



TOWARDS UNIVERSAL SPEECH DISCRETE TOKENS: A CASE STUDY FOR ASR AND TTS

Yifan Yang^{,1}, Feiyu Shen^{*,1}, Chenpeng Du¹, Ziyang Ma¹, Kai Yu¹, Daniel Povey^{†,2}, Xie Chen^{†,1}*

¹ MoE Key Lab of Artificial Intelligence, AI Institute

X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China

² Xiaomi Corporation, Beijing, China

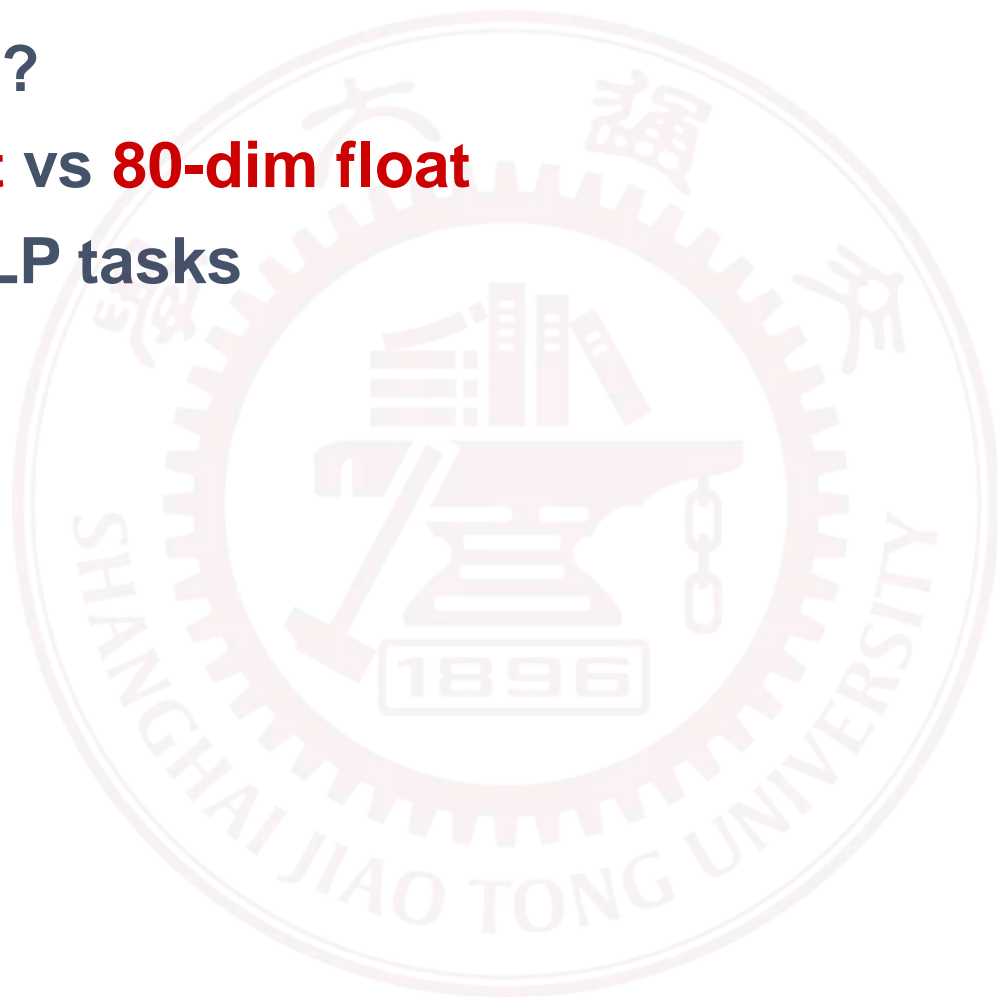
yifanyeung@sjtu.edu.cn

2024.04.17

