

Student number (6 digits):

Spoken Language Processing 2022/23

Second Exam

Duration: 90 minutes.

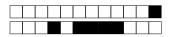
First and last name:

All n For q Othe	nultiple-ch questions 1 r incorrect	to 30, each co answers, mon	s have exactorrect answere than one	ctly one correver is worth 0 answer and	ect answer5 point. Very incoquestions left una	orrect answers are worth are worth 1.67 po	vorth -0.25 points 0 points.
Question	n 1 Wh	at is the distin	nguishing f	eature of fric	ative consonants?		
A Th	ney are pro	duced with co	omplete clo	osure of the v	ocal tract.		
B Th	ney are pro	duced with tu	rbulent air	flow through	a narrow constric	tion in the vocal tra	ict.
C Th	ney are pro	duced with ra	pid vibrati	on of the voc	al folds.		
D Th	ney are pro	duced with a	burst of air	r released fro	m a complete stop).	
Question	n 2 Wh	at are the diffe	erences be	tween vowel	and consonant sou	unds in speech?	
		ls are produce nstriction in th			n vocal tract, while	e consonant sounds	involve some
B Co	onsonant s	ounds are long	ger in dura	tion than vov	vel sounds.		
C Vo	wel sound	ls are louder t	han consor	nant sounds.			
D Vo	wel sound	ls are produce	d with voc	al cord vibra	tion while conson	ant sounds are not.	
Question	n 3 Wh	ich of the foll	owing orga	ans are includ	led in the phonato	ory system?	
A To	ongue, lips,	, and teeth.			C Nasal cavity,	, pharynx, and epigi	lottis.
B Tr	achea, bro	nchi, and lung	gs.		D Larynx, voca	al folds, and glottis.	
Question	n 4 Wh	at is the trans	fer function	n of an LTI s	ystem?		
A Th	ne Fourier	transform of t	he output	signal divideo	d by the Fourier tr	ansform of the inpu	ıt signal.
B Th	ne z-transfo	orm of the out	tput signal	divided by th	ne z-transform of t	the input signal.	
C Th	ne inverse	z-transform of	f the outpu	t signal divid	ed by the inverse	z-transform of the i	nput signal.
D Th	ne inverse l	Fourier of the	output sig	nal divided b	y the inverse Four	rier transform of the	input signal.
					system with input omial $Q(z)$ called?	X(z) and output $Y(z)$	z) and $H(z) =$
A Ze	eros of the	transfer funct	ion $H(z)$.		C Poles of the	transfer function H	(z).
_		output signal				input signal $X(z)$.	
	ap-add alg		as used in	Lab1. If the	window length is 3	that need to be reco	_
		A 25600	В	25856	C 26112	D 26368	

