Generative Spoken Language Modeling from Raw Audio

Коган Александра
Данг Нина
Степанов Никита
Станкевич Матвей

О чем статья?

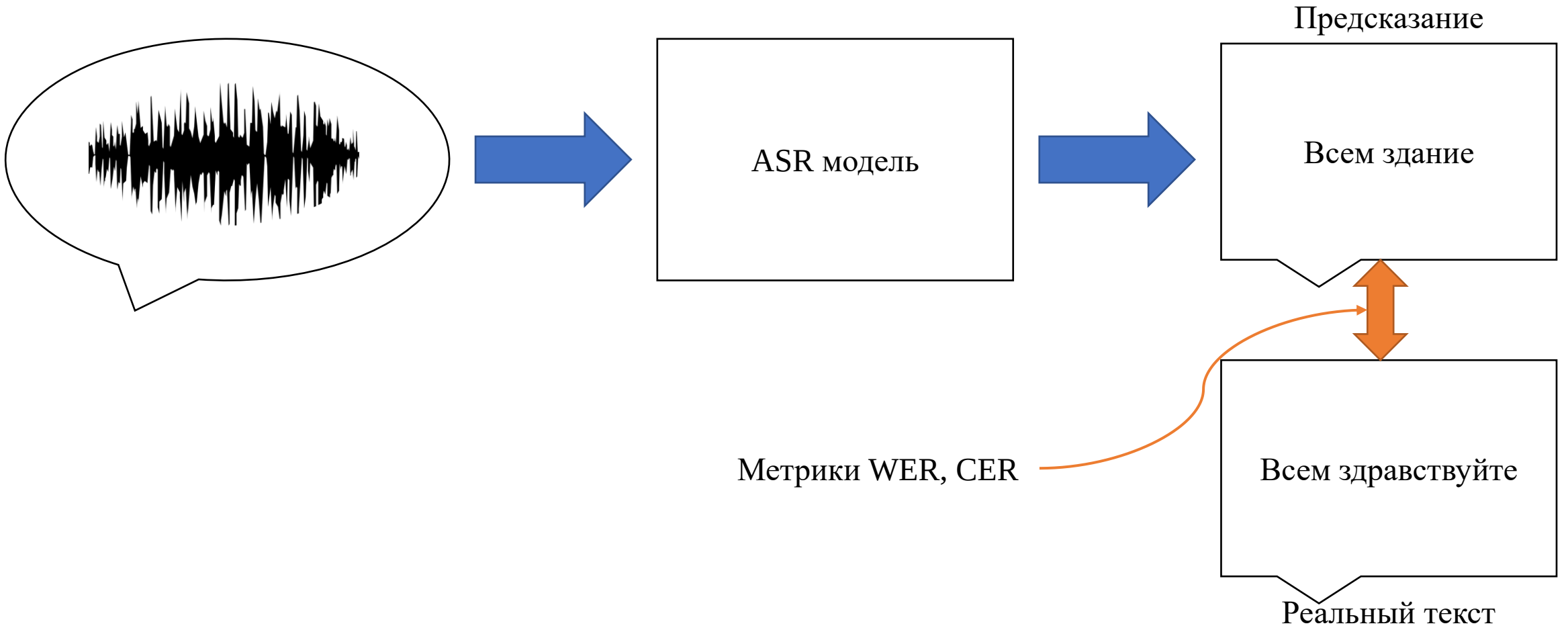
- Generative Spoken Language Modeling – задача обучения акустике и лингвистике только из аудио, без использования текстов или лейблов
- Метрики для оценки сгенерированного аудио и представлений на уровнях акустики и лингвистики.

(Не)много бэкграунда

Для тех, кто не ходил на курс Глубинное обучение в звуке

Automatic Speech Recognition (ASR)

Распознавание речи



Подробней можно прочитать в [Лекции по ASR]

WER

Word Error Rate

How to compute?