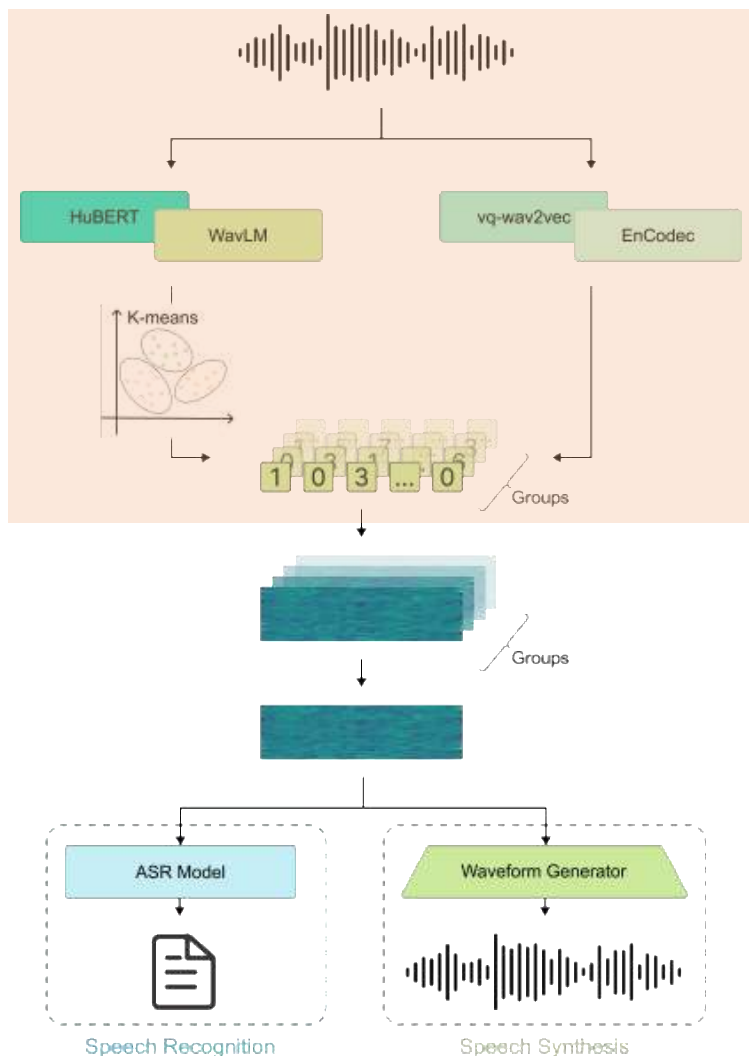
Motivation

Why speech discrete tokens?

- Lower bandwidth: **1 int** vs **80-dim float**
- Akin to BPE used in NLP tasks



Method at a Glance



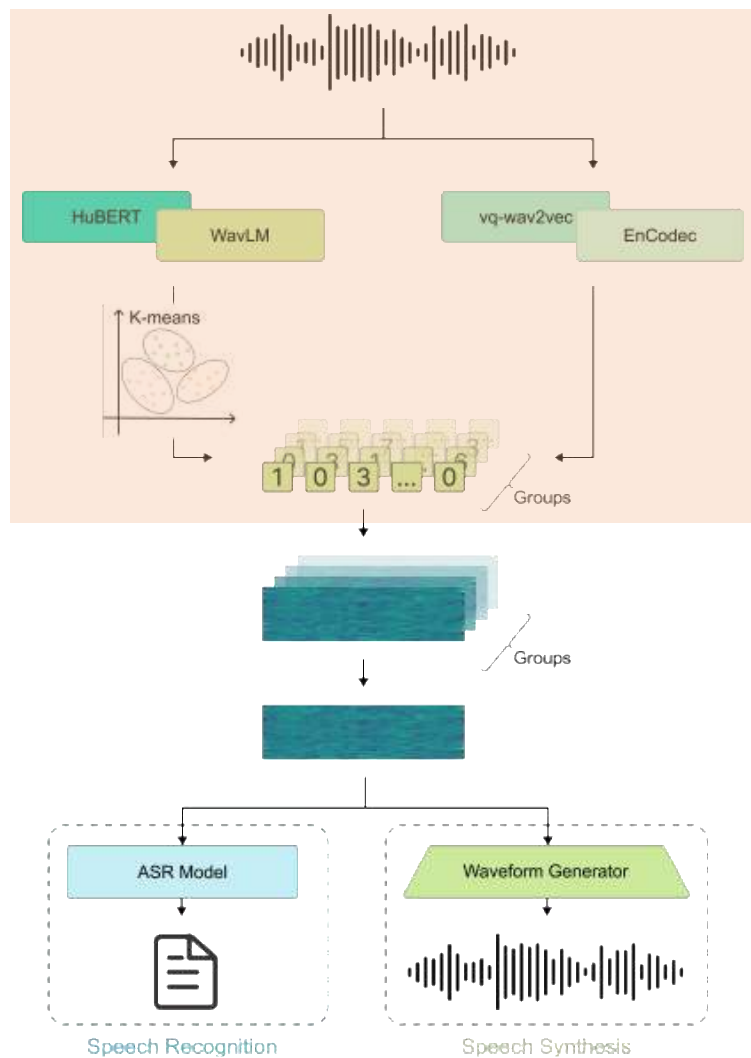
Semantic and acoustic discrete tokens from speech:

- Speech-based SSL models
 - e.g. **vq-wav2vec**; **HuBERT**; **WavLM**
- Neural Codec
 - e.g. **EnCodec**; **DAC**

Two way to discretize speech:

