Self-introduction

1. Senior researcher at Microsoft
2. Got PhD at NTU under Prof. Hung-yi Lee[1] in speech processing
3. Publish more than 20 first-author in ICASSP / Interspeech / TASLP / ACL/ ASRU / SLT, etc
4. Main contributor of [2]S3prl (2.2k+ GitHub stars)
5. Internships: Microsoft * 2 (speech and audio generation), Meta * 2 (speech enhancement), Amazon (Model compression), Tencent (Model compression)
6. [3]Google PhD Fellowship (1/75 over the world)
7. Visiting student at Tsinghua University and the Chinese University of Hong Kong

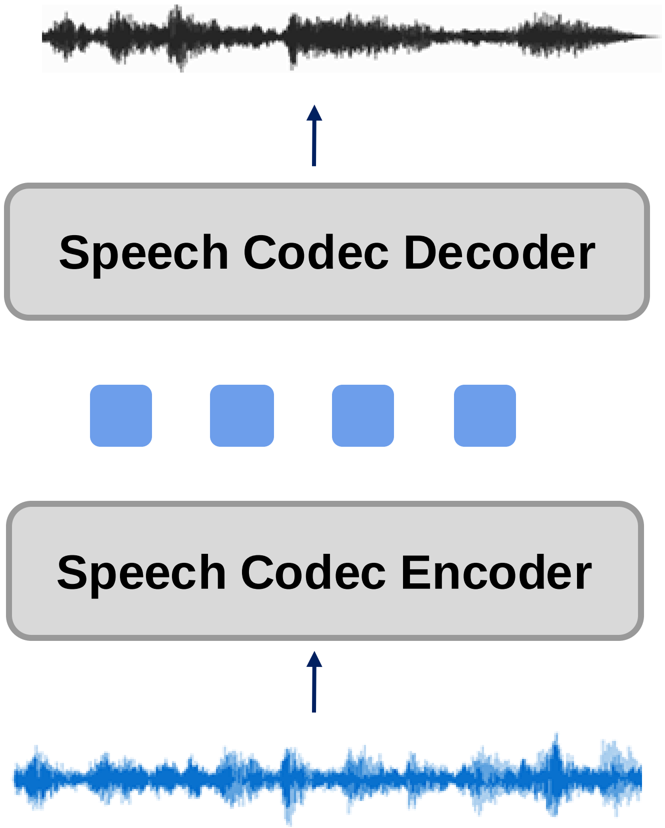
[1]Famous YouTuber

[2]S3prl: a speech self-supervised learning toolkit for all the speech tasks

[3]75 students get the award over the world every year

Neural audio codecs in the era of speech LMs

Neural audio codec - Brief recap



Audio compression

Discrete audio tokens

Neural audio codec models typically consist of an encoder, a vector quantization (VQ) module, and a decoder:

1. Encoder: down-sample the time-domain audio (16k) to extract frame-wise audio features (50 Hz).
2. The VQ module: convert each frame-wise audio feature into discrete tokens.
3. Decoder: reconstruct the time-domain audio signal from the discrete tokens.

Text language models -> speech language models

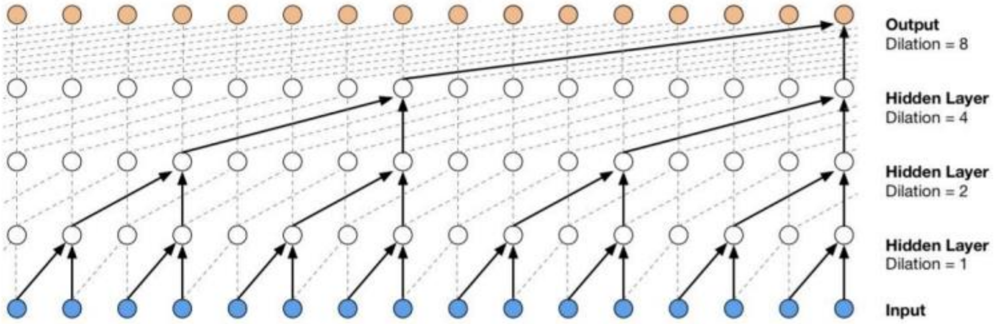
- Language modeling is successful in NLP domain
- Speech contains more information than text

https://www.youtube.com/watch?v=Zjpl84KCTvw&list=PLJV_el3uVTsNnC37JYD8kBcNDI7CNJgum&index=1

Next token prediction on raw audio signals - Wavenet

Pros:
- Can produce high-quality speech

Cons:
- Does not learn high-level structure (random babble)

