

On Generative Spoken Language Modeling from Raw Audio

Выполнила Стахова Екатерина, БПМИ 193

<https://arxiv.org/pdf/2102.01192.pdf>

Возможен ли 'textless NLP'?

Хотим: чтобы модель научилась "говорить" на языке, обучаясь исключительно на аудио данных.

А возможно ли это? В чем мотивация генерации речи, на основе аудио фрагментов?

Возможен ли 'textless NLP'?

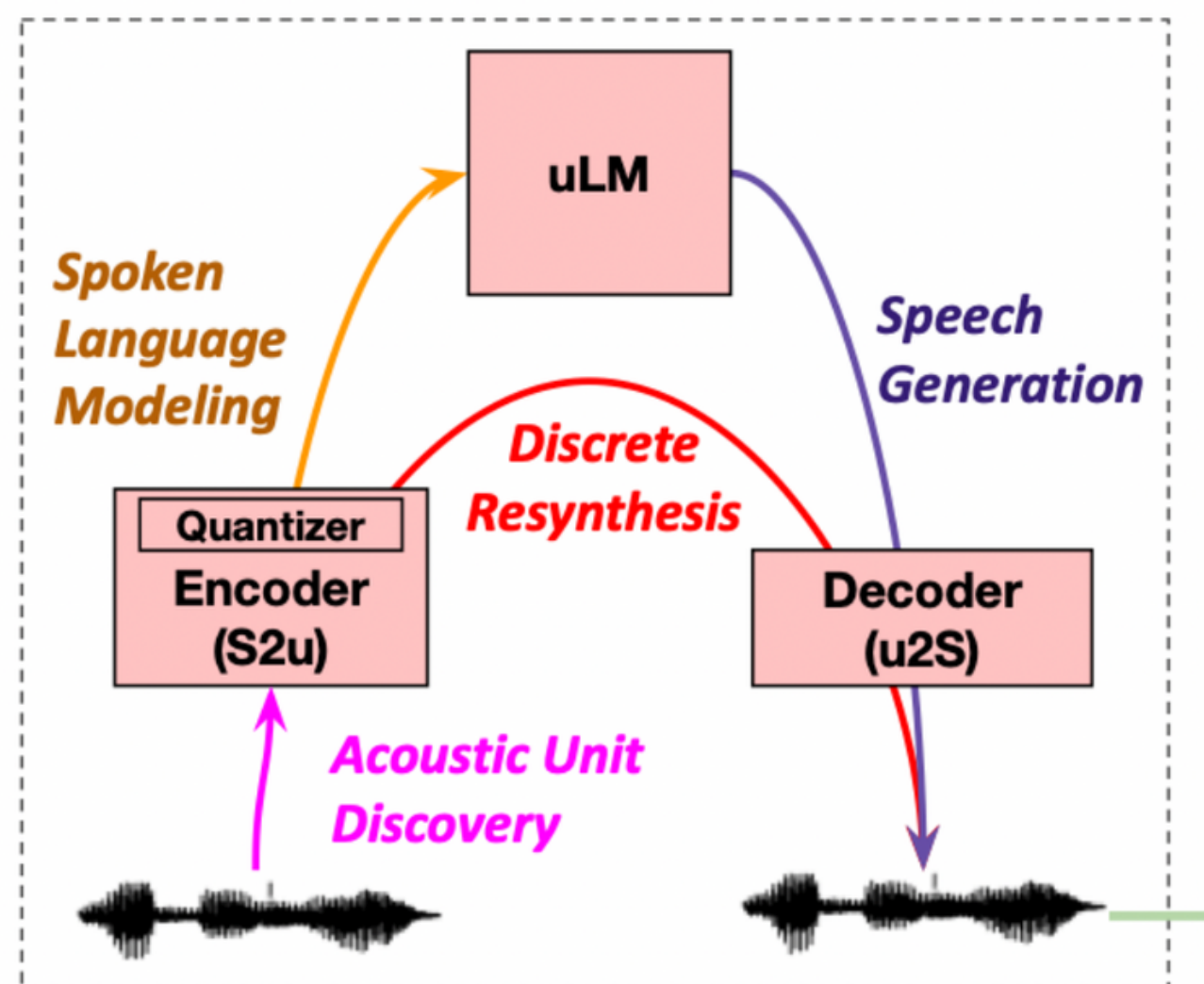
Хотим: чтобы модель научилась "говорить" на языке, обучаясь исключительно на аудио данных.