- **K-means**
- **VQ**

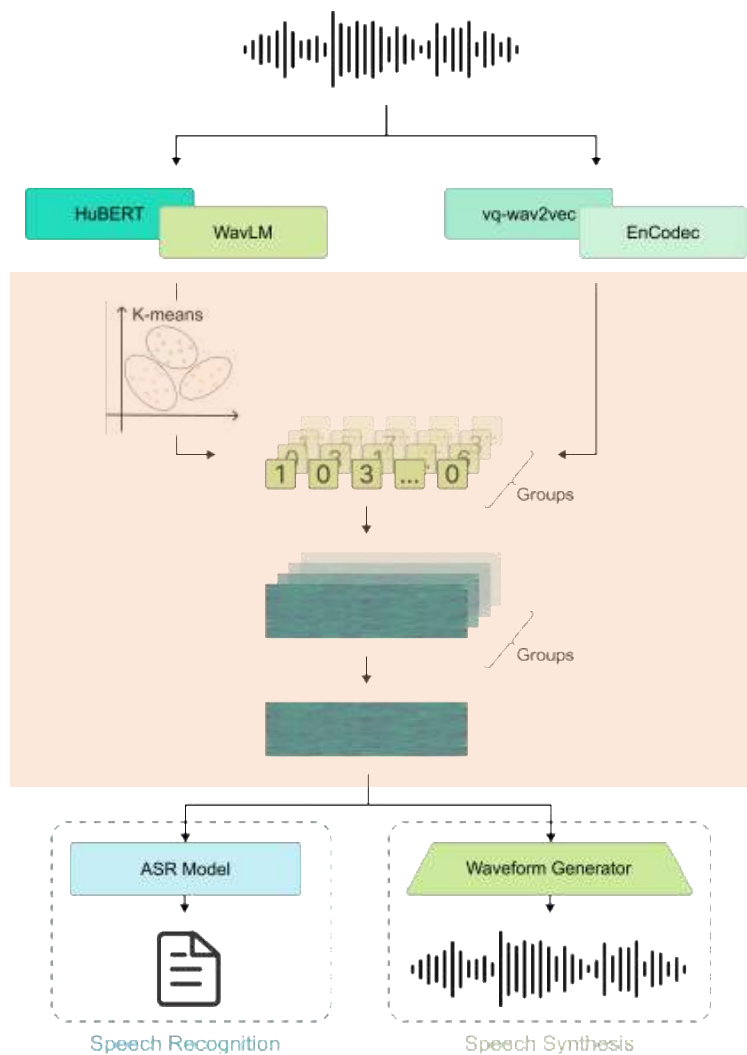
Method at a Glance



Discrete tokens for speech tasks:

- ASR
- TTS

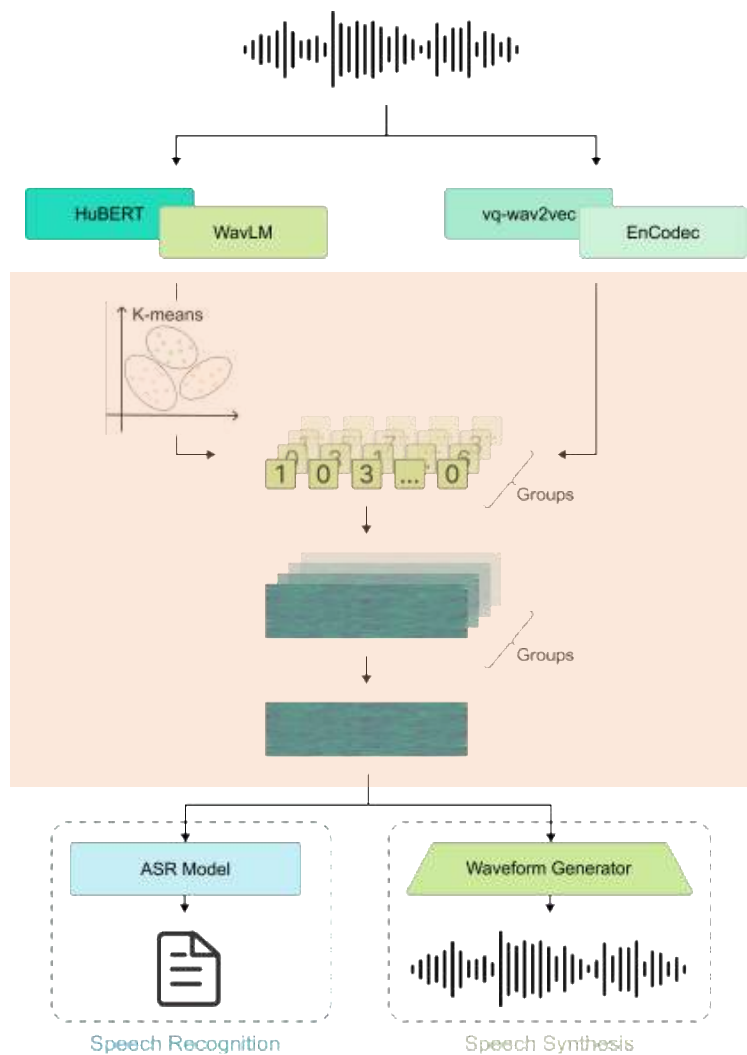
ASR with Discrete Tokens



ASR with Discretized Input

- Random initialized embedding
- If multiple groups exist:
 - Concatenate
 - Linear projection to compress
- Interpolated to a uniform 100 Hz rate