Edit path from reference to prediction

- ***S*** – substitution count
- ***D*** – deletions count
- ***I*** – insertions count
- ***C*** – correct count
- ***N*** – ***S*** + ***D*** +***C*** - Total word count in reference

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

CER

Character Error Rate

The same, but on character level

```
True: quick brown    fox jumped over a lazy dog
Pred: quick brow  an fox jumped over  lazy dog
```

Слайд взят из Лекции

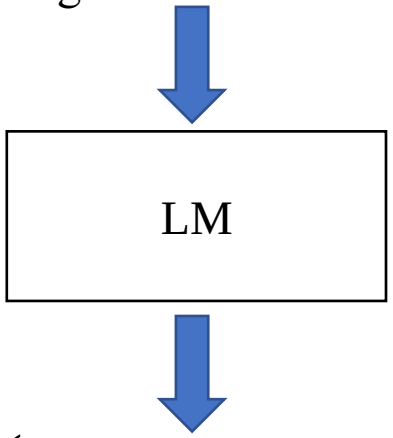
Language Models (LM) для ASR

Проблемы в ASR:
Написание слова зависит от контекста

Подсказки:
- Неразмеченные тексты получить легко
- ASR выдает тексты

LM - модель, подсчитывающая вероятность текста

let's go ***two*** a movie
let's go ***to*** a movie
let's go ***too*** a movie



P(let's go ***two*** a movie) = 0.01

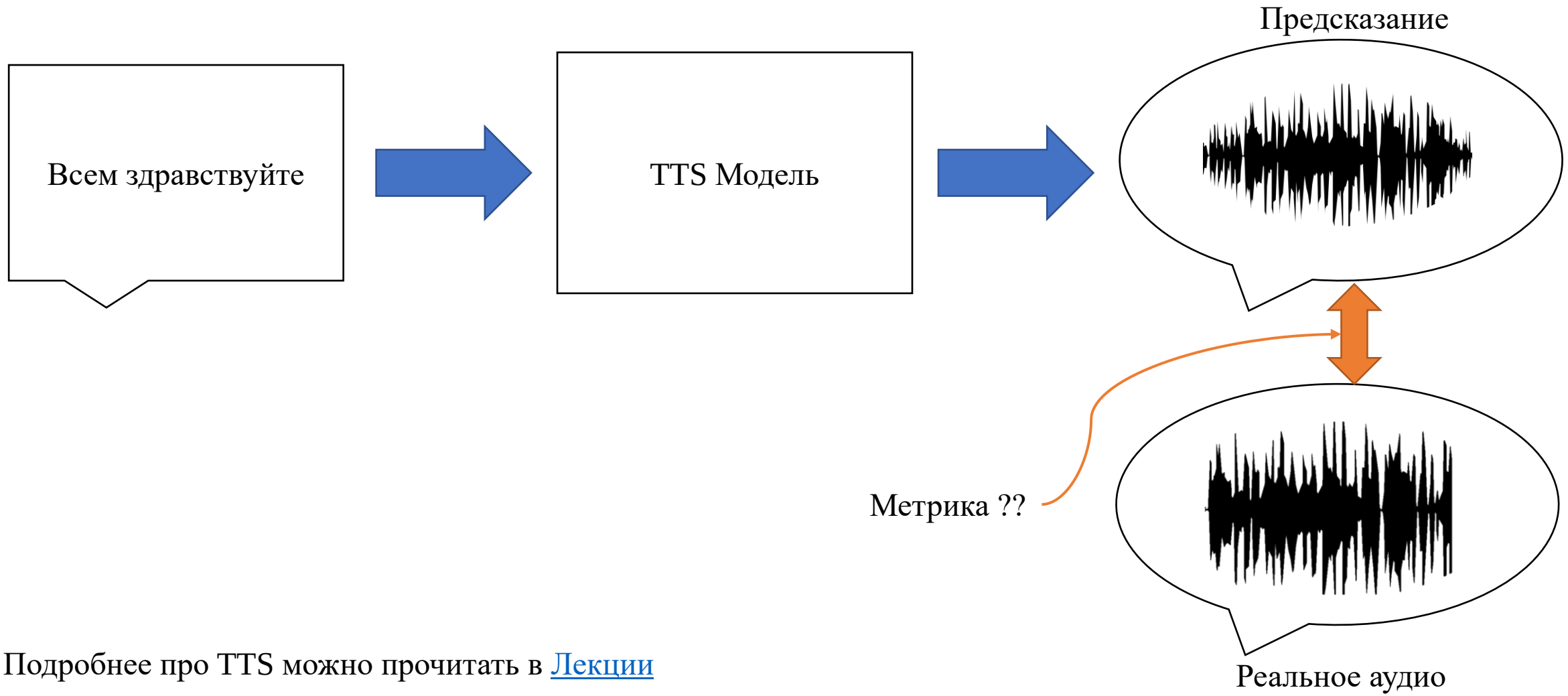
P(let's go ***to*** a movie) = 0.6

P(let's go ***too*** a movie) = 0.02

Подробнее про LM можно почитать в Лекции

Text-to-Speech (TTS)