		+1/2/59+
Question 7 In the source-filter	r model, the pulse tra	in is used to model
A the spectral envelope of spe B the filter excitation for voice		C the filter excitation for unvoiced fricatives.D the resonances in the voiced phonation.
Question 8 Which of the followersing?	owing is a common	application of linear prediction in speech signal pr
A Extracting the fundamental	frequency from a sp	beech signal.
B Removing background noise	se from a speech sign	nal.
C Estimating the spectral env	elope of a speech sig	gnal.
D Removing background nois	se from a speech sign	nal.
Question 9 Two words that ha	we the same spelling	but sound differently are:
A non-homographs and homo	ophones.	C homographs and non-homophones.
B homographs and homophor	nes.	D non-homographs and non-homophones.
Question 10 The Griffin-Lim	algorithm is used to:	
A estimate the original magni	itude information bas	sed on the phase of the spectrum.
B estimate the original speech		
		n the magnitude of the spectrum.
D estimate the original speech		
		el (GMM) of 16 mixtures with diagonal covarian
	parameters, includir	speech feature vectors of dimension 5. What is to ge weights, means, and covariances? C 96 D 80
model's total number of different A 176 Question 12 Consider a num shape (N, D). What is the size	parameters, including by array formed by of the array after ac	ng weights, means, and covariances?
model's total number of different A 176 Question 12 Consider a num	parameters, including by array formed by of the array after ac	reg weights, means, and covariances? C 96 D 80 N feature rows each with dimension D, that, is with
model's total number of different A 176 Question 12 Consider a num shape (N, D). What is the size features, a.k.a, double delta or accommodation [A] (N, 3D) Question 13 Consider a features	parameters, including the parameters, including the parameters, including the parameters of the array after acceleration: B (3N, D) The extraction frame-lisec. If this method is vectors? The vectors of the feature vectors.	C 96 D 80 N feature rows each with dimension D, that, is wilding/concatenating the first and second-order delection (N, 2D) D (2N, D) December 20 D (2N, D) December 30 D (2N, D) December 40 D (2N, D) D
Question 12 Consider a numshape (N, D). What is the size features, a.k.a, double delta or accordance (N, 3D) Question 13 Consider a feature 40 msec with a hop size of 20 ms is the resulting number of feature A It will produce 100 feature B It depends on the dimension C It will produce 200 feature D It depends on the sampling	parameters, including the parameters, including the array formed by of the array after acceleration: B (3N, D) The extraction frame-like the conventional MFC and the speech signs the conventional MFC array of the speech signs the conventional MFC array of the speech signs the conventional MFC array of the speech signs the speech speech signs the speech signs the speech signs the speech signs th	C 96 D 80 N feature rows each with dimension D, that, is wilding/concatenating the first and second-order decomposed method that analyses speech using windows applied to a speech signal of length 4 seconds, where the second method.
Question 12 Consider a numshape (N, D). What is the size features, a.k.a, double delta or accordance (N, 3D) Question 13 Consider a feature 40 msec with a hop size of 20 ms is the resulting number of feature A It will produce 100 feature B It depends on the dimension C It will produce 200 feature D It depends on the sampling Question 14 The last step of the contraction (DCT). One of the objectives	parameters, including the parameters, including the parameters, including the parameters of the array after acceleration: B (3N, D) The extraction frame-level of the second the feature of the speech single conventional MFC and the parameters of the DCT is:	© 96 D 80 N feature rows each with dimension D, that, is widding/concatenating the first and second-order delayed (N, 2D) D (2N, D) Deased method that analyses speech using windows applied to a speech signal of length 4 seconds, where the seconds with the seconds of the second of the seconds of the second of th
Question 12 Consider a numshape (N, D). What is the size features, a.k.a, double delta or accordance (N, 3D) Question 13 Consider a feature 40 msec with a hop size of 20 ms is the resulting number of feature A It will produce 100 feature B It depends on the dimension C It will produce 200 feature D It depends on the sampling Question 14 The last step of the tion (DCT). One of the objectives A to separate low/fast-varying	parameters, including the parameters, including the array formed by of the array after acceleration: B (3N, D) The extraction framelese. If this method is vectors? The vectors are of the feature frame of the speech single conventional MFC is of the DCT is: The contributions of the grant of the speech single contributions of the speech spee	© 96 D 80 N feature rows each with dimension D, that, is wilding/concatenating the first and second-order decomposition (N, 2D) D (2N, D) Decomposition of the property of t
Question 12 Consider a numshape (N, D). What is the size features, a.k.a, double delta or accordance (N, 3D) Question 13 Consider a feature 40 msec with a hop size of 20 ms is the resulting number of feature A It will produce 100 feature B It depends on the dimension C It will produce 200 feature D It depends on the sampling Question 14 The last step of the tion (DCT). One of the objectives A to separate low/fast-varying	parameters, including the parameters, including the parameters and the parameters are considered by the array after and collection: B (3N, D) The extraction frame-level of the second frame-level of the feature and the conventional MFC of the DCT is: The conventional MFC of the DCT is: The contributions of the second frame-level of the speech signs and the parameters are conventional MFC of the DCT is:	© 96 D 80 N feature rows each with dimension D, that, is widding/concatenating the first and second-order delayed (N, 2D) D (2N, D) Deased method that analyses speech using windows applied to a speech signal of length 4 seconds, where the seconds with the seconds of the second of the seconds of the second of th

