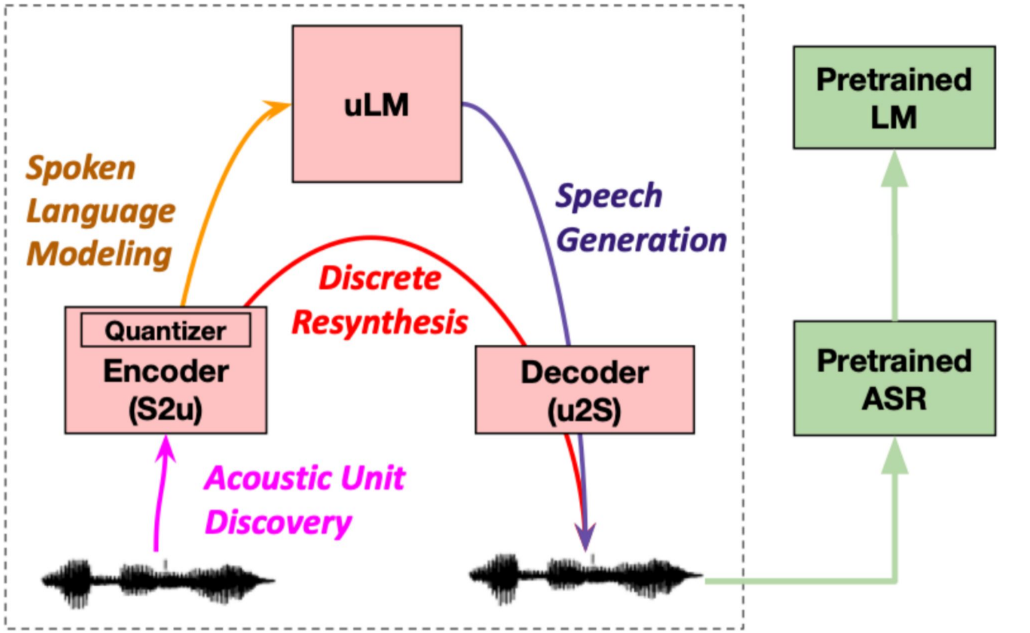
Generative Spoken Language Modeling from Raw Audio

