0%

Question 20 CTC was proposed as a method to train an acoustic model without requiring frame-level alignments. To do so, it defines the CTC alignment concept. Considering CTC alignment and an input audio sequence of length 15 frames, which of the following CTC alignments is valid for the word parrot (consider the symbol ε as the blank character):

■ ppεaaεrrεrrεoεt
[B] ppεaεaεrrεrεoot
[C] pppppppparrrrot
[D] ppaaεrεrεrεoεtt

Question 21 Consider a speech classification model based on a Transformer encoder, with a stack of 8 multi-head self-attention modules and 4 attention heads. How many softmax operations are computed within the model in connection to the multi-head self-attention blocks, when processing an input sequence of 10 elements?

[A] 100 [B] 10 ■ 320 [D] 3200

Question 22 Consider the computations associated with the dot-product self-attention operation. Consider also an input sequence of four vectors [[2,0,0,0], [2,8,0,0], [2,0,8,0], [2,0,0,8]], 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] $[8, 8, 8, 8]$ [B] $[2, 8, 8, 8]$ ■ $[2, 2, 2, 2]$ [D] $[8, 0, 0, 0]$

Question 23 Consider the original Transformer model, proposed for sequence-to-sequence NLP tasks like machine translation. Which of the following architectural components **DOES NOT** correspond to learned parameters in the model?

[A] Projection matrices used to compute queries, keys, and value.s
[B] Feed-forward transformations after the multi-head attention operations.
[C] Input and output token embeddings.
■ Positional embeddings.

Question 24 Which of the following architectures **DOES NOT** support Automatic Speech Recognition (ASR) tasks?

[A] OpenAI Whisper
■ VALL-E
[C] SpeechT5
[D] Wav2vec and other similar encoder models, combined with a downstream text decoder

Question 25 Consider speech representation models like DiscreteBERT, pre-trained with objectives that resemble masked language modeling. Why is the pre-training of these models based on masking spans of consecutive tokens, rather than individual tokens?

[A] Make the pre-training task simpler, this way facilitating training.
[B] Facilitate the combination with contrastive loss functions.
[C] Improve computational efficiency in the computation of the loss function.
■ Avoid exploring local smoothness in nearby audio signals.

Question 26 Consider encoder-decoder versus decoder-only transformer models. Which of the following architectural components is **NOT COMMON** to both families of models?

[A] Embeddings and positional encodings.
■ Cross-attention to consider context.
[C] Masked self-attention for causal/auto-regressive masking.
[D] Multi-head self-attention operations.

Question 27 Consider the general architecture of modular task-oriented dialogue systems. Which of the following tasks is typically **NOT** considered part of the natural language understanding module?

[A] Domain identification
[B] Slot filling
[C] User intent detection
■ Dialogue state tracking

Question 28 Consider the OpenAI Whisper multitask speech model. Which of the following statements is false?

- [A] The system predicts the language being spoken through a specific output token.
- ■ The system can predict the speaker of a given utterance (from a set of speakers) through a specific output token.
- [C] The system can predict the start of a speech event through a token that encodes the time relative to the current audio segment.
- [D] The system predicts non-speech segments through a specific output token.

Question 29 Which of the following aspects corresponds to an advantage of BERTScore over BLEU?

- [A] Consider explicit penalties for very short generations.
- [B] Direct training to approximate human quality judgments for language generation.
- ■ Consider semantic comparisons instead of exact word/n-gram matches.
- [D] Avoid the need for ground-truth references.

Question 30 Consider the Sparrow system introduced in the classes. Which of the following statements is **wrong**?

- [A] The system uses a large language model to guide the interaction with an external search engine.
- ■ The system can interact with different external databases and tools.
- [C] The system can consider a broad conversational domain.
- [D] The system uses a large language model for response generation.

Question 31
Consider the first utterance of the Harvard set:

`The birch canoe slid on the smooth planks`

with the phonetic transcription:

ðə bɜːʧ kəˈnuː slɪd ɒn ðə smuːð plæŋks

Considering that a voiced region is a sequence of voiced phones, how many voice regions are in the utterance? Identify the boundaries of each voiced region.

Question 32 In the lectures, automatic speech recognition (ASR) research was described as an open scientific field that has been the focus of remarkable developments since the 50s. Briefly describe the main generations of ASR systems. Mention their main characteristics. Finally, explain what the two main alternatives in current modern ASR systems are.

Question 33 Consider the BLEU metric, as used in the labs, for evaluating automatically generated responses in dialogue systems. Discuss the main problems and limitations associated with the use of this metric.

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 ■ C D

Question 2: ■ B C D

Question 3: A B C ■

Question 4: A ■ C D

Question 5: A B ■ D

Question 6: A B ■ D

Question 7: A ■ C D

Question 8: A B ■ D

Question 9: A B ■ D

Question 10: A B ■ D

Question 11: ■ B C D

Question 12: ■ B C D

Question 13: A B ■ D

Question 14: ■ B C D

Question 15: A ■ C D

Question 16: A B ■ D

Question 17: A B C ■

Question 18: A B C ■

Question 19: A B ■ D

Question 20: ■ B C D

Question 21: A B ■ D

Question 22: A B ■ D

Question 23: A B C ■

Question 24: A ■ C D

Question 25: A B C ■

Question 26: A ■ C D

Question 27: A B C ■

Question 28: A ■ C D

Question 29: A B ■ D

Question 30: A ■ C D

Question 31: 0 1 2 3 4 ■

The phonetic transcription can be annotated in voiced (+) and unvoiced (-) phones:

`The birch canoe slid on the smooth planks`

ðə bɜːʧ kəˈnuː slɪd ɒn ðə smuːð plæŋks

`++ ++- -+++ -+++ ++ ++ -+++ -+++-`

If the word boundaries are removed:

`++++--+++-+++++++-+++-+++--`

It is now clear that there are 5 voiced regions in the utterance. The first region ends before the ʧ sound. The second starts in the ə of the word "canoe" and ends in the s of "slid". The third starts at ɪ and runs until the s of "slid". The fourth is the sequence muːð in the word "smooth". The last region is the læŋ in "planks".

QUESTION 32: 0 1 2 3 4

We can distinguish 4 main generations in ASR development. The first systems were based on heuristics and on dedicated analog systems and hardware (1G: 1950-1970). The second generation corresponds to the introduction of non-statistical pattern-matching approaches, especially, dynamic time warping (DTW) (2G: 1970-1980). The great evolution of ASR came with the 3rd generation that tackled the problem as a statistical problem and introduced the extremely influential HMM/GMM approach (3G: 1980-2010). HMM-based systems have been the standard de facto in ASR until the arrival of more recent end-to-end approaches (4G: 2010-now), with systems such as the transformer. Nowadays, attention-encoder-decoder-based approaches dominate the field, however, in limited-resourced scenarios, HMM-based systems (using DNNs as the acoustic model) can still provide very competitive performances.

QUESTION 33: 0 1 2 3 4

The BLEU metric is based on comparisons against ground-truth references and, in the context of dialogue, this is a serious limitation because there are usually many acceptable responses to an input context, while only a few of these possibilities are collected as references. As a consequence, the BLEU metric is known to correlate poorly with human judgments of dialogue response quality, and recent developments have proposed the use of learned metrics that avoid the use of ground-truth references. Another problem relates to the fact that BLEU uses word/n-gram overlaps to assess similarity against the ground-truth references, which may be misleading due to the possibility of phrasing the same semantic content in different ways. Metrics based on assessing similarity through (contextual) word embeddings can offer an alternative that avoids this issue.