UTDUSS: UTokyo-SaruLab System for Interspeech2024 Speech Processing Using Discrete Speech Unit Challenge

Wataru Nakata*, **Kazuki Yamauchi***, Dong Yang, Hiroaki Hyodo, Yuki Saito (*E. contribution)
The University of Tokyo

What is DIscrete Speech Unit Challenge?[1]

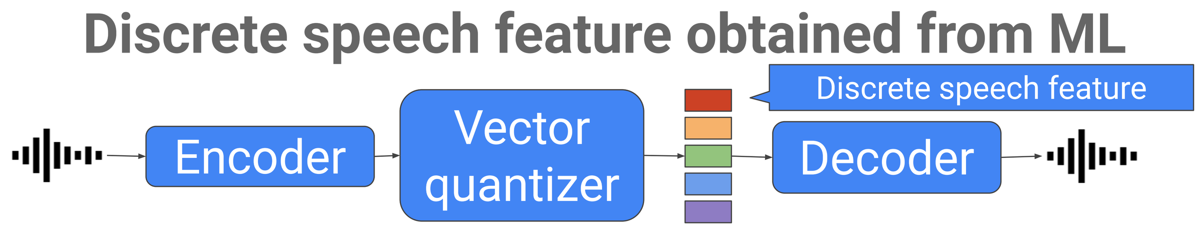
Traditional Speech processing paradigm Mel-spectrogram as a speech representation

Example of Mel-spectrogram in TTS model

😆 Works surprisingly well for most tasks

🤔 Is this human designed feature optimal?

New Approach Discrete speech feature obtained from ML



Process of discrete speech feature learning using VQVAE

😆 Optimal feature is obtained in End-to-End

🤔 How does this new feature perform on speech processing

Interspeech2024 Speech processing using discrete speech unit challenge (discrete challenge)

Goal: Promote research and **compare the results** in the speech processing **with discrete speech representation**

Four tracks:
- ASR (Automatic speech recognition)
- Vocoder
- TTS
- SVS

Vocoder, TTS: Tracks we perticipated

Our method: UTDUSS (The University of Tokyo Discrete Unit Speech Synthesizer)

UTDUSS performance on discrete challenge

🏆 1st place in TTS track

🥈 2nd place in Vocoder track

UTDUSS Discrete speech unit aquisition

Backbone model: Descript Audio Codec (DAC)[2]
- RVQGAN based discrete speech feature acquisition model
- Implement techniques to improve the performance discussed on Vocder track section
- DAC decoder is also used as a Vocoder

DAC model architecture: Improved RVQGAN

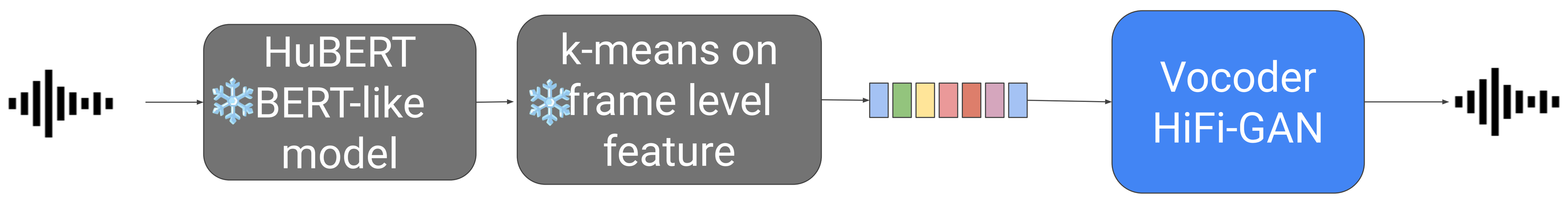
- Residual Vector quantizer (RVQ) for avoiding codebook collapse
- Adversarial training similar to HiFi-GAN
- Widely used for speech/audio discretization

UTDUSS Vocoder: 😃SMILEY

Vocoder: Task objective

Vocoder: recovers speech waveform from discretized speech representation

Baseline model: HuBERT-kmeans & HiFi-GAN[Polyak+21]



Rules

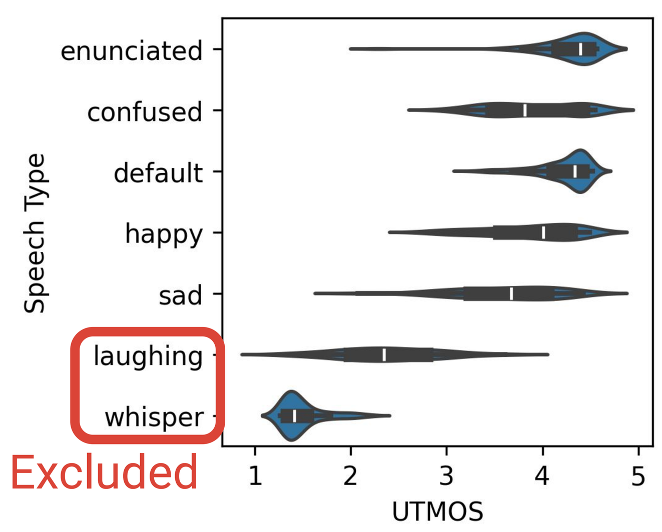
Data: EXPRESSO dataset[3]
- train/val/test split provided by the organizer
- English multi-speaker dataset
- Includes diverse speaking styles (whisper, laughter)

Evaluation metrics
- UTMOS[4] : Predicted Naturalness MOS ≠ Human evaluated
- Bitrate : The bitrate of the discretized speech

Achieve Highest UTMOS score with low bitrate as possible

Techuniqus applied for improving UTMOS

- Hyper-parameter tuning
  - Original DAC is configured for audio
- Data exclusion
  - UTMOS performs poorly on non-read style speech
- Matching sampling rate to UTMOS
  - UTMOS doesn't consider sampling rate higher than 16kHz