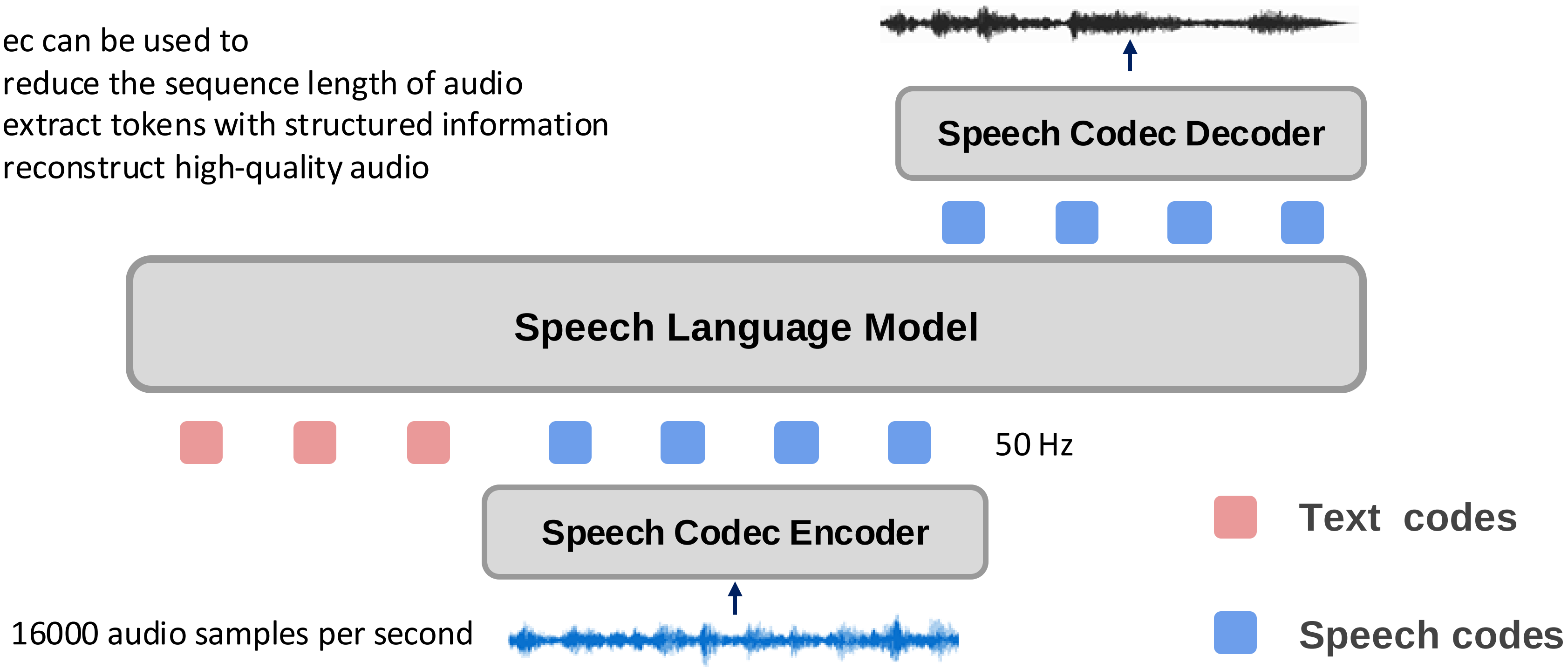


https://arxiv.org/abs/1609.03499
https://www.youtube.com/watch?v=Zjpl84KCTvw&list=PLJV_el3uVTsNnC37JYD8kBcNDI7CNJgum&index=1

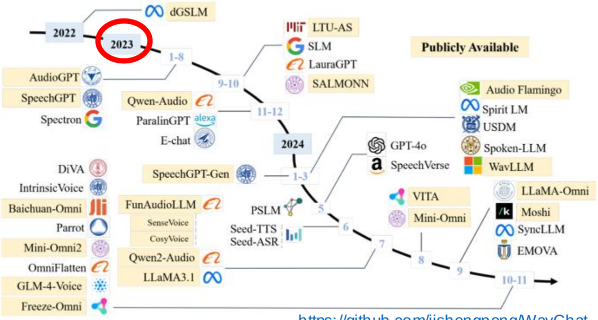
Codec codes for speech language modeling

Codec can be used to

- reduce the sequence length of audio
- extract tokens with structured information
- reconstruct high-quality audio

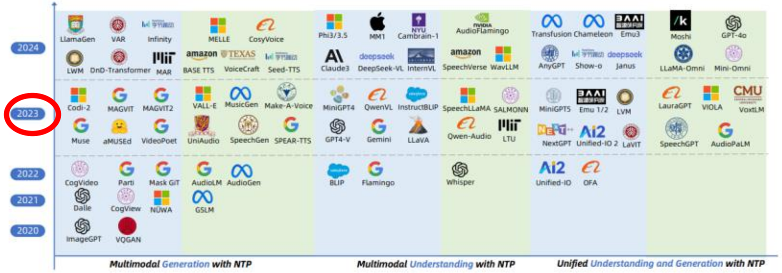


Numerous speech LMs have been proposed



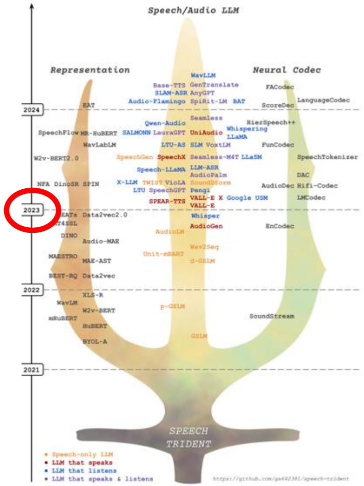
https://github.com/jishengpeng/WavChat

Chat model survey



https://arxiv.org/pdf/2412.18619

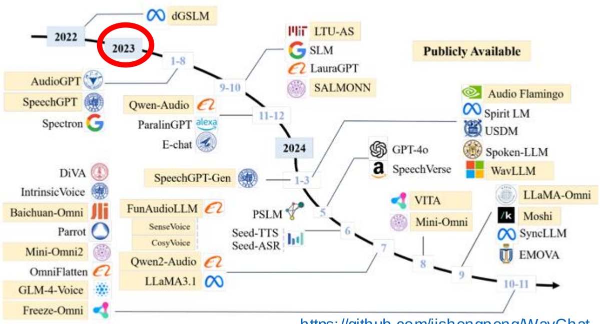
Multi-modal next token prediction survey