А возможно ли это? В чем мотивация генерации речи, на основе аудио фрагментов?

Какие проблемы это решит:

- Для языков и их узких диалектов, у которых не достаточно текстовых данных для обучения, да и нет так таковых правил орфографии.
- Для популярных языков такой подход позволит учитывать различия в произношении и написании слова, такие особенности как тона, интонации.

Архитектура модели



**Model architecture
and tasks**

Компоненты:

speech-to-unit - из аудиозаписи делает псевдотекст

unit-based language model

unit-to-speech - из векторного представления синтезирует речь произвольным голосом

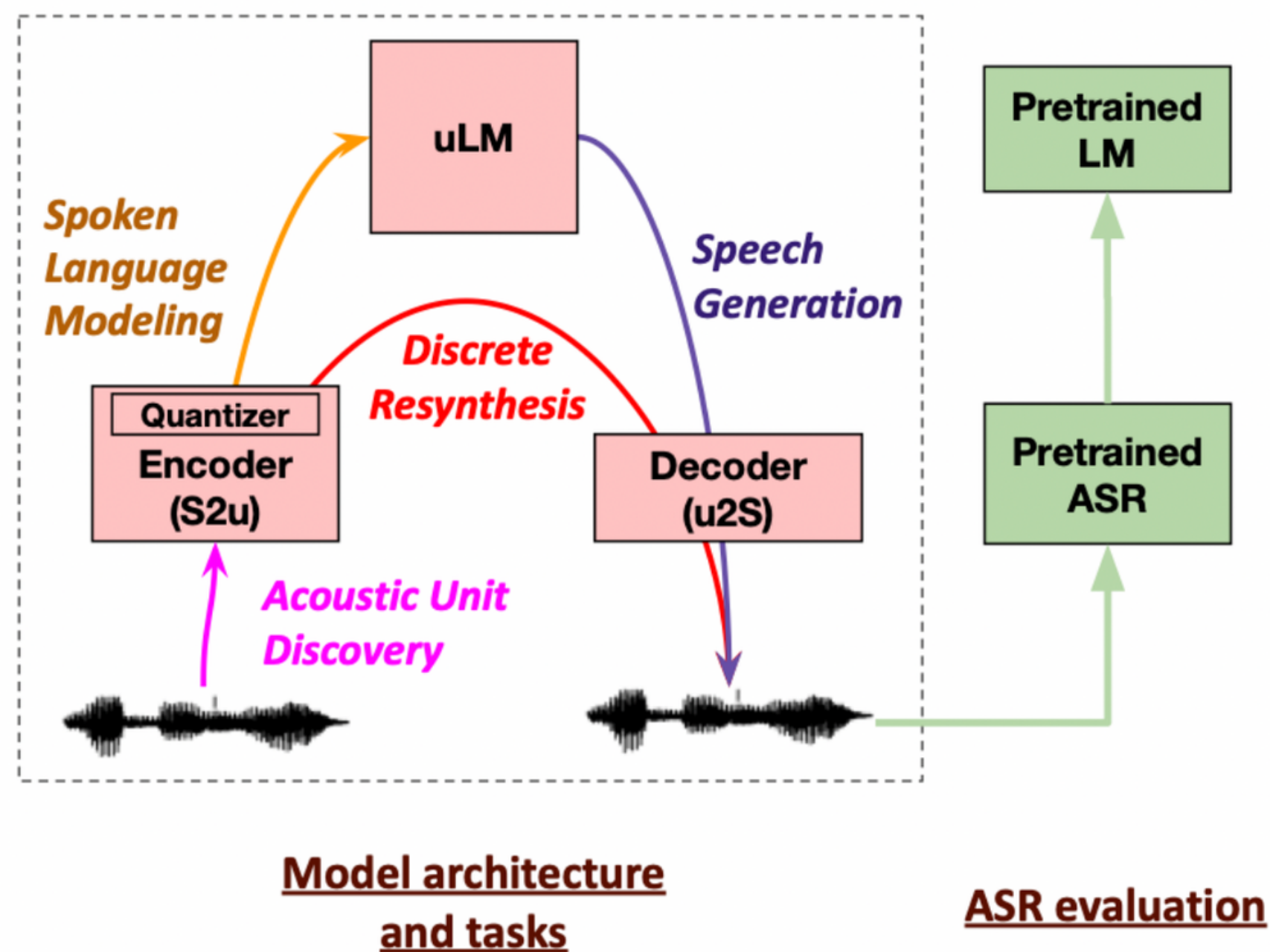
Как оценивать результаты?