Синтез речи



Подробнее про TTS можно прочитать в Лекции

Метрика

Важные аспекты:

- Сходство с настоящим аудио
- Общее впечатление
- Усилие для прослушивания
- Разборчивость
- Натуральность
- Приятность
- Интонация, паузы
- Эмоциональность

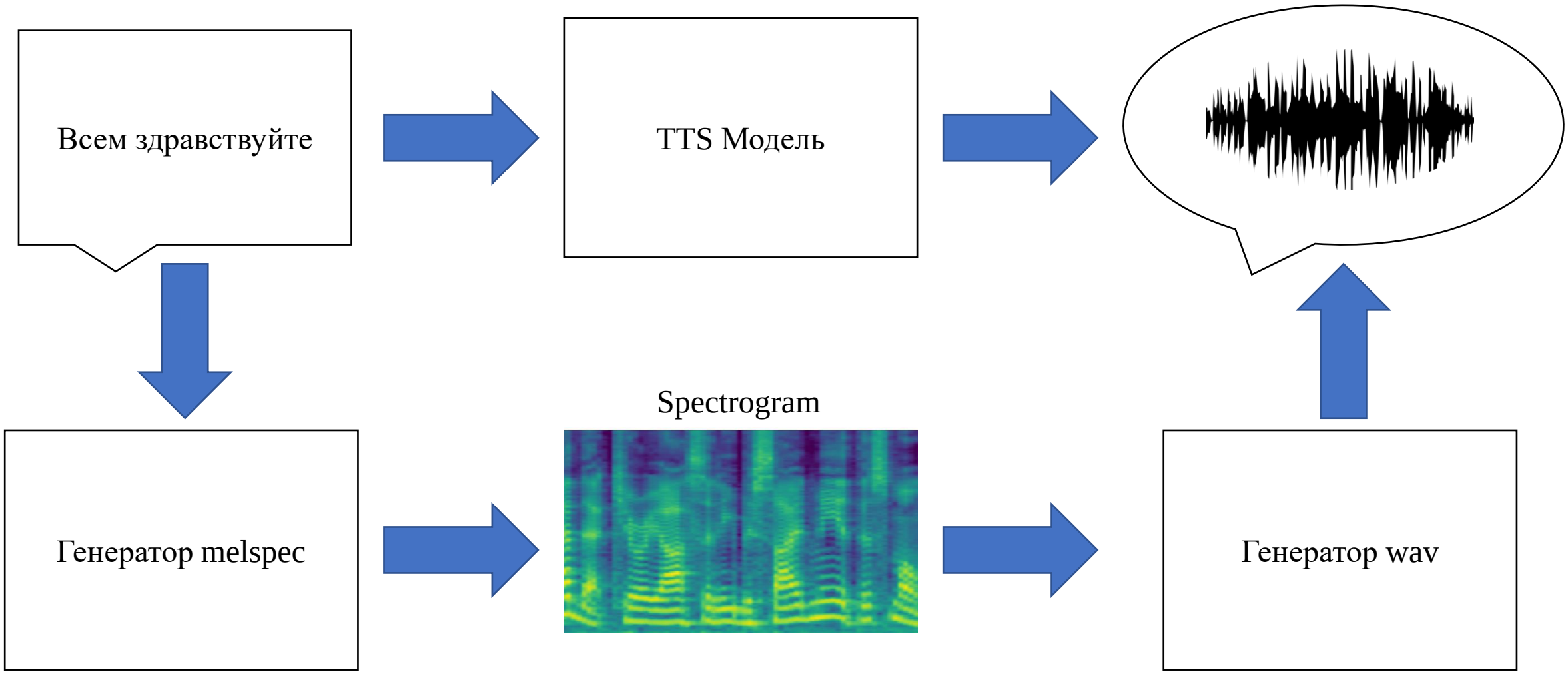
Mean Opinion Score (MOS)

- Люди оценивают аудио от 1 до 5
- Оценки усредняются

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Подробнее про MOS можно прочитать в Лекции