https://github.com/ga642381/speech-trident

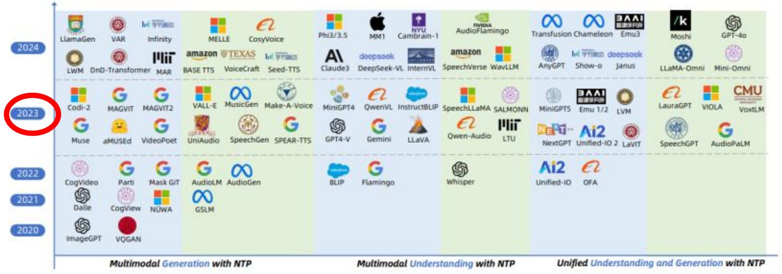
Speech trident survey

Tons of neural audio codecs come into stage



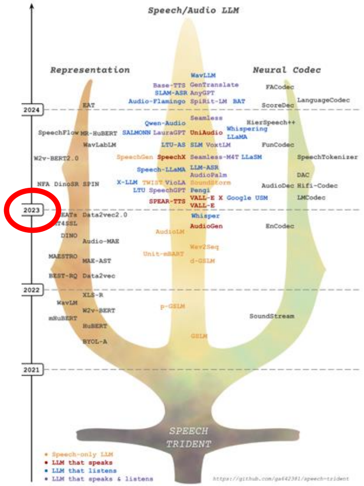
https://github.com/jishengpeng/WavChat

Chat model survey



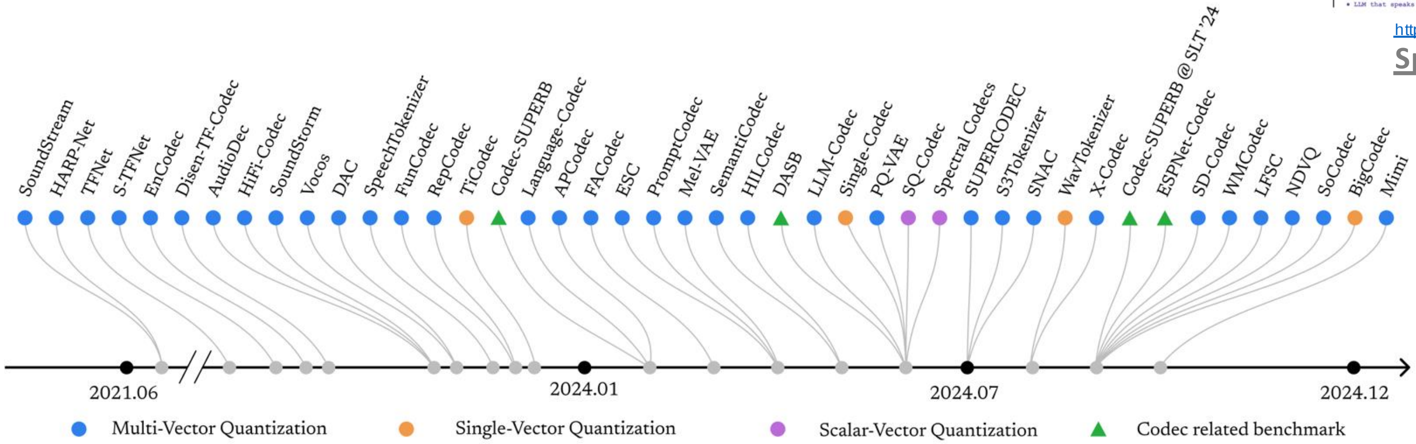
https://arxiv.org/pdf/2412.18619

Multi-modal next token prediction survey

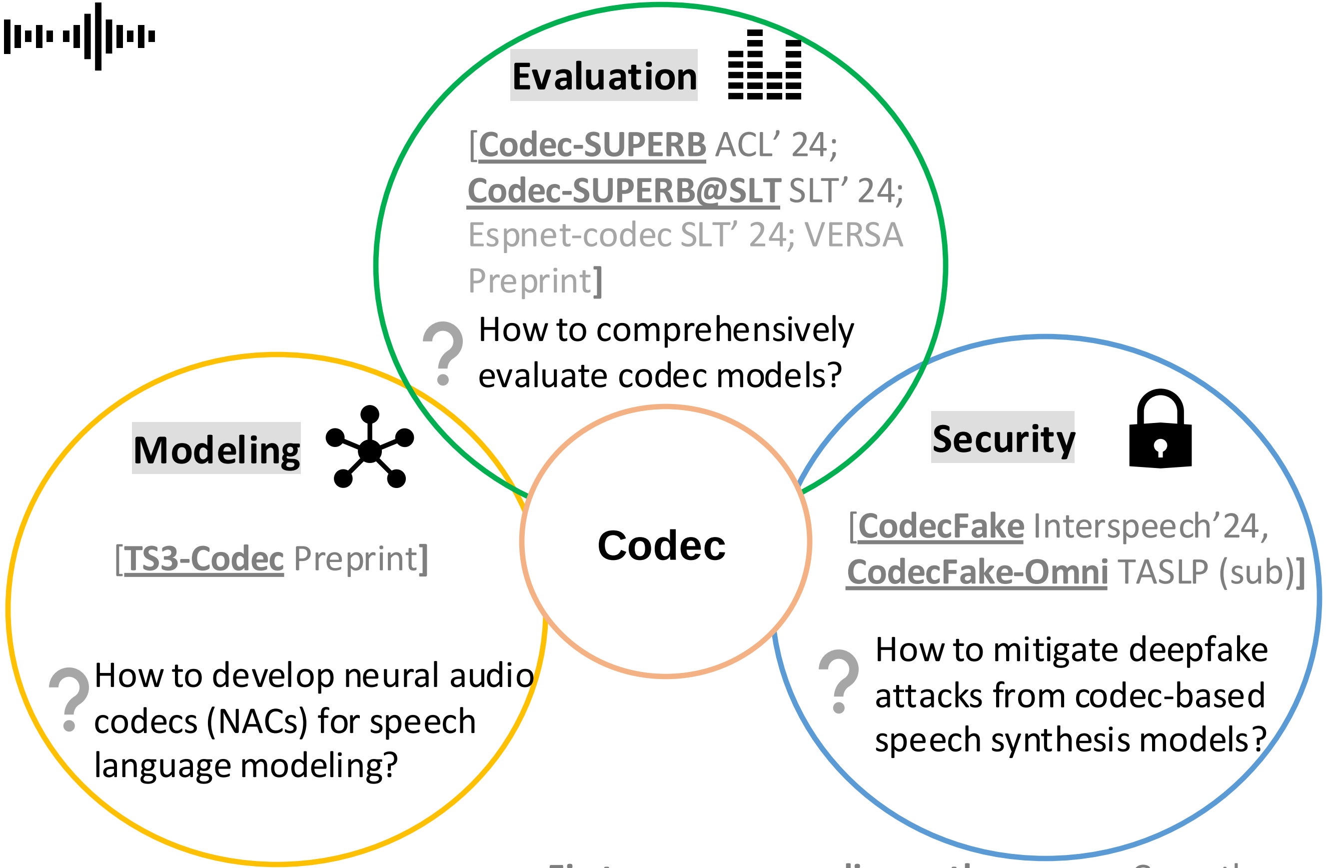


https://github.com/ga642381/speech-trident

Speech trident survey



Research roadmap and open questions



First- or corresponding-author paper Co-author paper