						+1/	3/58	i+
				_				ely classified according espond to one of thes
A	Spectral	В	ilobal	C	High-level		D	Prosodic
adapt it to the chara	cteristics of h speaker. T	each targ	get spea	ker, in co	ontrast to pre	vious s	trate	ll background model t gies that train a mod of the following is no
A It permits usin	g larger GM	IM model	s.					
B It permits usin	g less enrol	ment/adap	otation d	ata.				
C It keeps mean	•	•						
D It keeps corres	spondence a	mong mea	ans for d	lifferent s	peakers.			
Question 17 In H	Iybrid HMM	I/DNN sy	stems fo	or automa	itic speech red	cognitio	on:	
A The language	model is alv	vays a DN	IN-base	d model t	rained on ma	nual tra	nscr	iptions.
B The observation	on probabili	ties of the	HMM 1	nodels ar	e provided by	y GMM	I.	
C The decoding	is performe	d by a DN	N enco	der-decod	ler architectu	re.		
D The transition HMM/GMM		es of the I	HMM m	odels are	provided by	a conv	enti	onal previously traine
Question 18 In (LVCSR), which of				ot a com		?	peec	h recognition system
B Pronunciation	model			_	Speaker mo			
		following	alianm				nd th	ne hypothesis generate
by an ASR system: REF: speech RECC HYP: speech IGNI The word error rate	OGNITON is	s known s known	as THE as **	task o		bing a	audi	o INTO ** text
A	23.1%	В	30.8%	C	33.3%	D] 25	5.0%
alignments. To do s	so, it defines ength 15 fra	s the CTC mes, which	Calignm ch of the	ent conc	ept. Conside	ring C	TC a	it requiring frame-leve dignment and an inpu- lid for the word parro
A ppeaaerreree				D	ppppppppa ppaaererere			
В рреававтвево								
Question 21 Con 8 multi-head self-att	tention mod	ules and	4 attenti	on heads	. How many	softma	ax op	ncoder, with a stack operations are compute ssing an input sequence

•			+1	/4/57+
queries, keys, a [0,0,0,1]] (i.e., self-attention o	put sequence of four vect and values are all comput diagonal 4x4 matrices wi	tors [[2,0,0,0], [2,3] ed through the proof the same values) ent in the sequence	8,0,0], [2,0,8,0], jection matrix [). What would be e? Recall that the	self-attention operation. Con- [2,0,0,8]], and consider that [1,0,0,0], [0,1,0,0], [0,0,1,0], the result of the dot-product e softmax operation returns a in all dimensions.
A	[8, 8, 8, 8] B [2	2, 8, 8, 8] C	[2, 2, 2, 2]	D [8,0,0,0]
				uence-to-sequence NLP tasks s DOES NOT correspond to
A Projection	n matrices used to comput	e queries, keys, and	d value.s	
B Feed-forv	ward transformations after	the multi-head atte	ention operations.	
C Input and	output token embeddings	s.		
D Positiona	l embeddings.			
Question 24 (ASR) tasks?	Which of the following a	rchitectures DOES	S NOT support A	utomatic Speech Recognition
A OpenAI	Whisper			
B VALL-E				
C SpeechT:	5			
D Wav2vec	and other similar encoder	models, combined	with a downstre	am text decoder
		hy is the pre-trainir		re-trained with objectives that ls based on masking spans of
A Make the	pre-training task simpler,	this way facilitatin	g training.	
	the combination with con			
C Improve	computational efficiency is	n the computation	of the loss function	on.
D Avoid ex	ploring local smoothness i	n nearby audio sign	nals.	
Question 26 architectural co	Consider encoder-decode mponents is NOT COMN			odels. Which of the following
A Embeddi	ngs and positional encodir	igs.		
	ention to consider context.			
	elf-attention for causal/au		ing.	
	ad self-attention operation	_	J	
	•	nitecture of modular		alogue systems. Which of the erstanding module?
Question 27 following tasks	is typically NOT consider	red part of the natu		C
following tasks	is typically NOT consider	_		_
following tasks	is typically NOT consider dentification	C	User intent detection Dialogue state tr	etion



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.
- B The system can predict the speaker of a given utterance (from a set of speakers) through a specific output token.
- 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.
- C 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.
- B The system can interact with different external databases and tools.
- The system can consider a broad conversational domain.
- D The system uses a large language model for response generation.

Ouestion 31

Consider the first utterance of the Harvard set:

The birch canoe slid on the smooth planks

with the phonetic transcription:

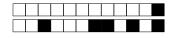
ða batt ka nu slid on ða smuð plænks

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.