Докладчик: Дмитрий Кириллов
Хакер: Петр Молодык

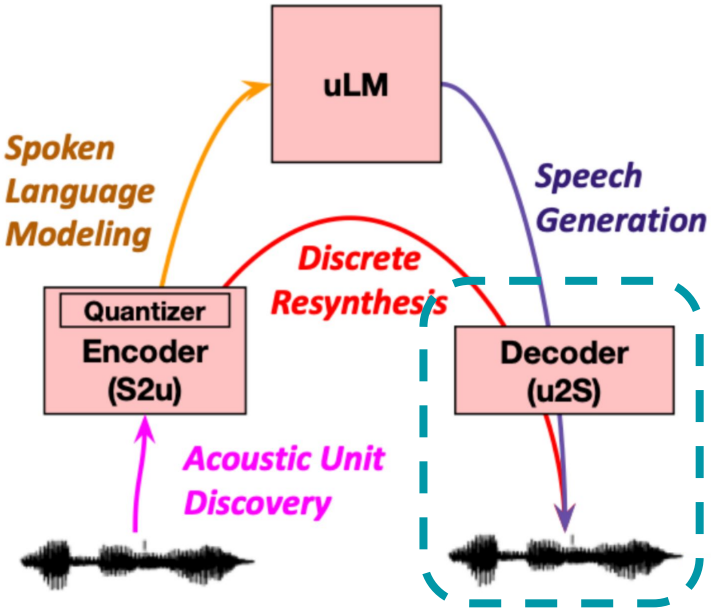
План

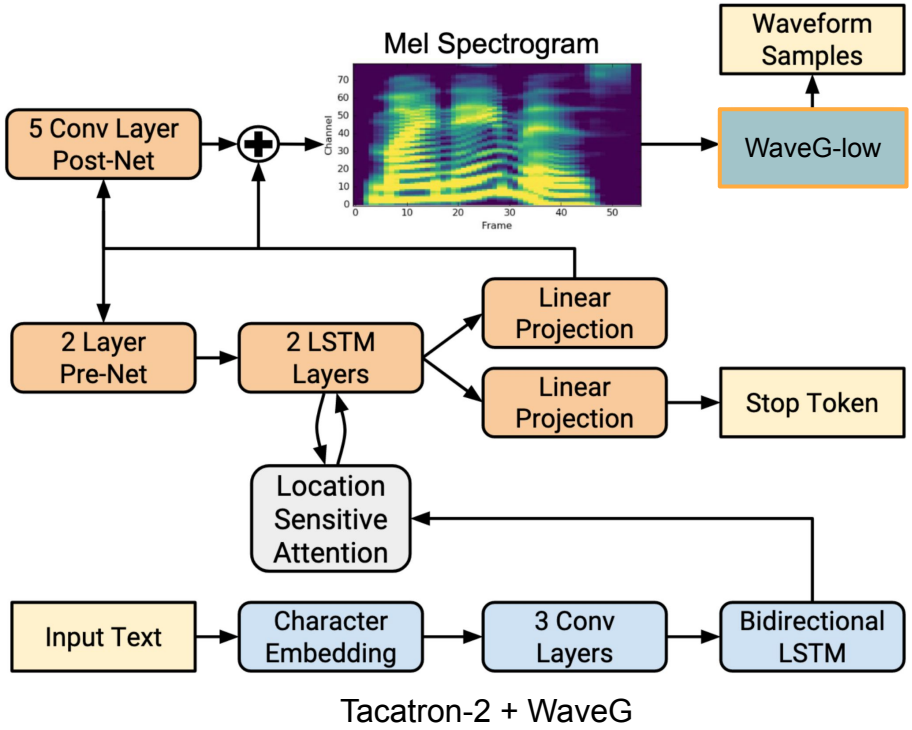
1. Какую задачу решаем
2. Как измеряем качество
3. Какие результаты

Общий сетап


uLM
Spoken Language Modeling
Speech Generation
Discrete Resynthesis
Quantizer
Encoder (S2u)
Decoder (u2S)
Acoustic Unit Discovery
Pretrained LM
Pretrained ASR
Model architecture and tasks
ASR evaluation