Research roadmap and open questions

Modeling

[TS3-Codec Preprint]

? How to develop neural audio codecs (NACs) for speech language modeling?

Codec

First- or corresponding-author paper Co-author paper

TS3-Codec: Transformer-Based Simple Streaming Single Codec

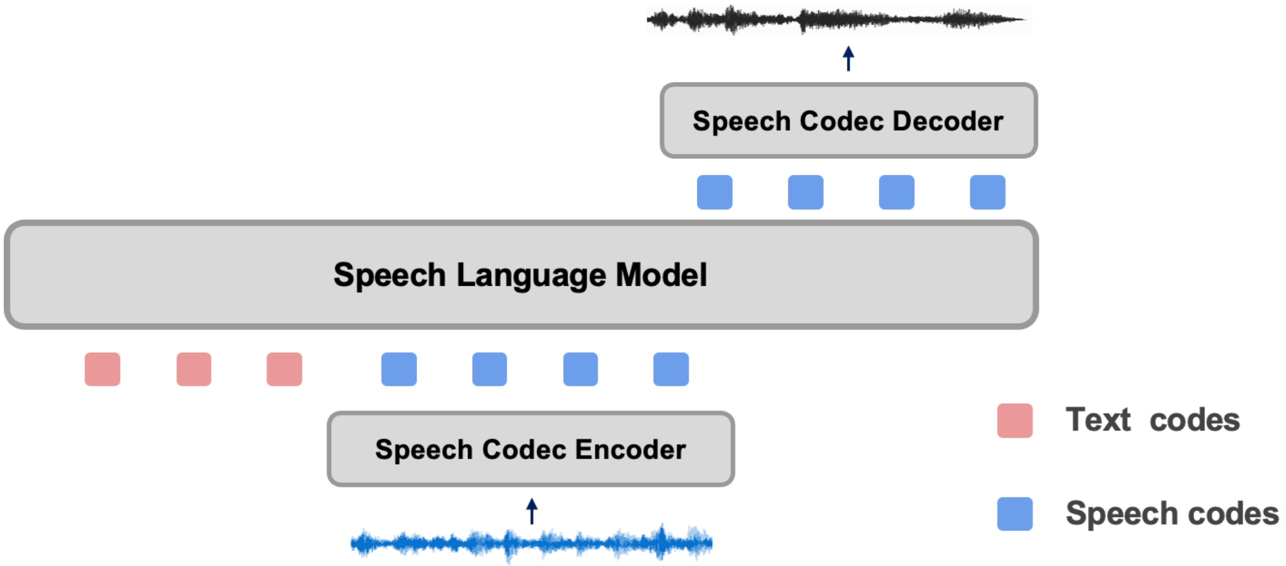
Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, Jinyu Li

Microsoft, USA

haibinwu@microsoft.com

TL;DR: The first attempt to develop a convolution-free, transformer-only NAC.

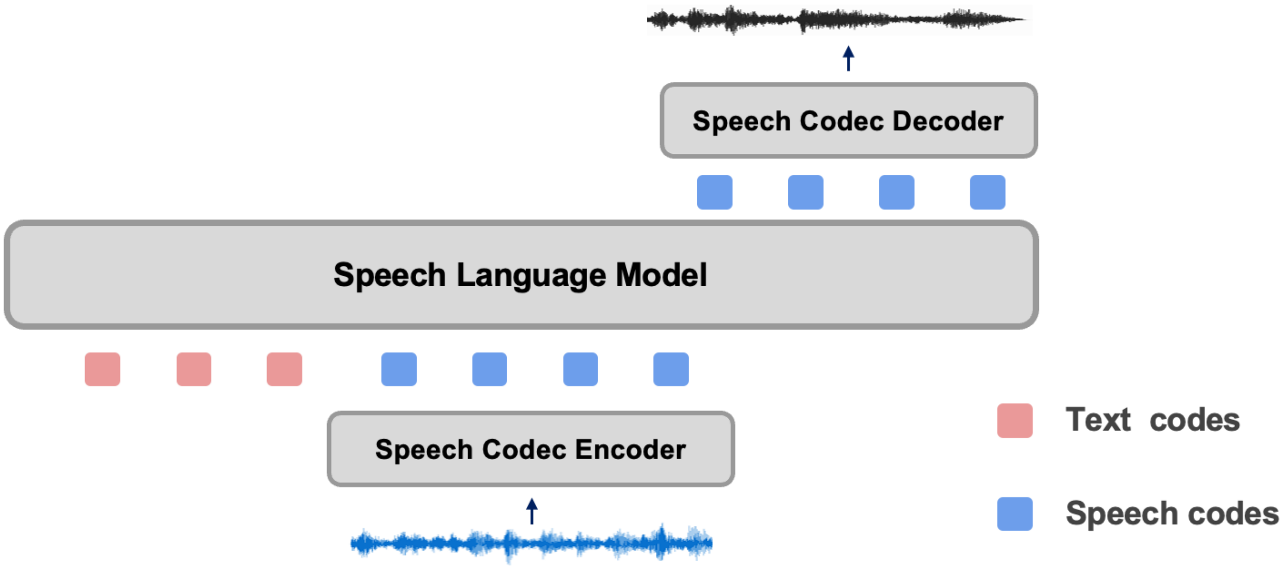
Recap the functionality of codec in speech LMs