Level	Encoding		Generation		
	Task	Automatic metric	Task	Automatic metric	Human
Language	Spoken LM	Spot-the-word , Syntax-Acc	Speech Gen.	AUC-of-VERT/PPX , cont-BLEU, PPX@o-VERT	MMOS
Acoustic	Acoustic Unit Disc.	ABX-across , ABX-within	Resynthesis	PER-from-ASR , from-ASR CER-	CER , MOS

GSLM оценивают по 2-м уровням:

- Акустический (членораздельность)
- Языковой (осмысленность)

Generation: ASR metrics



Основная идея - выход модели с помощью ASR переводить в текст и использовать метрики для текстовых данных.

Model + ASR - проверка на произношение

Model + ASR + LM - проверка на лексический смысл выражения

Speech resynthesis intelligibility: ASR-PER

Вспомним текстовые метрики

WER - word error rate

CER - character error rate

- **Substitution error:** Misspelled characters/words
- **Deletion error:** Lost or missing characters/words
- **Insertion error:** Incorrect inclusion of character/words

STEAM

STEAM

STEAM


STEAL

TEAM

STREAM

 Substitution

 Deletion

 Insertion

Speech resynthesis intelligibility: ASR-PER

1. *m*itten → *f*itten (substitute **m** with **f**)
2. *f*itten → *f*itt**i**n (substitute **e** with **i**)
3. *f*ittin → *f*itt**g** (insert **g** at the end)

$$CER = \frac{S + D + I}{N}$$

Character Error Rate (CER) formula

where:

- **S** = Number of Substitutions
- **D** = Number of Deletions
- **I** = Number of Insertions
- **N** = Number of characters in reference text (aka ground truth)

Speech resynthesis intelligibility: ASR-PER

Но мы будем использовать

PER - phone error rate, который вычисляется аналогично, но относительно фонем.

Speech generation quality and diversity: AUC on Perplexity and VERT

Хотим оценивать качество генерации речи относительно ее осмысленности и разнообразия.

Как правило, это зависит от гиперпараметров LM:

low temperature - речь осмысленна, но не вариативна.

high temperature - речь вариативна, но не обладает смысловой нагрузкой.

Как же найти trade-off между этими двумя показателями?

Speech generation quality and diversity: AUC on Perplexity and VERT

$$\text{auto-BLEU}(u, k) = \frac{\sum_s \mathbb{1} [s \in (NG_k(u) \setminus s)]}{|NG_k(n)|}$$

u - выражение

k -грамм

Числитель - считаем кол-во k -грамм, которые встречались хотя бы 2 раза.

Знаменатель - кол-во k -грамм в выражении.