Unit2speech

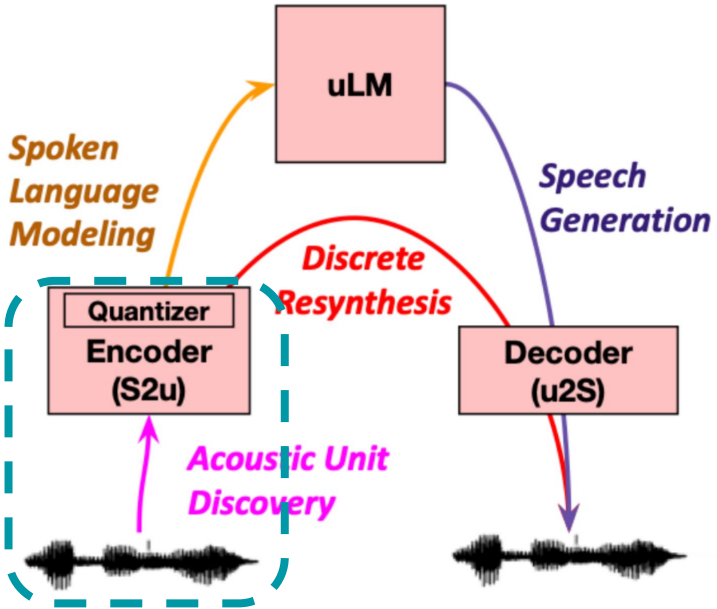


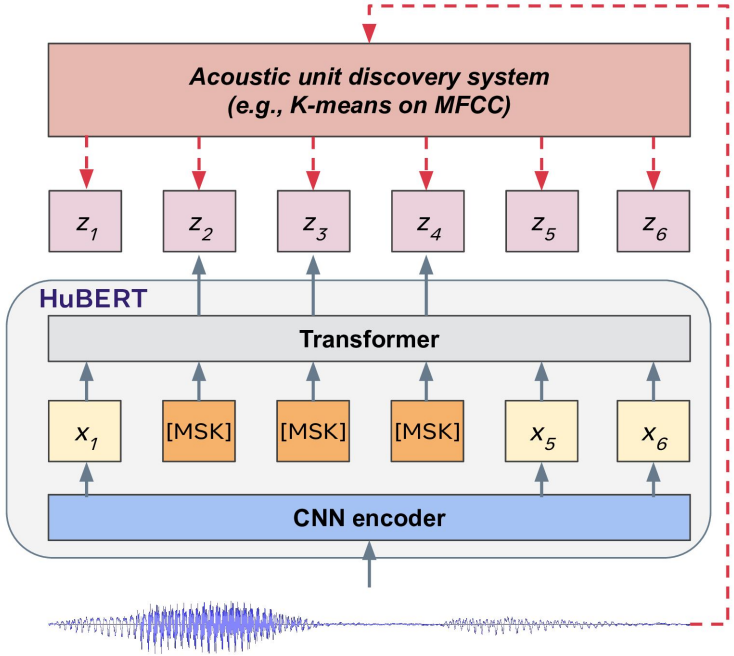


Tacatron-2 + WaveG

[HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units]

Speech2unit





[HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units]

План

1. Какую задачу решаем
2. Как измеряем качество
3. Какие результаты

Ручные метрики

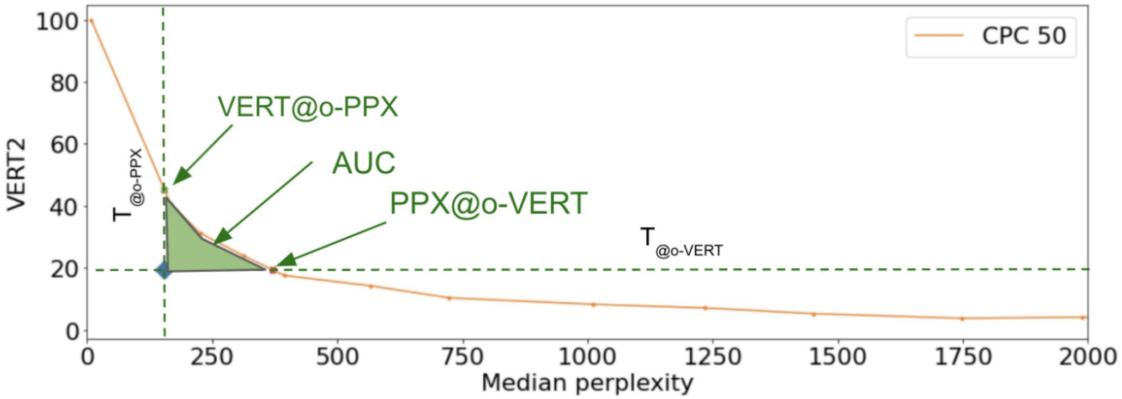
1. Mean Opinion Scores (MOS) – асессоров просили оценить **понятность** генерируемого аудио
2. CER на ручной разметке
3. meaningfulness-MOS (MMOS) – асессоров просили оценить **естественность** генерируемого аудио (для этой метрики подбиралась температура семплирования на тесте)

Автоматические на основе ASR

1. PER-from-ASR, CER-from-ASR – ошибка в фонемах(символах) после распознавания сгенерированного аудио предобученным ASR
2. PPX – перплексия текста после ASR
3. self-BLEU – BLEU между разными сгенерированными предложениями. Чем больше значение, тем меньше разнообразие
4. auto-BLEU – доля n-грамм в предложении, которые повторились хотя бы k раз
5. $\mathrm{VERT} = \sqrt{\mathrm{autoBLEU} \times \mathrm{selfBLEU}}$

AUC-of-VERT/PPX

1. Посчитать PPX и VERT реального текста
2. Найти температуры для которых после ASR получаются такие же значения PPX/VERT



3. Посчитать площадь под кривой VERT-PPX между этими границами
4. Чем AUC-of-VERT/PPX меньше, тем ближе модель к реальной речи

Zero-shot метрики

Для токенов ***x*** и ***a***, принадлежащих категории ***A*** и токена ***b*** из другой категории ***B***, ABX – вероятность того, что ***x*** ближе к **a**, чем к ***b***

Вероятность оценивается **S2u** моделью для звука

ABX-within

Категории – слова из трех букв, отличающихся только центральной

A = *beg*, ***B*** = *bag*, ***a*** – произношение *beg*, ***x*** – другое произношение, ***b*** – произношение *bag*