- Listen to the speech - Frontend
- Speak out the response - Backend

Codec should have some good properties to support the speech LM

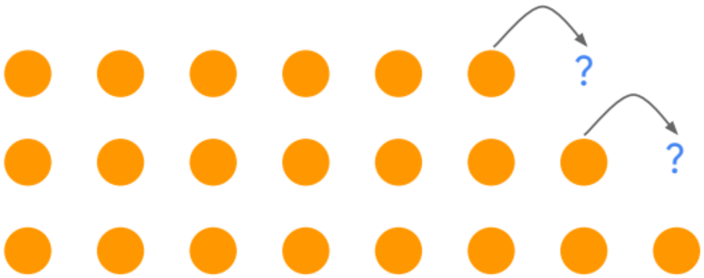
Good property for speech LM - Low computation



<u>Low-computation</u> NACs enable

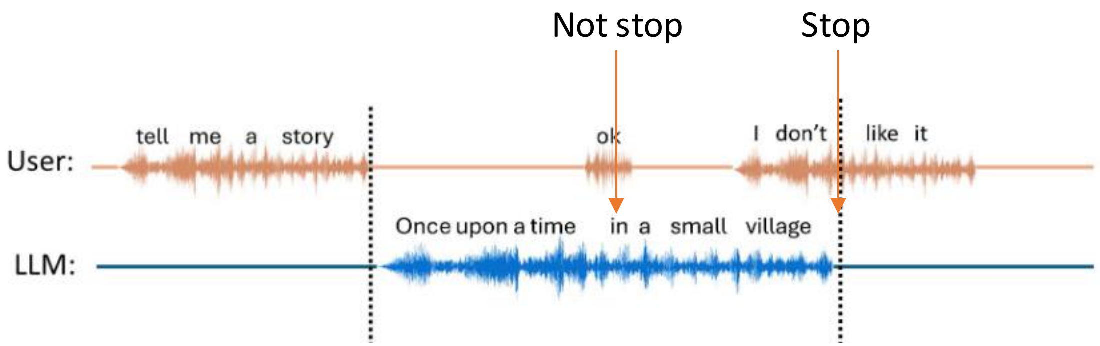
- fast encoding and decoding
- reducing computational costs and leaving more computation resources available for SLMs

Good property for speech LM - Low token rate



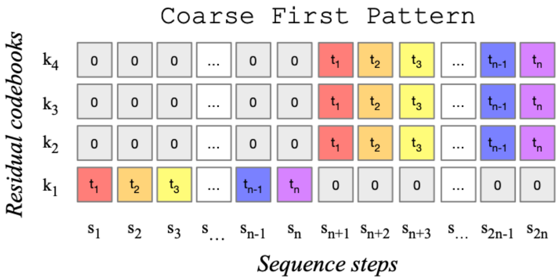
<u>Low token rate</u>:
Long sequences generally make LLM training slow and unstable. Therefore, it is preferable to use low-token-rate NAC models for SLM.

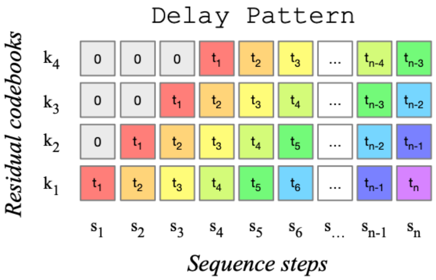
Good property for speech LM - Streaming

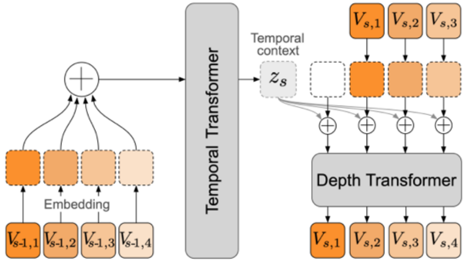


- Full-duplex communication, where users and machines can speak and respond simultaneously, is a popular and ongoing challenge in the SLM field.
- The speech LM should listen and speak in the same time with fast responses.
- To enable seamless real-time interactions, the codec should support streaming processing, allowing it to encode user speech and generate speech response with low latency.

Good property for speech LM - Single-Codebook







A single codebook-based model is preferable to a multiple-codebook-based model, because the latter introduces additional complexity to the architecture of SLMs:

- The combination of auto-regressive and non-autoregressive models (Valle)
- Delay pattern (MusicGEN)
- the temporal and depth transformers (uniaudio)