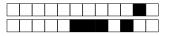
Question 32 In the lectures, automatic speech recognition (ASR) research was described as an open scientific field that has been the focus of remarkable developments since the 50s. Briefly describe the main generations of ASR systems. Mention their main characteristics. Finally, explain what the two 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	Answer Sheet
	Answers must be given exclusively on this sheet. Answers given on
2 2 2 2 2 2	other sheets will be ignored.
3 3 3 3 3 3	No corrections are allowed on this sheet.
4 4 4 4 4 4	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.
5 5 5 5 5	ich, starting with o ir it has just 3 digns, and write your name octow.
6 6 6 6 6	First and last name:
7 7 7 7 7 7	
8 8 8 8 8	
9 9 9 9 9 9	
QUESTION 1: A B C	D QUESTION 16: A B C D
Question 2: A B C	D QUESTION 17: A B C D
QUESTION 3: A B C	D QUESTION 18: A B C D
QUESTION 4: A B C	D QUESTION 19: A B C D
QUESTION 5: A B C	D QUESTION 20: A B C D
QUESTION 6: A B C	D Question 21: A B C D
QUESTION 7: A B C	D QUESTION 22: A B C D
QUESTION 8: A B C	D QUESTION 23: A B C D
QUESTION 9: A B C	D QUESTION 24: A B C D
Question 10: A B C	QUESTION 25: A B C D
Question 11: A B C	Question 26: A B C D
Question 12: A B C	QUESTION 27: A B C D
Question 13: A B C	QUESTION 28: A B C D
QUESTION 14: A B C	
Question 15: A B C	QUESTION 30: A B C D
Question 31:	0 1 2 3 4 5
I	



Question 32:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5



Student number (6 digits):

Spoken Language Processing 2022/23

Second Exam Duration: 90 minutes.

First and last name:

	All multiple-choice question For questions 1 to 30, each of Other incorrect answers, mo	ns have exactorized answore than one	the answer sheet. Answers given on other sheets will be ignored. ctly one correct answer. ver is worth 0.5 point. Very incorrect answers are worth -0.25 points answer and questions left unanswered are worth 0 points. each. Open questions 32 and 33 are worth 1.67 points.
Que	estion 1 What are the diff	ferences bet	ween vowel and consonant sounds in speech?
A	Nowel sounds are produc	ed with voc	al cord vibration while consonant sounds are not.
E	Wowel sounds are product degree of constriction in		elatively open vocal tract, while consonant sounds involve some act.
C	Vowel sounds are louder	than conson	ant sounds.
Γ	O Consonant sounds are lor	iger in durat	tion than vowel sounds.
Que	estion 2 What is the disti	nguishing fo	eature of fricative consonants?
A	They are produced with o	omplete clo	osure of the vocal tract.
E	They are produced with t	urbulent air	flow through a narrow constriction in the vocal tract.
C	They are produced with r	apid vibrati	on of the vocal folds.
	They are produced with a	burst of air	released from a complete stop.
Que	estion 3 Which of the fol	lowing orga	ans are included in the phonatory system?
A	Tongue, lips, and teeth.		C Nasal cavity, pharynx, and epiglottis.
E	B Larynx, vocal folds, and	glottis.	D Trachea, bronchi, and lungs.
_			n of an LTI system with input $X(z)$ and output $Y(z)$ and $H(z)$ = of the polynomial $Q(z)$ called?
A	Poles of the input signal 2	X(z).	$\boxed{\mathbf{C}}$ Zeros of the output signal $Y(z)$.
=	Zeros of the transfer func		$\boxed{\mathbf{D}}$ Poles of the transfer function $H(z)$.
Que	estion 5 What is the trans	sfer function	n of an LTI system?
A	The Fourier transform of	the output s	signal divided by the Fourier transform of the input signal.
=	_		divided by the z-transform of the input signal.
	_		nal divided by the inverse Fourier transform of the input signal.
1	D The inverse z-transform of	of the output	t signal divided by the inverse z-transform of the input signal.
Que	estion 6 In the source-file	er model, th	ne pulse train is used to model
Α	the resonances in the voice	ed phonatic	on. C the filter excitation for voiced phonation.
E	the spectral envelope of s	peech sound	ds. D the filter excitation for unvoiced fricatives.