ABX на ZeroSpeech 2017

Zero-shot метрики

Для токенов ***x*** и ***a***, принадлежащих категории ***A*** и токена ***b*** из другой категории ***B***, ABX – вероятность того, что ***x*** ближе к **a**, чем к ***b***

Вероятность оценивается **S2u** моделью для звука

ABX-across

Категории – автор, произносящий текст

a – *beg*, сказанное *первым* автором, ***x*** – *bag*, сказанное *первым* автором, ***b*** – *beg*, сказанное *вторым* автором

ABX на ZeroSpeech 2017

Zero-shot метрики

spot-the-word accuracy – доля верно различенных пар настоящее-ненастоящее слово

Например *p(sound_of(**brick**)) > p(sound_of(**blick**))*

Вероятности беруться для последовательности на выходе после S2u→uLM

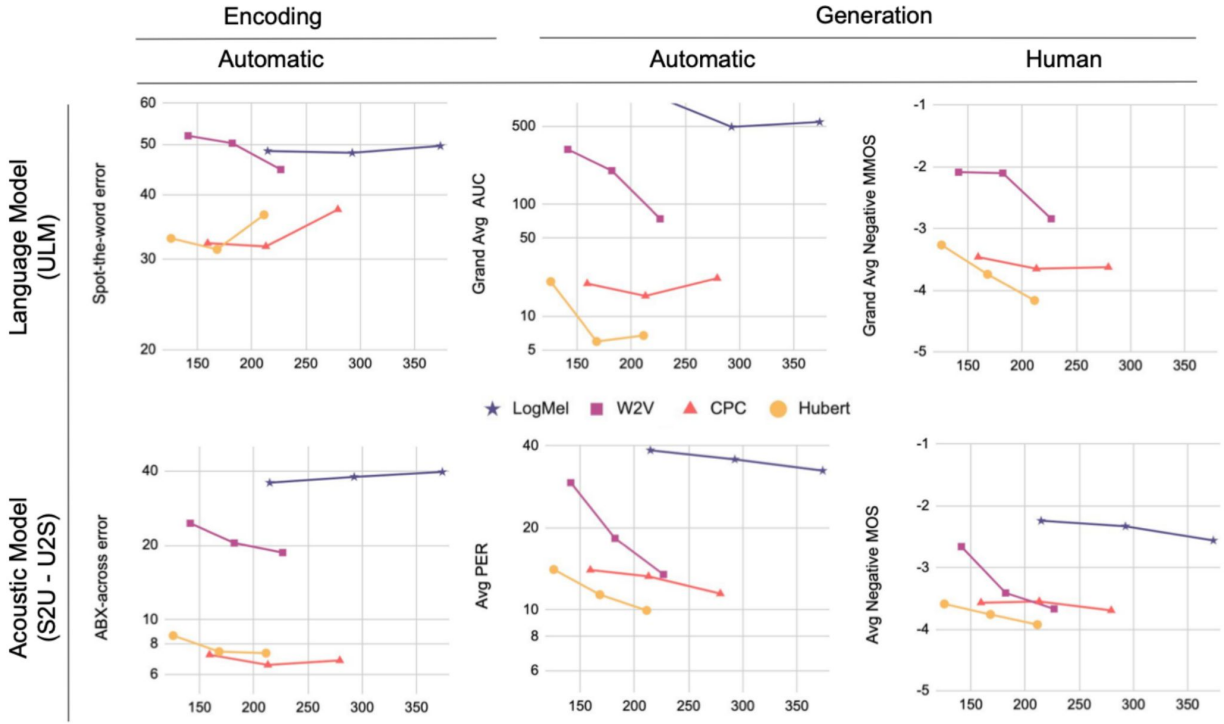
Все метрики вместе

	Кодирование речи		Генерация		
Уровень	Задача	Автоматическая метрика	Задача	Автоматическая метрика	Ручная метрика
Текстовый	Spoken LM	**Spot-the-word**	Speech Generation	**AUC-of-VERT/PPX**, BLEU, PPX@o-VERT	**MMOS**
Звуковой	Acoustic Unit	**ABX-across**, ABX-within	Resynthesis	**PER-from-ASR**, CER-from-ASR	CER, **MOS**