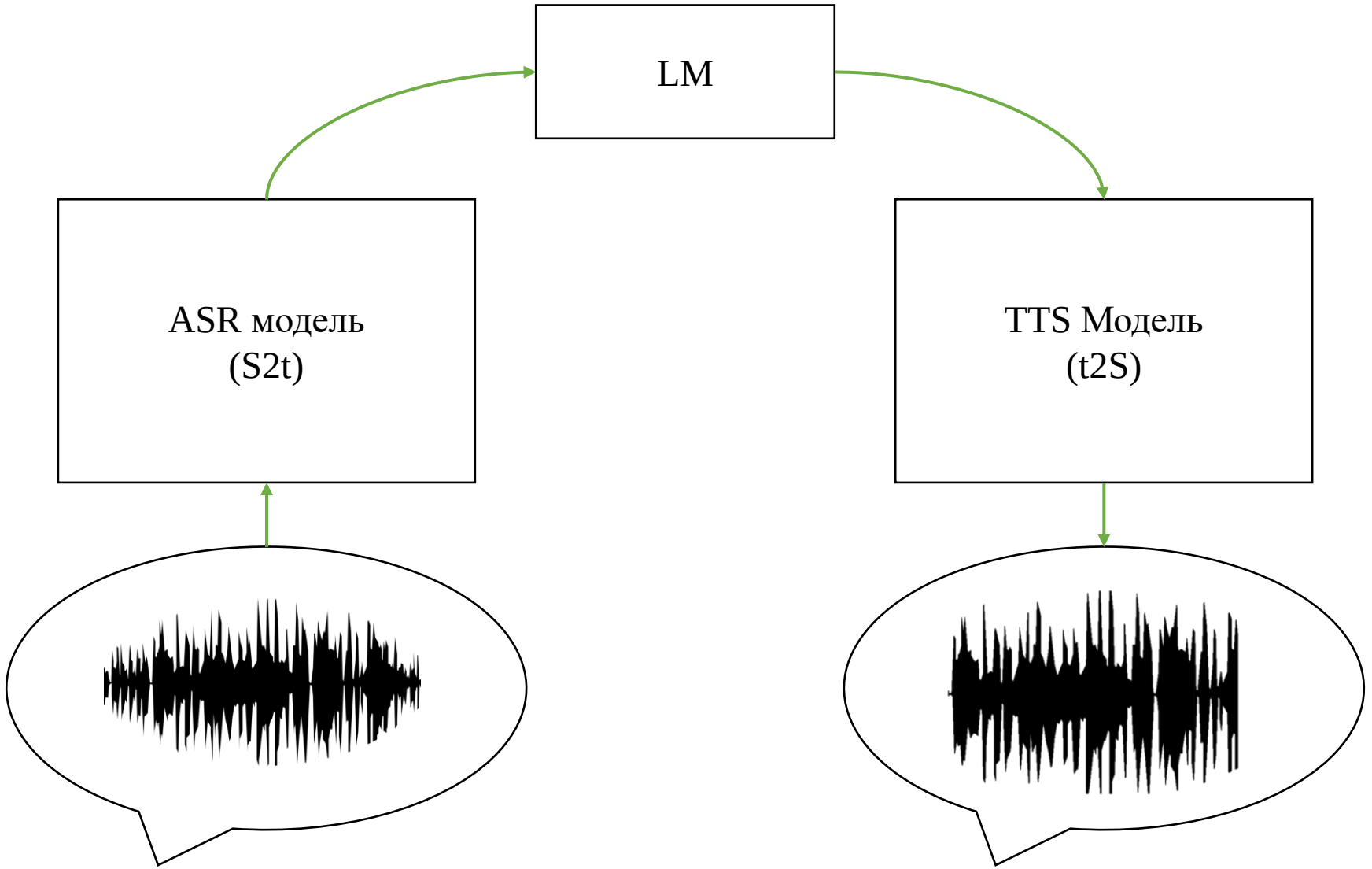
Разделение TTS модели



Про spectrogram можно почитать в Лекции

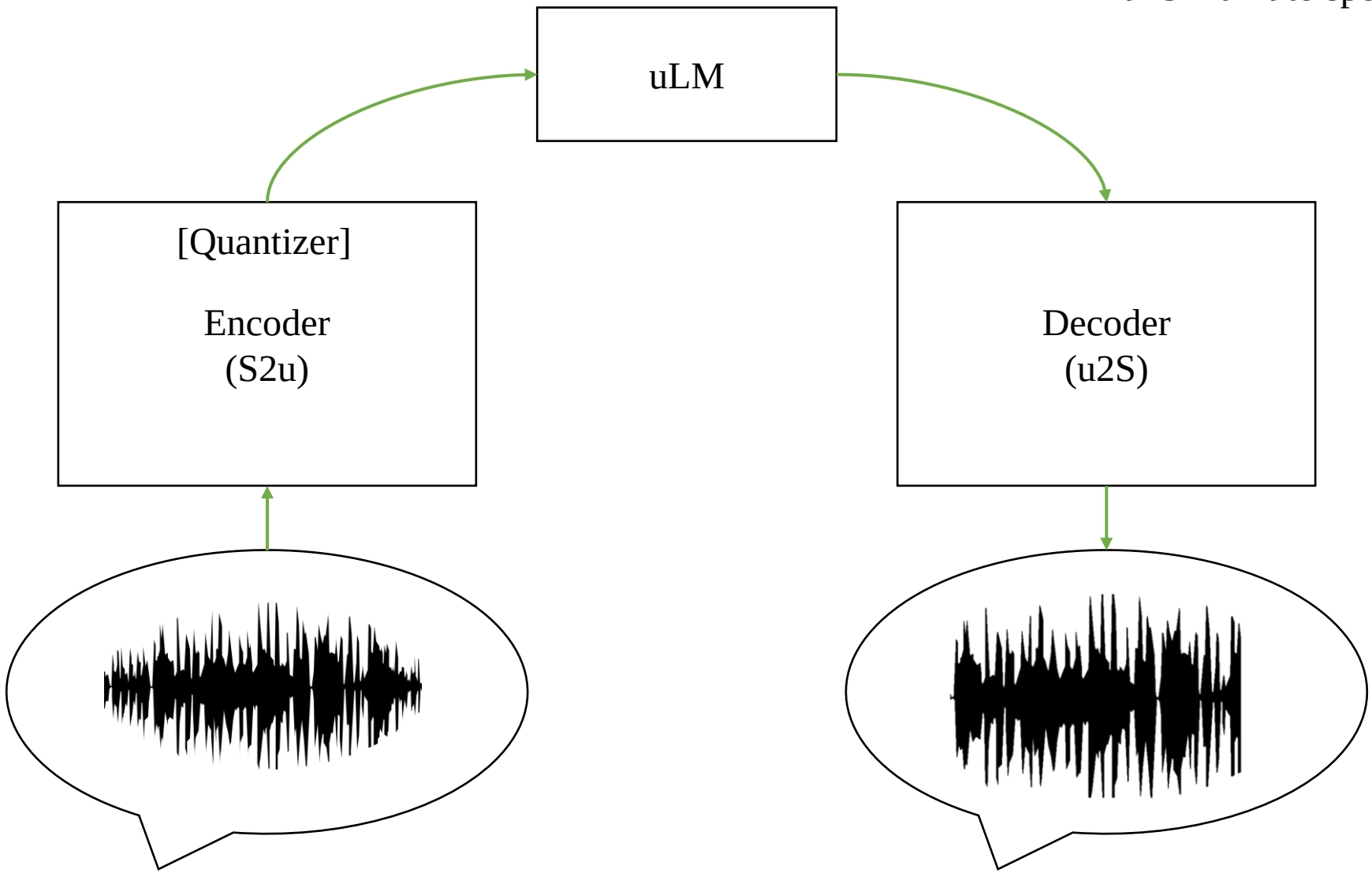
Микс из ASR, LM и TTS

Обозначу:
S2t – Speech to text
t2S – text to Speech



Возвращаемся к статье

Обозначили:
S2u - Speech to unit
uLM - unit-based language model
u2S – unit to speech



Архитектура S2u

3 state-of-the-art unsupervised энкодера «из коробки»:

(с теми же гиперпараметрами)

- **CPC** (Contrastive Predictive Coding)
- **wav2vec 2.0** LARGE
- **HuBERT** BASE 12

+ LogMel (туториал по Filter banks)

Квантизация: k-means

Архитектура uLM

Трансформер (transformer_lm_big из fairseq)

Гиперпараметры:
N = 12
h = 16
d_{model} = 1024
d_{ff} = 4096
P_{drop} = 0.1

Предварительные эксперименты: удаление unit-повторов улучшает качество

В работе: повторы удаляются

Архитектура u2S

Адаптация модели **Tacotron-2** (создание spectrogram). Берет на вход unit (псевдо-текст)

Модификации:

- Добавление специального токена «end-of-input» (EOI)
- Обучение модели с использованием chunk-ов

Pre-trained **WaveGlow** (создание wav)

Предложенные метрики

Автоматические:

- ASR метрики - для оценивания аудио
- Zero-shot probe метрики – для оценивания представлений

«Ручками»:

- Human evaluation метрики - для оценивания аудио

ASR метрики

Задача 1
Воспроизводство (resynthesis) речи:
Акустика
S2u → u2S
Нужно измерять разборчивость (intelligibility)

Пх’нглуи мглв’нафх Ктулху Р’льех вгах’нагл фхтагн

Задача 2
Генерация (generation) речи:
Лингвистика
S2u → uLM→u2S
Нужно измерять осмысленность (meaningfulness)

Где кончается безумие и начинается реальность?