https://arxiv.org/pdf/2306.05284
https://arxiv.org/pdf/2410.00037

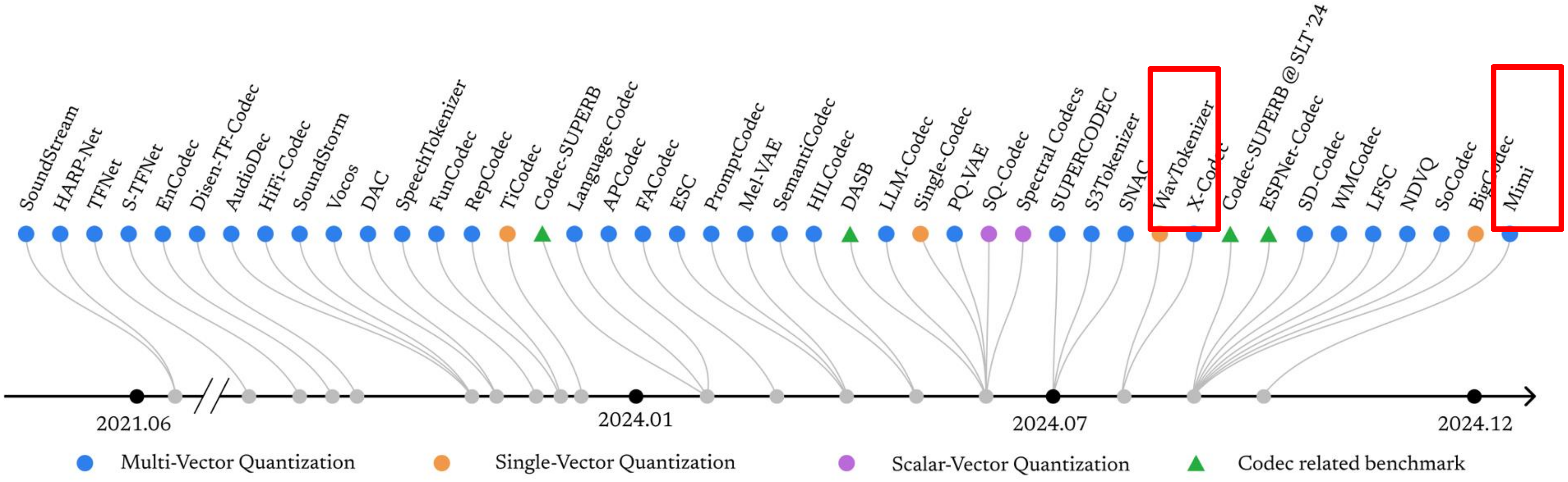
Limitations of current neural audio codecs

1. High token rate (long token sequence)
- e.g., 6kbps Descript audio codec has 600 tokens per second
- Make auto-regressive modeling challenging and computational expensive

2. Poor reconstruction quality at low bit rate (e.g., 1 kbps)
- Most previous studies work on bit rate > 2kbps
- Can we go further under 1.0 kbps?

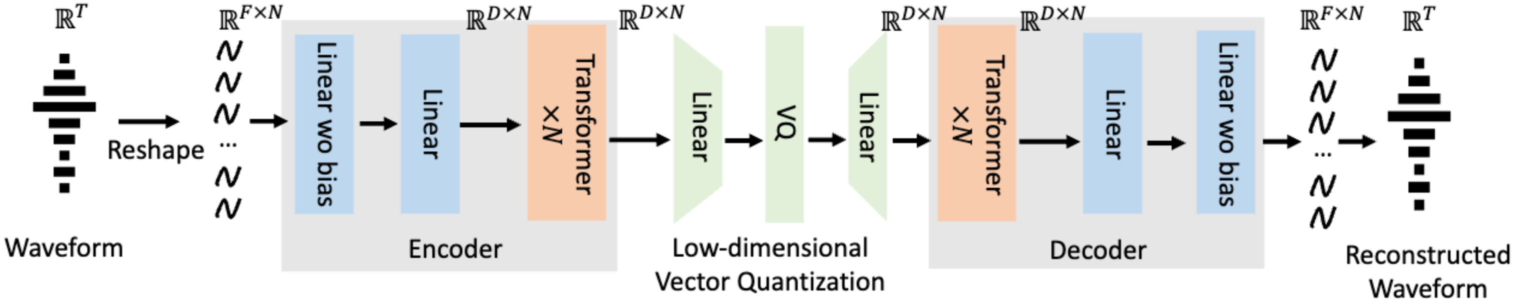
3. Less works explore the streaming capacity of codecs

Few research of Transformers in NAC domain



1. Most NAC models rely on CNNs as the dominant architecture, with only a few incorporating transformers as intermediate layers within the CNN encoder-decoder framework.
2. However, the performance of a purely transformer-based and convolution-free architecture in NACs remains unexplored.

TS3-Codec



Architecture

- WaveNeXt frond and back-ends[1] [1]
- Transformer-only architecture[2] with self-attention left fixed window
- Reduce the codebook embedding dimension to enlarge codebook usage [3]
- Enlarge the codebook size to 65k / 130k

[1]Better performance than Vocos [2]
[2]Reason 1: Low computation. E.g. TS3-Codec (**1.6B** paras): 60.52G MACs, while BigCodec (**160M** paras): 61.1G MACs
[2]Reason 2: Transformers offer **simplicity** in model design

[1] Okamoto, Takuma, et al. "WaveNeXt: ConvNeXt-based fast neural vocoder without iSTFT layer." *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023.
[2] Siuzdak, Hubert. "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis." arXiv preprint arXiv:2306.00814 (2023).
[3] Kumar, Rithesh, et al. "High-fidelity audio compression with improved rvqgan." Advances in Neural Information Processing Systems 36 (2024).

Why Transformer rather than CNNs

1. CNNs are well-known for their parameter efficiency and reusability. On the other hand, for similar parameter sizes, CNNs typically require significantly more computation than transformers.
2. Convolutions have inherent biases. Convolutions apply fixed weighted-sum weights across all intermediate feature maps across different time stamps.
3. Transformers offer simplicity in model design. CNNs require careful tuning of kernels and up- and down-sampling mechanisms due to their inherent biases.