План

1. Какую задачу решаем
2. Как измеряем качество
3. Какие результаты

Эксперименты



LogMel – kMeans поверх Mel спектрограмм *LibriSpeech clean-100h*

W2V – Wav2vec 2.0

Корреляция с ручной разметкой

	Zero-shot			ASR-based				Human			
	ABX within	ABX across	spot-the word	avg PER	avg CER	AUC uncond	AUC prompted	avg CER	avg MOS	MMOS uncond	MMOS prompted
ABX within				0.904	0.896	0.893	0.806	0.901	0.883	0.935	**0.881**
ABX across	0.970			**0.944**	**0.938**	**0.962**	**0.910**	**0.905**	**0.924**	**0.941**	**0.881**
spot-the-word	0.937	0.853		0.767	0.760	0.753	0.639	0.806	0.743	0.902	0.808

Выводы

- Авторы предложили набор метрик для оценки self-supervised моделирования устной речи
- Измерили качество для нескольких SotA Speech2unit моделей
- Показали скоррелированность предложенных автоматических метрик с ручной разметкой

Источники

Demo

Wav2vec 2.0

HuBERT

CPC

The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling

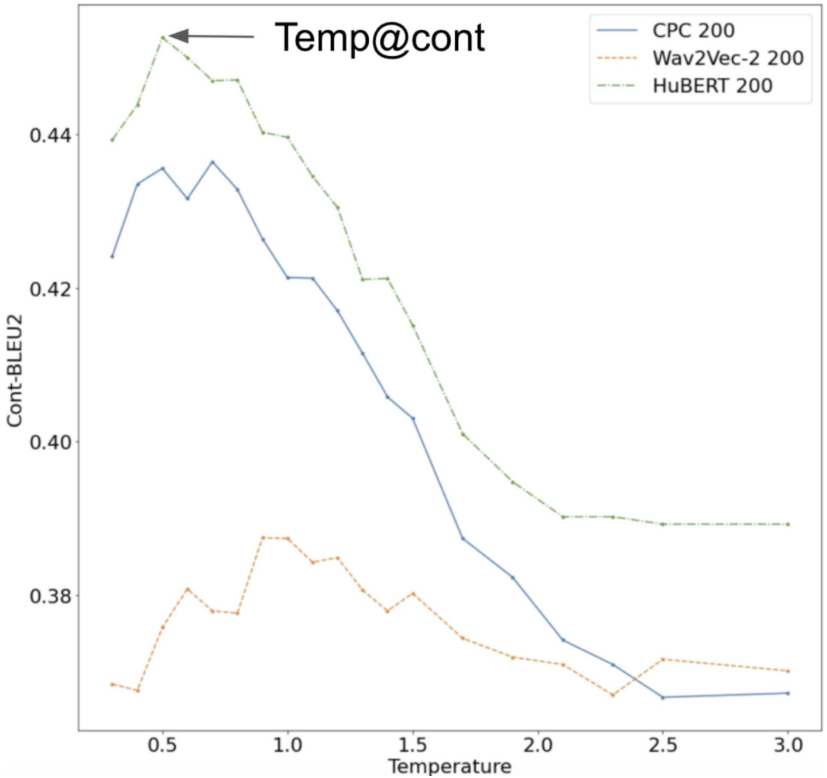
Формулы для метрик

Перплексия (PP) для корпуса (W):