(1) ASR-PER (resynthesis intelligibility)

Идея:

- Берем state-of-the-art pretrained ASR модель **без** LM
- Считаем PER (Phone Error Rate – это как CER, только для фонем)

Идея предварительных экспериментов*:

- Берем state-of-the-art pretrained ASR модель **с** LM
- Считаем WER, CER

*Результаты двух идей схожи, поэтому результаты первых экспериментов опускаются

(2) AUC on Perplexity and VERT

Идея аналогична: оцениваем текст, распознанный ASR моделью

При оценивании генерации текста обычно оценивают:

- Качество (quality). Например, mean perplexity или negative log likelihood после LM)
- Разнообразие (diversity) Например, self-BLEU

Обычно: trade-off между этими показателями (при изменении гиперпараметра температуры T при сэмплировании из LM)

(2) AUC on Perplexity and VERT

Проблема: При маленькой T self-BLEU не увеличивается, но модель повторяет слова в предложении

Решение:

- Новая метрика auto-BLEU, которая измеряет разнообразие внутри предложения:

$u - utterance,$
$s \in NG_k(u) - k - grams\ with\ u$

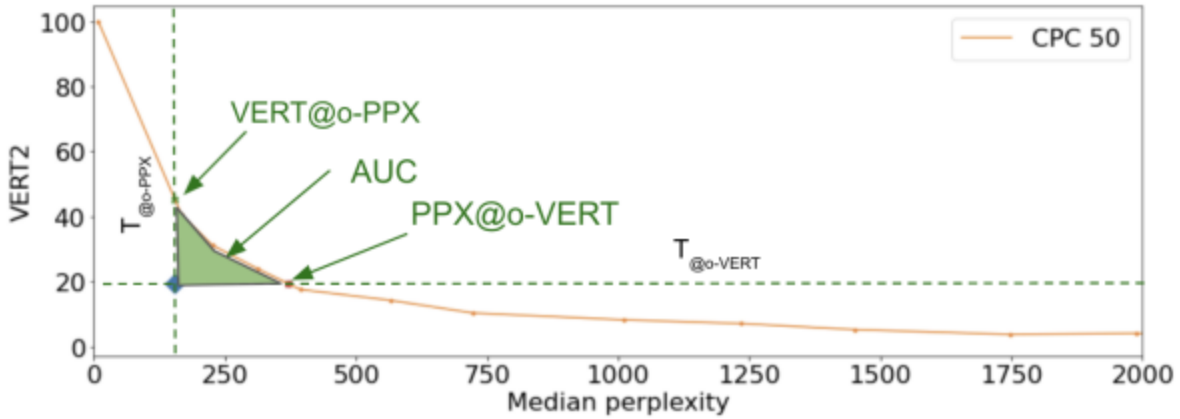
$$\texttt{auto-BLEU}(u,k) = \frac{\sum_s \mathbb{1}\left[s \in (NG_k(u) \backslash s)\right]}{|NG_k(n)|} \quad (1)$$

- Новая метрика $VERT = \sqrt{auto - BLEU \times self - BLEU}$

(2) AUC on Perplexity and VERT

Проблема: «критическая» T, при которой выход нормальный сильно зависит от модели

Решение: использовать Perplexity(PPX) и VERT для oracle текста в качестве границ и считать AUC между ними



Comparison of **diversity** and **perplexity** of the generated speech. We plot VERT vs. Median perplexity. The blue diamond corresponds to the oracle reference point. It defines two cut-offs on the curve: VERT @oracle-PPX and PPX @oracle-VERT. The green area corresponds to the AUC metric.