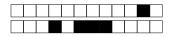
				+2/2/51+
the overlap-ado		s used in Lab1. If th	ne window length is	that need to be recombined us. 512 samples and the hop length
	A 25856	B 25600	C 26368	D 26112
Question 8 cessing?	Which of the follo	wing is a common	application of linea	ur prediction in speech signal p
A Extractin	ng the fundamental	frequency from a sp	peech signal.	
B Removir	ng background noise	e from a speech sign	nal.	
C Removir	ng background noise	e from a speech sign	nal.	
D Estimation	ng the spectral enve	lope of a speech sig	gnal.	
Question 9	Two words that have	ve the same spelling	g but sound differen	tly are:
A homogra	aphs and non-homo	phones.	C homographs	and homophones.
B non-hom	nographs and homo	phones.	D non-homogr	raphs and non-homophones.
Question 10	The Griffin-Lim a	lgorithm is used to:	:	
A estimate	the original speech	waveform based or	n the complex cepst	rum.
			on the magnitude of	
			sed on the phase of	
			n the complex spect	•
Question 11 to the type of categories?		-		n be coarsely classified accord s not correspond to one of th
[.	A High-level	B Prosodic	C Spectral	D Global
Question 12	-		CC extraction pipeli	ne is a discrete cosine transform
tion (DCT). Or	ne of the objectives	of the DCT is:		
	ve type-varying char			
A to remov	ve type-varying char	nnel effects.	e spectrum envelope	e.
A to remov B to separa	ve type-varying char	nnel effects.	e spectrum envelope	e.
A to remov B to separa C to remov	ve type-varying characte low/fast-varying	nnel effects. contributions of the		2 .
A to remov B to separa C to remov D to conver Question 13 shape (N, D).	ve type-varying char ate low/fast-varying ve convolution noise rt convolutive signa Consider a nump	nnel effects. contributions of the c. l contributions into by array formed by f the array after ac	additive ones. N feature rows eac	h with dimension D, that, is w
A to remov B to separa C to remov D to conver Question 13 shape (N, D).	ve type-varying char ate low/fast-varying ve convolution noise rt convolutive signa Consider a nump What is the size o	nnel effects. contributions of the c. l contributions into by array formed by f the array after ac	additive ones. N feature rows eac	h with dimension D, that, is we the first and second-order do
A to remov B to separa C to remov D to conver Question 13 shape (N, D). features, a.k.a, Question 14 40 msec with a	ve type-varying character low/fast-varying ve convolution noise rt convolutive signa Consider a nump What is the size of double delta or acce A (2N, D) Consider a feature	contributions of the contributions of the contributions into by array formed by afthe array after acceleration: B (3N, D) e extraction frame-lec. If this method is	additive ones. N feature rows eac dding/concatenating (N, 3D) based method that a	h with dimension D, that, is we the first and second-order do
A to remove B to separa C to remove D to conver Question 13 shape (N, D). features, a.k.a, Question 14 40 msec with a is the resulting	ve type-varying character low/fast-varying ve convolution noise rt convolutive signa Consider a nump What is the size of double delta or acc A (2N, D) Consider a feature a hop size of 20 mse	contributions of the contributions of the contributions into a contribution in contributions in contr	additive ones. N feature rows eac dding/concatenating (N, 3D) based method that as applied to a speech	h with dimension D, that, is we the first and second-order do D (N, 2D) unalyses speech using windows
A to remove B to separa C to remove D to convert Question 13 shape (N, D). features, a.k.a, Question 14 40 msec with a is the resulting	ve type-varying character low/fast-varying ve convolution noise rt convolutive signa Consider a nump What is the size of double delta or acc A (2N, D) Consider a feature a hop size of 20 mse number of feature versus and the size of feature of the size of	contributions of the contributions of the contributions into by array formed by a frag after acceleration: B (3N, D) e extraction frame-ec. If this method is vectors?	additive ones. N feature rows eac dding/concatenating (N, 3D) based method that as applied to a speech	h with dimension D, that, is we the first and second-order do D (N, 2D) unalyses speech using windows
A to remove B to separa C to remove D to conver Question 13 shape (N, D). features, a.k.a, Question 14 40 msec with a is the resulting A It depend B It will pr	ve type-varying charate low/fast-varying ve convolution noise rt convolutive signa Consider a nump What is the size of double delta or acc A (2N, D) Consider a feature a hop size of 20 mse number of feature versions.	nnel effects. contributions of the contributions of the contributions into by array formed by afthe array after acceleration: B (3N, D) e extraction frame-lec. If this method is vectors? nality of the feature vectors.	additive ones. N feature rows eac dding/concatenating (N, 3D) based method that a sapplied to a speech	h with dimension D, that, is we the first and second-order do D (N, 2D) unalyses speech using windows