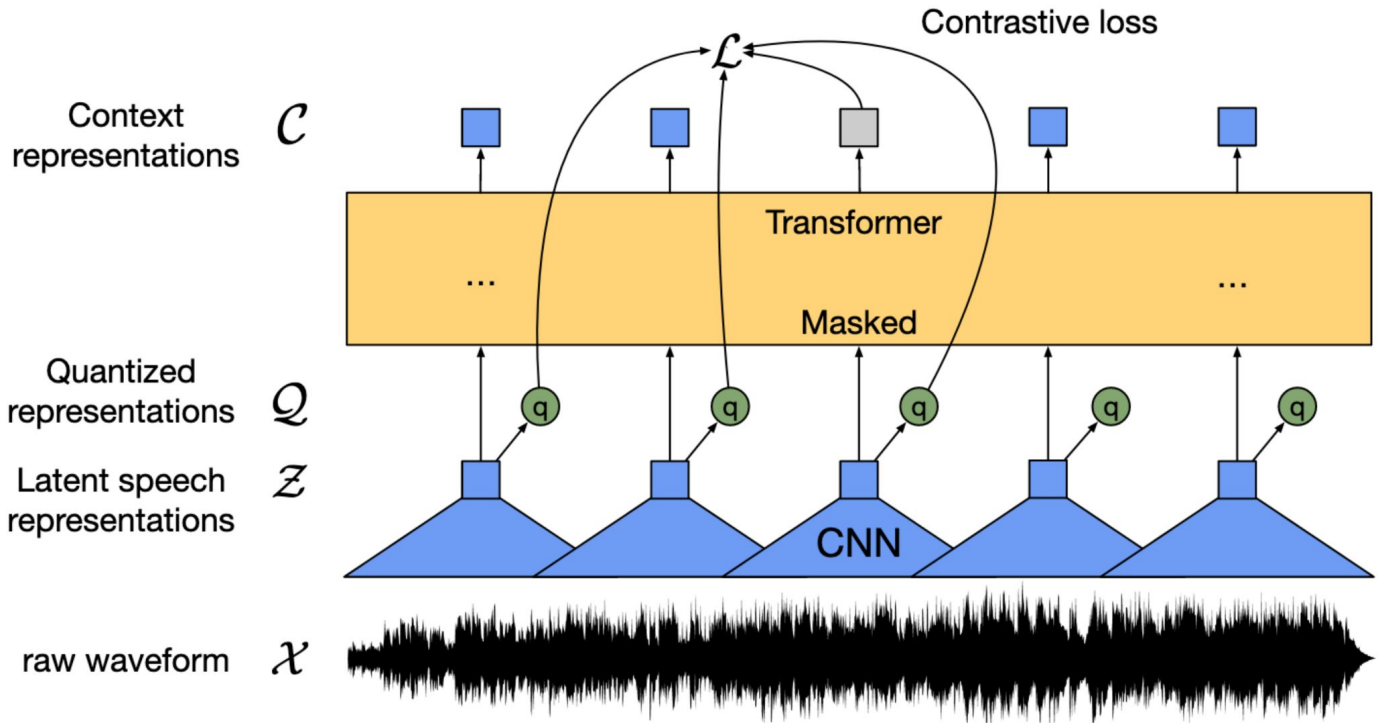
1. $PP(W) = \frac{1}{P(w_1, w_2, \ldots, w_N)^{\frac{1}{N}}}$ P – вероятность слов из модели
2. $PP(W) = 2^{H(W)}$ H – энтропия предсказаний языковой модели

- $\texttt{auto-BLEU}(u, k) = \frac{\sum_s \mathbb{1}\left[s \in (NG_k(u) \backslash s)\right]}{|NG_k(n)|}$
- $\mathbf{ABX}(\mathbf{x}, \mathbf{y}) = \frac{1}{m(m-1)n} \sum_{a \in S(\mathbf{x})} \sum_{b \in S(\mathbf{y})} \sum_{x \in S(\mathbf{x}) \backslash \{a\}} \left(\mathbb{I}_{d(a,x) < d(b,x)} + \frac{1}{2} \mathbb{I}_{d(a,x) = d(b,x)} \right)$, где $S(\mathbf{y})$ – множество звуков категории $\mathbf{y}$

Подбор температуры MMOS



Wav2vec 2.0


Contrastive loss
$\mathcal{L}$
Context representations
$\mathcal{C}$
Transformer
...
...
Masked
Quantized representations
$\mathcal{Q}$
q
q
q
q
q
Latent speech representations
$\mathcal{Z}$
CNN
raw waveform
$\mathcal{X}$

Эксперименты (S2u→u2S)