(2) AUC on Perplexity and VERT

Алгоритм:

- Находим PPX, VERT для oracle текста (o-PPX, o-VERT)
- Находим T_{o-PPX} - температуру, когда PPX сгенерированного текста = o-PPX
- Находим T_{o-VERT} - температуру, когда VERT сгенерированного

текста = o-VERT

- Считаем AUC-PPX-VERT между T_{o-PPX} и T_{o-VERT}

AUC ↓ ⇒ ближе к oracle ⇒ лучше

Zero-shot probe метрики

Цель – оценить качество представлений в пайплайне (S2u → uLM) на уровнях:

1. Акустики,
2. Лингвистики

Метрики zero-shot, так как не требуют обучения никакого классификатора

(1) Zero-shot probe метрики для акустики

ABX score:

$$A, B \ - \text{категории}$$
$$a, x \ \in A; b \in B$$
$$ABX = P(x \text{ ближе к } a, \text{чем к } b)$$

ABX-within Категории – 3-звучия, которые отличаются только центральной фонемой. Например, bit-bet-bat

ABX-across Категории – разные говорящие

+**bitrate** $B(U) = n \sum_{i=1}^{n} \frac{p(s_i)\, log_2\, p(s_i)}{D}$ $U = [s_1, s_2, \dots, s_n],$ $D \ - len\ in\ seconds$

(2) Zero-shot probe метрики для лингвистики

spot-the-word accuracy – доля правильно выделенных слов из пары «настоящее» - «ненастоящее» utterance (транскрипции).

Например, ‘brick’ и ‘blick’

Вычисление: считается вероятность utterances после uLM

Human evaluation метрики

Аналогично ASR-метрикам, оценивается аудио на уровнях:

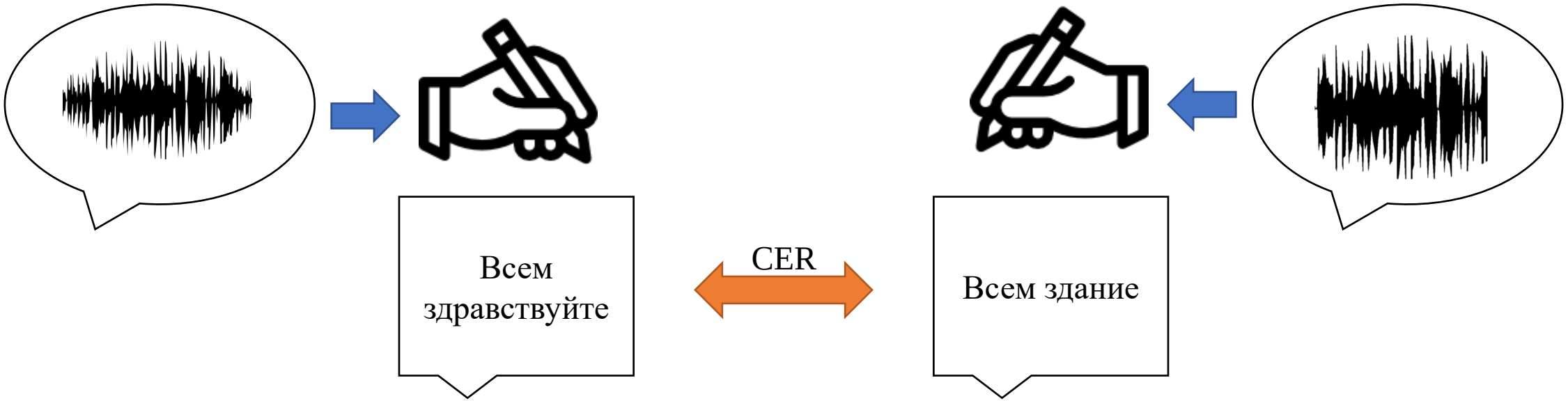
1. Акустики (разборчивость)
2. Лингвистики (осмысленность)

(1) Human evaluation метрики

Оценка разборчивости:

i. **MOS** – насколько разборчиво? (От 1 до 5 с шагом 1)

ii. **CER** – считается на транскрипциях ⇒ более объективный тест на разборчивость



(2) Human evaluation метрики

Оценка осмысленности:

meaningfulness-MOS (MMOS) – насколько естественно (грамматика и смысл)? (От 1 до 5 с шагом 1)

Выбор T:

Предварительные эксперименты: люди любят маленькую T (меньше diversity, больше quality)