				+2/3/50+	
matrices. The mo	del has been train	ned to fit MFCC s	speech feature	6 mixtures with diagonal covectors of dimension 5. Whens, and covariances?	
	A 80	B 240	C 176	D 96	
adapt it to the cha	racteristics of each speaker. This	h target speaker, approach introduc	in contrast to 1	uses a universal background previous strategies that train stages. Which of the followi	a model
B It permits us C It keeps mea	ing larger GMM ing less enrolment parameters unclesspondence amon	t/adaptation data. nanged.	ent speakers.		
Question 17 Co	onsidering the follo	owing alignment b	oetween a text r	eference and the hypothesis g	generated
by an ASR system	: COGNITON is kn NITION is kn	lown as THE ta lown as ** ta	sk of transc	ribing audio INTO **	
	A 33.3%	B 23.1%	C 30.8%	D 25.0%	
B The observa C The transition HMM/GMM	tion probabilities of	of the HMM model	els are provided ls are provided	by a conventional previousl	y trainec
Question 19 C alignments. To do	TC was proposed o so, it defines the length 15 frames.	as a method to tree CTC alignment, which of the fol	ain an acoustic concept. Cons	model without requiring fra idering CTC alignment and ignments is valid for the wo	an inpu
А рреаеаеттега	eoot		C ppppppp	parrrrot	
B ppaaererere	oett		D рреаает	rerreoet	
Question 20 It (LVCSR), which o				itinuous speech recognition ile?	system
A Speaker mod	del		C Decoder		
B Pronunciation	on model		D Languag	e model	
	anguage modeling	g. Why is the pre-		BERT, pre-trained with object se models based on masking	
Avoid explo	ring local smooth	ness in nearby aud	lio signals.		
	e combination with				
	nputational efficie	•			
D Make the pr	e-training task sim	ipier, this way fac	ilitating training	ζ.	

	+2/4/49+
8 multi-head self-attention modules and 4 attention	odel based on a Transformer encoder, with a stack of heads. How many softmax operations are computed f-attention blocks, when processing an input sequence
A 100 B 3200	C 320 D 10
sider also an input sequence of four vectors [[2,0,0 queries, keys, and values are all computed through [0,0,0,1]] (i.e., diagonal 4x4 matrices with the same	ed with the dot-product self-attention operation. Con- 0,0], [2,8,0,0], [2,0,8,0], [2,0,0,8]], and consider that the projection matrix [[1,0,0,0], [0,1,0,0], [0,0,1,0], values). What would be the result of the dot-product equence? Recall that the softmax operation returns a ors have the same values in all dimensions.
A [2, 8, 8, 8] B [2, 2, 2, 2]	[C] [8,8,8,8] [D] [8,0,0,0]
Question 24 Which of the following architectures (ASR) tasks?	S DOES NOT support Automatic Speech Recognition
like machine translation. Which of the following a learned parameters in the model? A Projection matrices used to compute queries, k B Input and output token embeddings. C Feed-forward transformations after the multi-h D Positional embeddings.	model, proposed for sequence-to-sequence NLP tasks architectural components DOES NOT correspond to eys, and value.s ead attention operations.
C Cross-attention to consider context.	
D Masked self-attention for causal/auto-regressiv	e masking.
Question 27 Which of the following aspects corre	esponds to an advantage of BERTScore over BLEU?
A Consider semantic comparisons instead of exact	ct word/n-gram matches.
B Avoid the need for ground-truth references.	
C Direct training to approximate human quality j	
D Consider explicit penalties for very short gener	rations.
Question 28 Consider the general architecture of following tasks is typically NOT considered part of t	modular task-oriented dialogue systems. Which of the he natural language understanding module?
A Domain identification	C Slot filling

B User intent detection

Dialogue state tracking



Question 29 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.
- B The system can predict the start of a speech event through a token that encodes the time relative to the current audio segment.
- C The system predicts non-speech segments through a specific output token.
- D The system can predict the speaker of a given utterance (from a set of speakers) through a specific output token.

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

Ouestion 31

Consider the first utterance of the Harvard set:

The birch canoe slid on the smooth planks

with the phonetic transcription:

ða battf ka'nut slid on ða smutð plænks

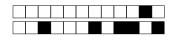
Considering that a voiced region is a sequence of voiced phones, how many voice regions are in the utterance? Identify the boundaries of each voiced region.

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.

+2/6/47+

Student number (6 digits) 0 0 0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 5 5 5 6 6 6 6 6 6 6 7 7 7 7 7 7 7 8 8 8 8 8 8 8 9 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 QUESTION 2: A B C QUESTION 3: A B C QUESTION 4: A B C QUESTION 5: A B C QUESTION 6: A B C QUESTION 7: A B C QUESTION 8: A B C QUESTION 9: A B C QUESTION 10: A B C QUESTION 11: A B C QUESTION 12: A B C QUESTION 13: A B C QUESTION 14: A B C QUESTION 14: A B C	D Question 27: A B C D D Question 28: A B C D D Question 29: A B C D
QUESTION 31:	0 1 2 3 4 5



Question 32:	0 1 2 3 4 5
Question 33:	0 1 2 3 4 5
Question 33:	0 1 2 3 4 5
Question 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
Question 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5
QUESTION 33:	0 1 2 3 4 5