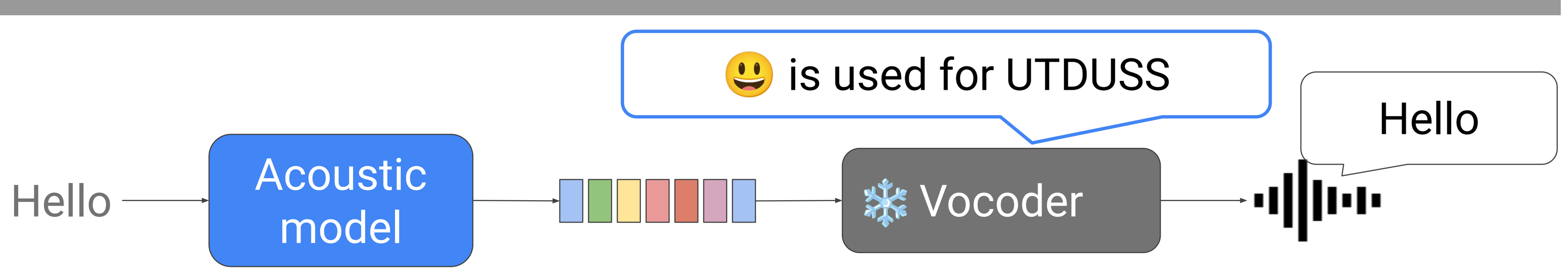
Ablation study result

Model type	Bitrate	UTMOS
baseline	448	2.310
DAC (official)	24046	3.560
😃	670	3.582
😃 w/o hyper-parameter tuning	670	3.578
😃 w/o data exclusion	670	3.568
😃 w/o matching sampling rate	1003	3.622
Ground truth		3.579

- 😃 outperform baseline, DAC and Ground truth.
- hyperparameter-tuning and data exclusion were effective
- Matching sampling rate degraded UTMOS

UTDUSS TTS

TTS task objective



Rules

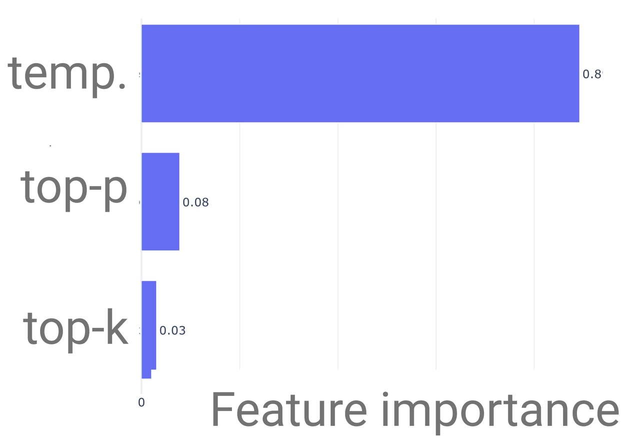
Data: LJSpeech dataset[5]
- train/val/test split provided by the organizer
- 24 hours, English single-speaker corpus

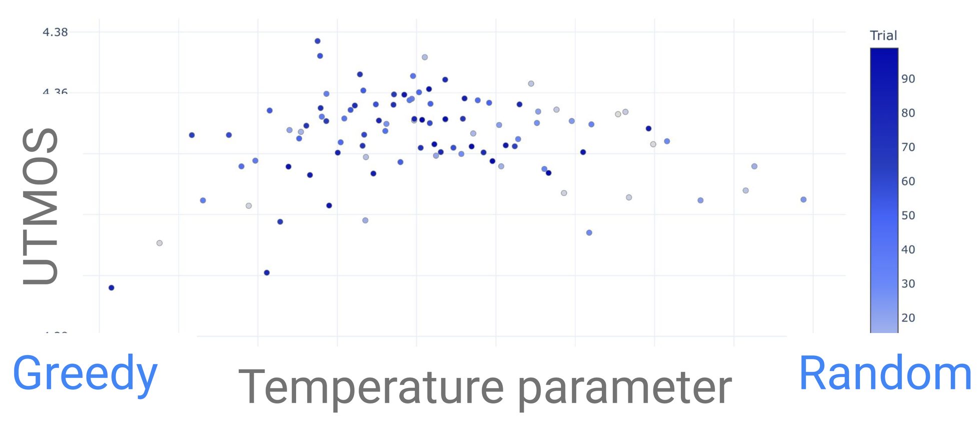
Evaluation metrics
- UTMOS[4] : Predicted Naturalness MOS ≠ Human evaluated
- Bitrate : The bitrate of the discretized speech

Techniques applied for improving UTMOS

Model architecture: Transformer[6] Encoder-decoder model
Vocoder: SMILEY with codebook size of 256, 512, 1024

Hyperparameter tuning for the sampling parameters. top-p, top-k and temperature
Objective: maximize UTMOS on valid set





Results

	Bitrate(↓)	UTMOS(↑)	Rank
baseline (FastSpeech2)	448.3	3.73	9
Ours w/ codebook size of 1024	351.1	4.29	8
Ours w/ codebook size of 512	313.8	**4.36**	**1**
Ours w/ codebook size of 256	**277.6**	4.33	2
Ground truth	-	4.43	-

UTDUSS is comparable to ground truth in terms of UTMOS

References

[1] X. Chang et al,. Interspeech, 2024
[2] R. Kumar et al,. NeurIPS, 2023.
[3] T. NGuyen et al., Interspeech 2023.
[4] T. Saeki et al., Interspeech, 2022.
[5] K. Ito et al., 2017
[6] A. Vaswani et al, NeurIPS 2017