Experiments - Baselines

Table 1: *Comparison between baseline codecs.* ***SEM*** *represents semantic distillation.* ***RVQ*** *and* ***single*** *means residual and single vector quantization, respectively.* ***SA*** *means self-attention.*

Codec	**SEM**	**Streaming**	**VQ type**	**Architecture**
Encodec [15]	✗	✓	RVQ	Conv + LSTM
DAC [36]	✗	✗	RVQ	Conv
SpeechTokenizer [37]	✓	✗	RVQ	Conv + LSTM
Mimi [11]	✓	✓	RVQ	Conv + Transformer
BigCodec [24]	✗	✗	Single	Conv + LSTM
WavTokenizer [10]	✗	✗	Single	Conv + LSTM + SA

Experiments - 1000bps non-streaming

	Properties					Complexity		Intelligibility		Distortion			Naturalness
Model Tag	**Streaming**	**Bitrate**	**Codebook Layer**	**Frame Rate**	**Token Rate**	**MACs**	**Paras**	**WER↓**	**STOI↑**	**PESQ↑**	**MCD↓**	**SPK-SIM↑**	**UTMOS↑**
Ground Truth	-	-	-	-	-	-	-	2.0	1.000	4.64	0.00	1.00	4.09
DAC (A)	✗	1500	2	75	150	55.6G	74.1M	7.2	0.829	1.48	4.83	0.47	1.68
SpeechTokenizer (B1)	✗	1000	2	50	100	17.1G	103.7M	3.9	0.768	1.21	6.30	0.33	2.32
BigCodec (C)	✗	1040	1	80	80	67.1G	159.4M	2.8	0.935	2.68	3.01	0.84	4.11
WavTokenizer (D1)	✗	975	1	75	75	6.3G	80.6M	6.8	0.886	2.05	4.00	0.59	3.89
Encodec (E1)	✓	1500	2	75	150	5.6G	14.9M	4.9	0.845	1.56	4.32	0.60	1.58
Mimi (F1)	✓	1100	8	12.5	100	8.1G	79.3M	**3.0**	0.905	2.22	3.81	**0.73**	3.60
BigCodec-S (G1)	✓	1040	1	80	80	7.1G	21.8M	4.6	0.888	1.96	3.80	0.56	3.41
BigCodec-S (G2)	✓	1040	1	80	80	61.1G	159.9M	3.8	0.906	2.17	3.52	0.65	3.73
TS3-Codec (X1)	✓	800	1	50	50	7.6G	203.6M	3.6	0.909	2.22	3.52	0.68	**3.85**
TS3-Codec (X2)	✓	850	1	50	50	7.6G	203.6M	3.6	**0.910**	**2.23**	**3.50**	0.68	3.84

Among the four non-streaming baselines, BigCodec (C) demonstrates the best performance at approximately 1000 bps, surpassing other codec models by a significant margin across all metrics.

Used for designing a streaming CNN codec baseline:

1. CNN-based model
2. The state-of-the-art single-codebook codec

Experiments - 1000bps non-streaming

Model Tag	Streaming	Bitrate	Codebook Layer	Frame Rate	Token Rate	MACs	Paras	WER↓	STOI↑	PESQ↑	MCD↓	SPK-SIM↑	UTMOS↑
	Properties					Complexity		Intelligibility		Distortion			Naturalness
Ground Truth	-	-	-	-	-	-	-	2.0	1.000	4.64	0.00	1.00	4.09
DAC (A)	✗	1500	2	75	150	55.6G	74.1M	7.2	0.829	1.48	4.83	0.47	1.68
SpeechTokenizer (B1)	✗	1000	2	50	100	17.1G	103.7M	3.9	0.768	1.21	6.30	0.33	2.32
BigCodec (C)	✗	1040	1	80	80	67.1G	159.4M	2.8	0.935	2.68	3.01	0.84	4.11
WavTokenizer (D1)	✗	975	1	75	75	6.3G	80.6M	6.8	0.886	2.05	4.00	0.59	3.89
Encodec (E1)	✓	1500	2	75	150	5.6G	14.9M	4.9	0.845	1.56	4.32	0.60	1.58
Mimi (F1)	✓	1100	8	12.5	100	8.1G	79.3M	**3.0**	0.905	2.22	3.81	**0.73**	3.60
BigCodec-S (G1)	✓	1040	1	80	80	7.1G	21.8M	4.6	0.888	1.96	3.80	0.56	3.41
BigCodec-S (G2)	✓	1040	1	80	80	61.1G	159.9M	3.8	0.906	2.17	3.52	0.65	3.73
TS3-Codec (X1)	✓	800	1	50	50	7.6G	203.6M	3.6	0.909	2.22	3.52	0.68	**3.85**
TS3-Codec (X2)	✓	850	1	50	50	7.6G	203.6M	3.6	**0.910**	**2.23**	**3.50**	0.68	3.84

WavTokenizer (D1) achieves good UTMOS scores, as highlighted in their paper, where they emphasize that reconstructed utterances from their models have strong naturalness.

Experiments - 1000bps streaming

	Properties					Complexity		Intelligibility		Distortion			Naturalness
Model Tag	**Streaming**	**Bitrate**	**Codebook Layer**	**Frame Rate**	**Token Rate**	**MACs**	**Paras**	**WER↓**	**STOI↑**	**PESQ↑**	**MCD↓**	**SPK-SIM↑**	**UTMOS↑**
Ground Truth	-	-	-	-	-	-	-	2.0	1.000	4.64	0.00	1.00	4.09
DAC (A)	✗	1500	2	75	150	55.6G	74.1M	7.2	0.829	1.48	4.83	0.47	1.68
SpeechTokenizer (B1)	✗	1000	2	50	100	17.1G	103.7M	3.9	0.768	1.21	6.30	0.33	2.32
BigCodec (C)	✗	1040	1	80	80	67.1G	159.4M	2.8	0.935	2.68	3.01	0.84	4.11
WavTokenizer (D1)	✗	975	1	75	75	6.3G	80.6M	6.8	0.886	2.05	4.00	0.59	3.89
Encodec (E1)	✓	1500	2	75	150	5.6G	14.9M	4.9	0.845	1.56	4.32	0.60	1.58
Mimi (F1)	✓	1100	8	12.5	100	8.1G	79.3M	**3.0**	0.905	2.22	3.81	**0.73**	3.60
BigCodec-S (G1)	✓	1040	1	80	80	7.1G	21.8M	4.6	0.888	1.96	3.80	0.56	3.41
BigCodec-S (G2)	✓	1040	1	80	80	61.1G	159.9M	3.8	0.906	2.17	3.52	0.65	3.73
TS3-Codec (X1)	✓	800	1	50	50	7.6G	203.6M	3.6	0.909	2.22	3.52	0.68	**3.85**
TS3-Codec (X2)	✓	850	1	50	50	7.6G	203.6M	3.6	**0.910**	**2.23**	**3.50**	0.68	3.84