Speech generation quality and diversity: AUC on Perplexity and VERT

VERT = `geom_mean(self-BLEU, auto-BLEU)`

VERT - среднее геометрическое от self-BLEU и auto-BLEU (2-граммные версии)

Speech generation quality and diversity: AUC on Perplexity and VERT

Perplexity = уровень озадаченности

Интуитивно, если модель присваивает тестовому набору высокую вероятность, это означает, что она не удивлена, увидев его (она не озадачена этим), что означает, что она хорошо понимает, как работает язык.

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_N)}}$$

Speech generation quality and diversity: AUC on Perplexity and VERT

Perplexity = уровень озадаченности

Test Set

“Yesterday I went to the cinema”
“Hello, how are you?”
“The dog was wagging its tail”

High probability
Low perplexity

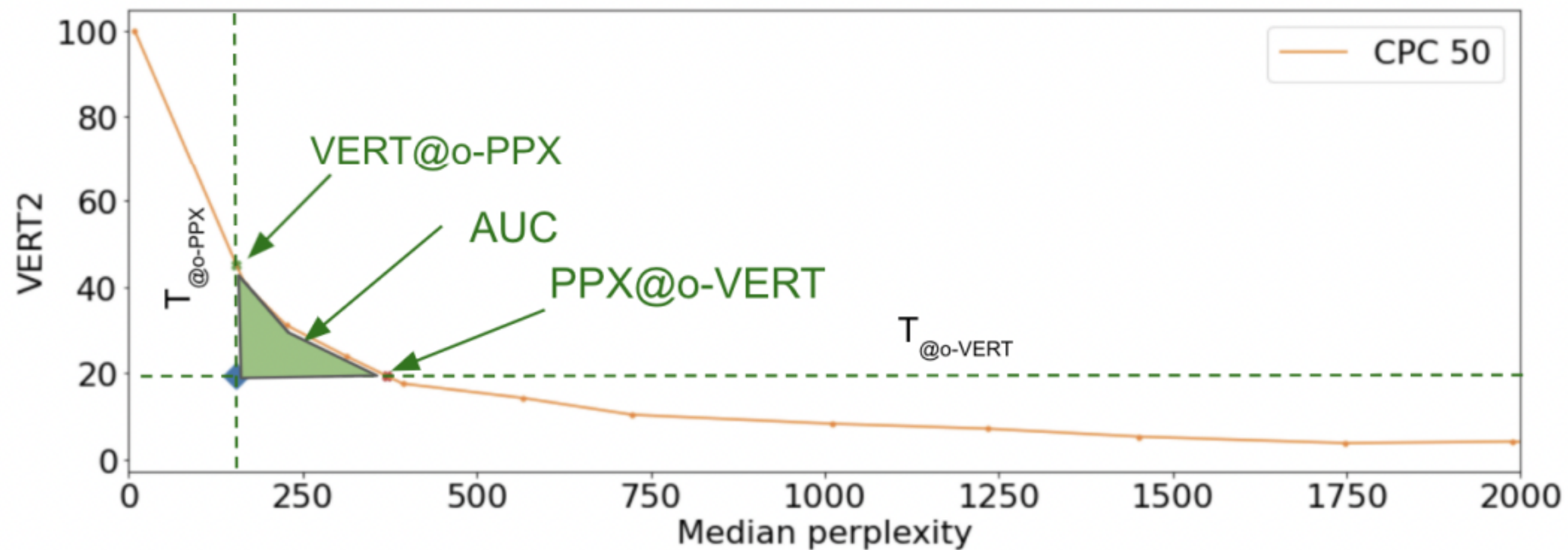
Fake/incorrect sentences

“Can you does it?”
“For wall a driving”
“She said me this”

Low probability
High perplexity

Speech generation quality and diversity: AUC on Perplexity and VERT

Trade-off осмысленности и разнообразия:



