SpeechLMScore

Evaluating Speech Generation using Speech Language Model

Soumi Maiti[1], Yifan Peng[1], Takaaki Saeki[1,2], Shinji Watanabe[1]

[1] Carnegie Mellon University, [2] The University of Tokyo

ICASSP 2023

Introduction

Objective & Subjective Evaluation

When evaluating **Speech Generation** and **Speech Enhancement** tasks, **Human Subjective Evaluation** is often relied on.

However, subjective evaluation is **Time-Consuming and Expensive**, so there are many objective methods used to replace humans to evaluate various characteristics of speech.

But, objective methods, although convenient and fast, does **NOT** show a **high correlation** with human evaluation scores.

Introduction (cont.)

Subjective Score Estimation

In order to truly replace human evaluation, many recent studies have collected **Speech and Corresponding Subjective Scores** to train scoring models.

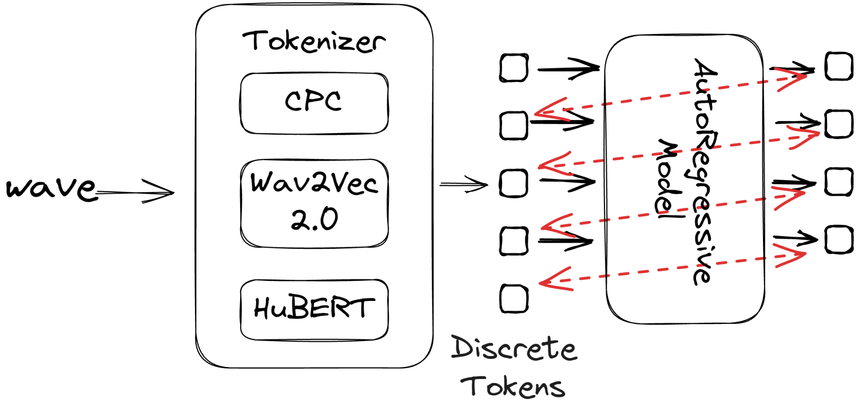
Nevertheless, **the Scarcity of Data** due to the high collection cost makes **the Generalization Ability** of these supervised models **Need to be Improved**.

This study refers to the unsupervised evaluation metrics of NLG and proposes SpeechLMScore.

Unlike past subjective score estimation methods, human scoring labels are not required when training the scoring model.

Speech Language Model

Discrete Units + Autoregression = LM



A tokenizer maps continuous speech signal $\mathbf{x}$ into a series of discrete units $\mathbf{d}$ as

$$\mathbf{Tok}(\mathbf{x}) = \mathbf{d} = [d_1, \ldots, d_T]\,,\ d_i \in \{1, \ldots, V\}$$

And models the probability distribution over the set of discrete tokens $\mathbf{d}$ as

$$p(\mathbf{d}|\theta) = \prod_t p(d_t|d_{<t}, \theta)$$

SpeechLMScore

Perplexity of Speech Units

$$\text{SpeechLMScore}(\mathbf{d}|\theta) = \frac{1}{T}\sum_{t} \log p(d_t|d_{<t}, \theta)$$

Measuring how perplexed a speech language model is given set of discrete tokens from speech $\mathbf{x}$.

- Lower perplexity, or higher log-likelihood, should correlate with human evaluations of higher speech quality.
- No need to collect expensive human evaluation scores as training data.
- No need to have reference speech to estimate score.

Expriments

VoiceMOS 2022 Challenge

A total of **7106 sentences** of synthetic speech and natural speech generated by **187 kinds of speech synthesis systems**

- were collected from Blizzard Challenges, Voice Conversion Challenges, ESPnet-TTS, and
- each speech was given by **8 listeners** with 1~5 naturalness assessment.

And evaluate the **Correlation** between the **Model Estimated** score and the **Subjective** score by

- Linear Correlation Coefficient (LCC),
- Spearman Rank Correlation Coefficient (SRCC), and
- Kendall Tau Rank Correlation (KTAU).

Experiments (cont.)

Speech Language Model Setup

This study uses GSLM composed of **Transformers (pretrained)** and **LSTM trained from scratch** as autoregressive model, and uses **HuBERT-Base-LS960H** as tokenizer.

- GSLM will **Remove the Repeated** speech units, and it was trained on a "clean" subset containing 6K hours speech selected from LibriLight 60K dataset.
- LSTM-Base model was trained on LibriLight medium segmented set with 5.6K hours of speech.
- LSTM-Large model trained on **16.8K (three times of LSTM-Base) hours** of speech randomly selected from the LibriLight 60K hour dataset.