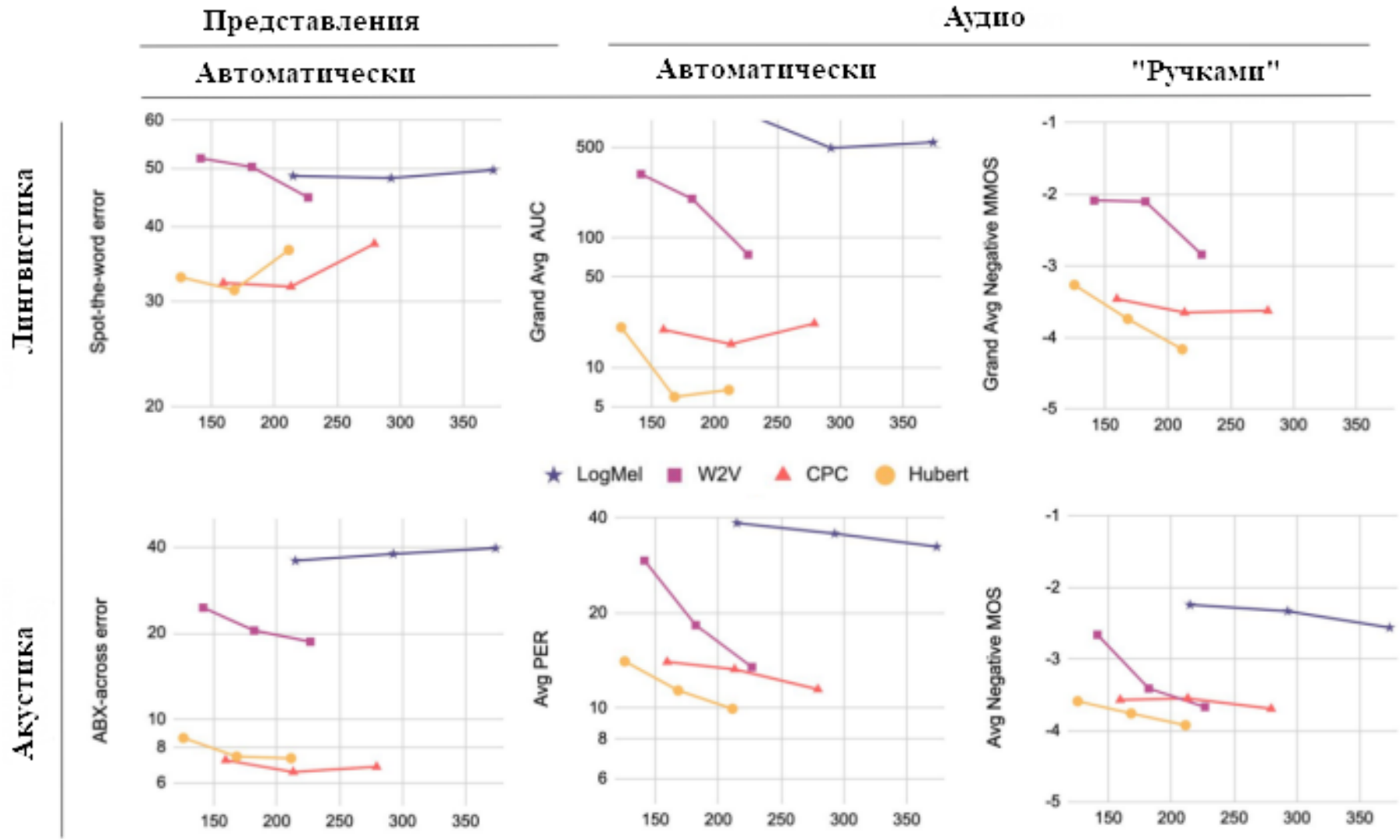
В работе: взяли короткие фразы из датасета Librespeech, прогнали с разными T, выбрали ту T, которая дает лучший BLEU-2 после ASR

Все метрики вместе

		Акустика	Лингвистика
Аудио	ASR метрики	ASR-PER	AUC (PPX-VERT)
	Human evaluation метрики	MOS, CER на транскрипциях	MMOS
Представления (units)	Zero-shot probe метрики	ABX-within, ABX-across, bitrate	spot-the-word accuracy

Результаты

Графики метрик в зависимости от bitrate
При разном количестве unit (50, 100, 200)
Чем меньше метрики, тем лучше.
Даны -MOS и -MMOS



Результаты - Аудио - Акустика

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

Результаты - Аудио - Лингвистика

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

Результаты - Представления

	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

Заключение

- Новая задача: Generative Spoken Language Modeling
- Метрики для данной задачи (ASR-based, Zero-shot probe, Human evaluation), которые измеряют качество представлений и аудио на уровнях акустики и лингвистики
- Вывод о корреляции автоматических и «ручных» метрик
- Хорошие результаты генерации (действительно возможно использовать только аудио)

Послушать аудиозаписи можно на сайте