ASR with Discrete Tokens



Discretized Input Augmentation Policy

- Time Warping
- Time Masking
- Embedding Masking
- Gaussian Noise

ASR with Discrete Tokens on LibriSpeech 100h

Method	Feature	# Units	Bandwidth (kbps)	test	
				clean	other
Chang et al. (2023) [12]	FBank	-	256.00	8.30	22.20
	WavLM-Large	2000	0.55	5.90	12.80
Ours	FBank	-	256.00	6.27	16.67
	WavLM-Large	2000	0.55	5.22	11.85
	HuBERT-Large	2000	0.55	5.19	11.94
	EnCodec	1024 ⁸	6.00	7.16	22.04
	vq-wav2vec	320 ²	1.66	11.76	33.08

- Discrete tokens **outperform** FBank on small-scale data
- Discrete tokens from **WavLM** work **best**

ASR with Discrete Tokens on LibriSpeech 960h

Method	Feature	# Units	test	
			clean	other
Chang et al. (2023) [12]	FBank	-	2.60	6.20
	WavLM-Large	2000	3.00	7.00
Ours	FBank	-	2.23	5.15
	WavLM-Large	2000	2.29	5.50
	HuBERT-Large	2000	2.29	5.73
	EnCodec	1024 ⁸	2.57	6.81
	vq-wav2vec	320 ²	3.19	9.44

- Discrete tokens are **competitive** with FBank
- Discrete tokens from **WavLM** still work **best**

ASR on GigaSpeech 1000h & AISHELL-1

Dataset	Feature	# Units	WER	
			dev	test
GigaSpeech	FBank	-	12.24	12.19
	WavLM-Large	2000	13.20	13.05
	HuBERT-Large	2000	14.45	14.79
AISHELL-1	FBank	-	4.27	4.49
	WavLM-Large	2000	8.41	8.83

Two reserved problems:

- **Noise robustness** still need to be improved
- **Language generalization** is highly associated with SSL model

TTS with Discrete Tokens on LibriTTS

Feature	# Units	Bandwidth (kbps)	MOS		SECS
			Naturalness	Similarity	
Ground-truth	-	-	4.48	4.18	0.843
Mel spectrogram	-	256.00	4.36	4.17	0.834
Encodec	1024 ⁸	6.00	3.83	3.85	0.834
DAC	1024 ⁸	4.00	4.41	4.30	0.841
vq-wav2vec	320 ²	1.66	4.36	4.21	0.842
HuBERT Large	2000	0.55	4.26	4.18	0.833
WavLM Large	2000	0.55	4.18	4.18	0.836

- Discrete tokens **outperform** mel-spectrogram features on subject & object metrics
- Discrete tokens from **DAC** work **best**

TTS with Discrete Tokens on LibriTTS: Demo

[Transcription] The investors in the enterprise were ready and anxious to meet the extra cost of putting the wires underground.

- Ground-truth



- Encodec



- DAC



- Mel-spectrogram



- vq-wav2vec



- HuBERT



- WavLM



Recent Advance about **ASR** with Discrete Tokens

Dataset	Benchmark	WER	
		test-clean	test-other
LibriSpeech 960h	FBank	2.21	4.79
LibriSpeech 960h	Discrete Tokens	2.00	4.12

Dataset	Benchmark	WER	
		dev	test
GigaSpeech 1000h	FBank	12.12	12.08
GigaSpeech 1000h	Discrete Tokens	11.24	11.27
GigaSpeech 10000h	FBank	10.31	10.50
GigaSpeech 10000h	Discrete Tokens	10.30	10.53

- Discrete tokens achieve **SOTA** based on our recent advance (68M)
 - LibriSpeech 960h
 - GigaSpeech 1000h & GigaSpeech 10000h

Conclusion

Discrete tokens outperform continuous features in **ASR** and **TTS**

Discrete tokens source matter

- Discrete tokens from **WavLM** work best for **ASR**
- Discrete tokens from **DAC** work best for **TTS**

Generalization of discrete tokens across languages is yet to be improved

Thank You!

Q&A

Feel free to contact me for any question

Yifan Yang

Shanghai Jiao Tong University

yifanyeung@sjtu.edu.cn

2024.04.17

Paper link