Experiments (cont.)

Performance Difference Caused by Tokenizer.

Table 1. Utterance and system-level correlation with MOS in VoiceMOS challenge 2022 dataset (7106 files) [7] with different configurations: layer number (L) to extract feature from Hubert and number of clusters (V). We use SpeechLMScore with pretrained uLM.

ID	V	L	Utterance-level LCC	Utterance-level SRCC	Utterance-level KTAU	System-level LCC	System-level SRCC	System-level KTAU
50_3	50	3	**0.472**	0.490	0.343	0.753	0.749	0.549
50_4		4	0.464	0.492	0.344	**0.760**	**0.755**	**0.562**
50_6		6	0.462	0.462	0.321	0.694	0.692	0.496
50_12		12	0.279	0.348	0.234	0.514	0.555	0.388
100_2	100	2	0.376	0.460	0.321	0.673	0.683	0.503
100_3		3	0.322	0.505	0.355	0.598	0.666	0.490
100_4		4	0.379	0.527	0.370	0.705	0.741	0.552
100_5		5	0.282	0.482	0.337	0.552	0.615	0.444
100_6		6	0.300	0.454	0.317	0.523	0.559	0.392
100_12		12	0.289	0.375	0.259	0.532	0.562	0.394
200_3	200	3	0.419	**0.538**	**0.380**	0.719	0.726	0.539
200_4		4	0.464	0.536	0.378	0.701	0.700	0.511
200_6		6	0.360	0.487	0.342	0.594	0.649	0.471

- Speech LMs trained with tokenizers with different numbers of clusters have advantages in different correlation metrics.
- Units taken from lower layers correlate better with human evaluations.

Experiments (cont.)

Performance Comparison

Table 2. Utterance and System-level correlation with MOS in VoiceMOS 2022 challenge testset (1066 files) and whole dataset (7106 files) with SpeechLMScore.

	test-set						**whole-set**					
	Utterance-level			**System-level**			**Utterance-level**			**System-level**		
Model	LCC	SRCC	KTAU	LCC	SRCC	KTAU	LCC	SRCC	KTAU	LCC	SRCC	KTAU
Matched training domain												
MOSNnet (pre)	0.454	0.480	0.339	0.481	0.459	0.323	0.415	0.432	0.302	0.518	0.497	0.356
MOSNet (ft)	**0.868**	**0.865**	**0.690**	**0.948**	**0.944**	**0.803**	-	-	-	-	-	-
Mismatched training domain												
DNSMOS (SIG)	**0.536**	**0.553**	**0.392**	**0.652**	**0.684**	**0.498**	**0.495**	**0.503**	**0.354**	**0.714**	**0.720**	**0.532**
DNSMOS (BAK)	0.266	0.298	0.204	0.370	0.410	0.282	0.293	0.317	0.219	0.429	0.410	0.280
DNSMOS (OVRL)	0.496	0.497	0.352	0.606	0.623	0.450	0.473	0.473	0.334	0.678	0.668	0.488
Unsupervised												
SpeechLMScore (Pre)	0.452	0.524	0.371	0.711	0.745	0.547	0.490	0.472	0.343	0.749	**0.754**	0.549
SpeechLMScore (LSTM)	0.538	0.539	0.383	0.720	0.728	0.531	0.497	0.499	0.350	0.753	0.748	0.554
SpeechLMScore (LSTM)+rep	0.582	0.572	0.410	**0.743**	**0.749**	**0.551**	**0.519**	**0.516**	**0.367**	**0.759**	0.739	**0.564**
SpeechLMScore (LSTM) (Large)	0.540	0.536	0.381	0.709	0.724	0.529	0.496	0.497	0.349	0.745	0.744	0.551
SpeechLMScore (LSTM)+rep (Large)	**0.586**	**0.584**	**0.419**	0.729	0.736	0.539	0.514	**0.516**	0.365	0.749	0.733	0.542

- Pre: using Generative Spoken Language Modelling (GSLM) as the pretrained speech language model.
- rep: without removing repeated units.
- Large: Trained on 16.8k hours of corpus.

Conclusions

Unsupervised Speech Quality Metrics

Proposed SpeechLMScore, an automatic metric for evaluating speech samples using speech language models.

- Easy to use and does **NOT require reference speech sample**.
- Trained using speech dataset only, and does **NOT need large-scale human evaluation data**.
- Has **better generalization ability** than existing supervised automatic evaluation models.