Systems			End-to-end ASR-based metrics				Human Opinion			
S2u architect.	Nb units	Bit-rate	PER↓ (LJ)	PER↓ (LS)	CER↓ (LJ)	CER↓ (LS)	MOS↑ (LJ)	MOS↑ (LS)	CER↓ (LJ)	CER↓ (LS)
Toplines										
original wav			-	-	-	-	4.83	4.30	8.88	6.73
orig text+TTS			7.78	7.92	8.87	5.14	4.02	4.03	13.25	10.73
ASR + TTS	27		9.45	8.18	9.48	5.30	4.04	4.06	15.98	11.56
Baselines										
LogMel	50	214.8	27.72	49.38	27.73	52.05	2.41	2.07	43.78	66.75
LogMel	100	292.7	25.83	45.58	24.88	48.71	2.65	2.01	37.39	62.72
LogMel	200	373.8	19.78	45.16	17.86	46.12	2.96	2.16	23.33	62.6
Unsupervised										
CPC	50	159.4	10.87	17.16	10.68	12.06	3.63	3.51	13.97	19.92
CPC	100	213.1	10.75	15.82	9.84	9.46	3.42	3.68	13.53	14.73
CPC	200	279.4	**8.74**	14.23	9.20	8.29	3.85	3.54	**9.36**	14.33
HuBERT-L6	50	125.7	11.45	16.68	11.02	11.85	3.69	3.49	14.54	13.14
HuBERT-L6	100	168.1	9.53	13.24	9.31	7.19	3.84	3.68	13.02	11.43
HuBERT-L6	200	211.3	8.87	**11.06**	**8.88**	**5.35**	**4.00**	**3.85**	11.67	**10.84**
wav2vec-L14	50	141.3	24.95	33.69	25.42	32.91	2.45	2.87	46.82	54.9
wav2vec-L14	100	182.1	14.58	22.07	13.72	17.22	3.50	3.32	23.76	28.1
wav2vec-L14	200	226.8	10.65	16.34	10.21	10.50	3.83	3.51	13.14	15.27

Эксперименты (S2u→uLM→u2S)

Systems		Generation based metrics						Human Opinion	
Encoder architect.	Nb units	unconditional PPX↓	unconditional VERT↓	unconditional AUC↓	prompt PPX↓	prompt VERT↓	prompt AUC↓	uncond. MMOS↑	prompt MMOS↑
Controls									
oracle text		154.5	19.43	-	154.5	19.43	-	4.02	4.26
ASR + LM		178.4	21.31	0.18	162.8	20.49	0.04	3.91	4.38
Baseline									
LogMel	50	1588.97	-	1083.76	-	-	-	-	-
LogMel	100	1500.11	95.50	510.26	-	-	-	-	-
LogMel	200	1539.00	-	584.16	-	-	-	-	-
Unsupervised									
CPC	50	374.26	46.26	19.68	323.9	39.92	18.44	3.31	3.61
CPC	100	349.56	41.797	15.74	294.7	42.93	14.06	3.65	3.65
CPC	200	362.84	40.28	16.46	303.5	43.42	26.67	3.58	3.67
HuBERT-L6	50	376.33	43.06	19.27	339.8	45.85	21.03	3.53	3.00
HuBERT-L6	100	**273.86**	**31.36**	**5.54**	**251.2**	**33.67**	**5.88**	3.95	3.53
HuBERT-L6	200	289.36	33.04	7.49	262.4	34.30	6.13	**4.01**	**4.32**
wav2vec-L14	50	936.97	-	307.91	1106.3	-	330.8	2.26	1.91
wav2vec-L14	100	948.96	79.51	208.38	775.1	-	205.7	2.28	1.92
wav2vec-L14	200	538.56	61.06	61.48	585.8	-	91.07	2.64	3.04

Эксперименты (zero-shot)

	Metrics	S2u		uLM	
System	Nb units	ABX with.↓	ABX acr.↓	spot-the-word↓	accept. judg.↓
Toplines					
ASR+LM		-	-	3.12	29.02
Baselines					
LogMel	50	23.95	35.86	48.52	46.78
LogMel	100	24.33	37.86	48.12	46.83
LogMel	200	25.71	39.65	49.62	47.76
Unsupervised					
CPC	50	5.50	7.20	32.18	45.43
CPC	100	**5.09**	**6.55**	31.72	44.35
CPC	200	5.18	6.83	37.40	45.19
HuBERT-L6	50	7.37	8.61	32.88	44.06
HuBERT-L6	100	6.00	7.41	**31.30**	**42.94**
HuBERT-L6	200	5.99	7.31	36.52	47.03
wav2vec-L14	50	22.30	24.56	51.92	45.75
wav2vec-L14	100	18.16	20.44	50.24	45.97
wav2vec-L14	200	16.59	18.69	44.68	45.70