Encoding: Zero-shot probe metrics

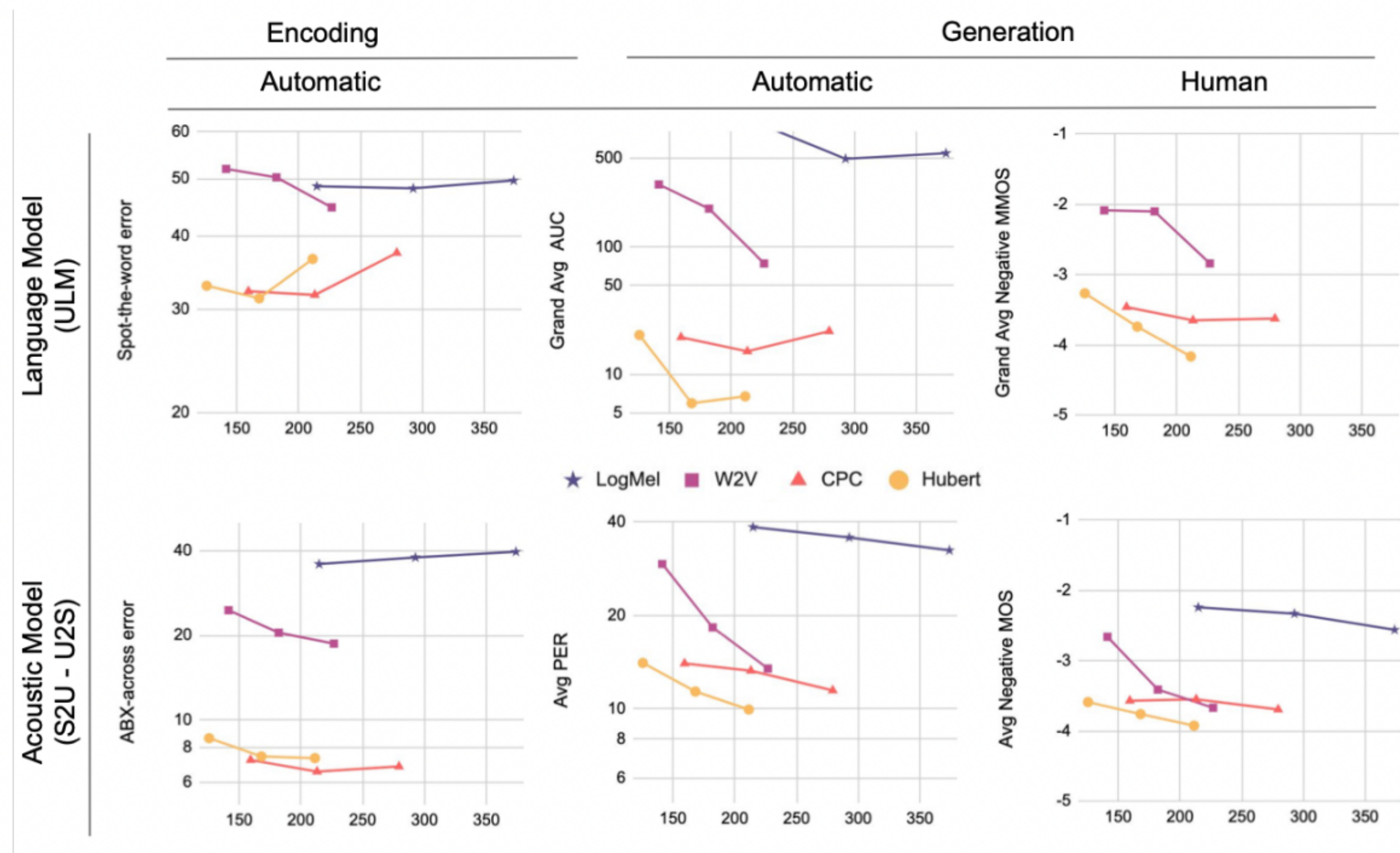
Поговорим про метрики для энкодинга:

Как оценить насколько информативны представления, которые поступают на вход LM?

Для акустических данных - **ABX score** между дикторами на эмбедингах.

Для языковых данных - **spot-the-word accuracy**

Результаты



Результаты

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)
<i>Toplines</i>										
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
<i>Baselines</i>										
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
<i>Unsupervised</i>										
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