- TS3-Codec models perform the best for STOI, PESQ, MCD, UTMOS, and the second best for WER and SPK-SIM.
- Mimi performs the best for WER, probably because of their inclusion of semantic distillation.

Experiments - 600bps

Model Tag	Streaming	Bitrate	Codebook Layer	Frame Rate	Token Rate	MACs	Paras	WER↓	STOI↑	PESQ↑	MCD↓	SPK-SIM↑	UTMOS↑
	Properties					Complexity		Intelligibility		Distortion			Naturalness
Ground Truth	-	-	-	-	-	-	-	2.0	1.000	4.64	0.00	1.00	4.09
SpeechTokenizer (B2)	✗	500	1	50	50	17.1G	103.7M	4.9	0.675	1.12	8.38	0.17	1.34
WavTokenizer (D2)	✗	520	1	40	40	3.4G	80.9M	8.0	0.868	1.88	4.32	0.57	3.77
Encodec (E2)	✓	750	1	75	75	5.6G	14.9M	29.0	0.770	1.23	5.66	0.25	1.25
Mimi (F2)	✓	687.5	5	12.5	62.5	8.1G	79.3M	**4.0**	0.872	1.82	4.40	0.58	3.27
BigCodec-S (G3)	✓	640	1	40	40	4.6G	21.8M	5.9	0.870	1.78	4.16	0.50	3.20
BigCodec-S (G4)	✓	640	1	40	40	39.6G	160.5M	5.4	0.889	1.96	3.97	0.58	3.68
TS3-Codec (X3)	✓	640	1	40	40	6.2G	204.4M	4.9	0.893	2.01	3.81	0.61	3.69
TS3-Codec (X4)	✓	680	1	40	40	6.2G	204.4M	4.5	**0.897**	**2.06**	**3.75**	**0.63**	**3.73**

- TS3-Codec models achieve the best performance in STOI, PESQ, MCD, SPK-SIM, and UTMOS, while securing the second-best WER
- TS3-Codec also outperforms the two non-causal baselines across all metrics.
- SpeechTokenizer performs poorly in most metrics, but its WER is relatively decent. Upon listening, some male voices are distorted to sound like female robotic speech, yet the content remains intelligible.

Experiments - TS3-Codec vs BigCodec

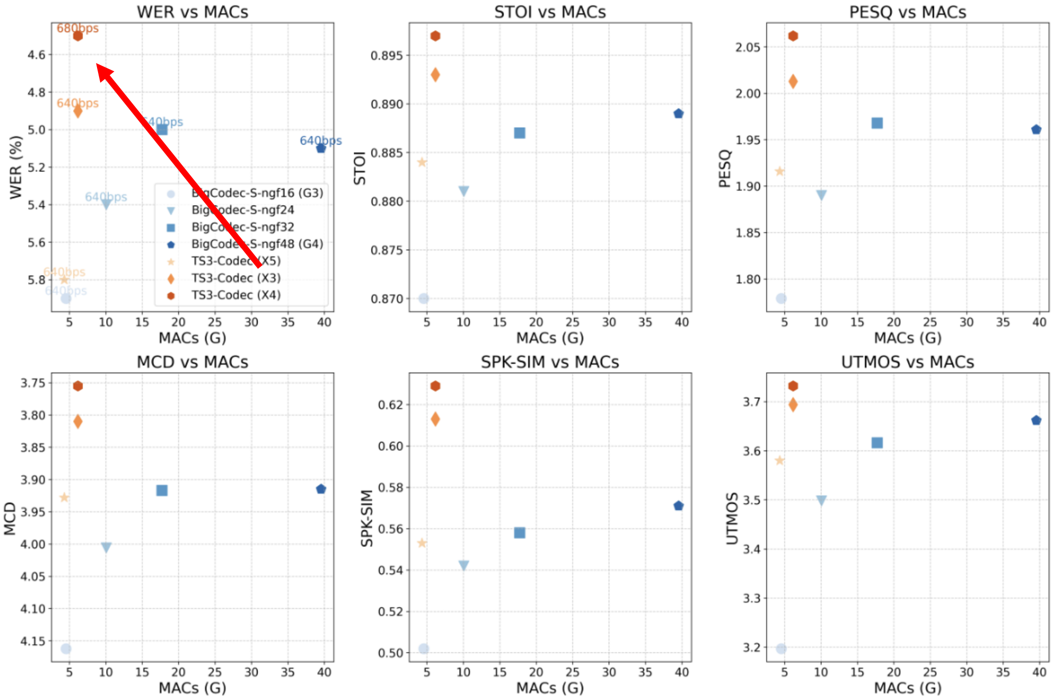


Figure 3: *Comparison between BigCodec-S and TS3-Codec (Bitrate ≈ 600 bps). To enhance visualization, the y-axes for WER and MCD are inverted, so that model points in the upper-left corner exhibit the best performance with the least computational cost.*

- Cold color ones: BigCodec-S
- Warm color ones: TS3-Codec

Experiments - TS3-Codec vs BigCodec

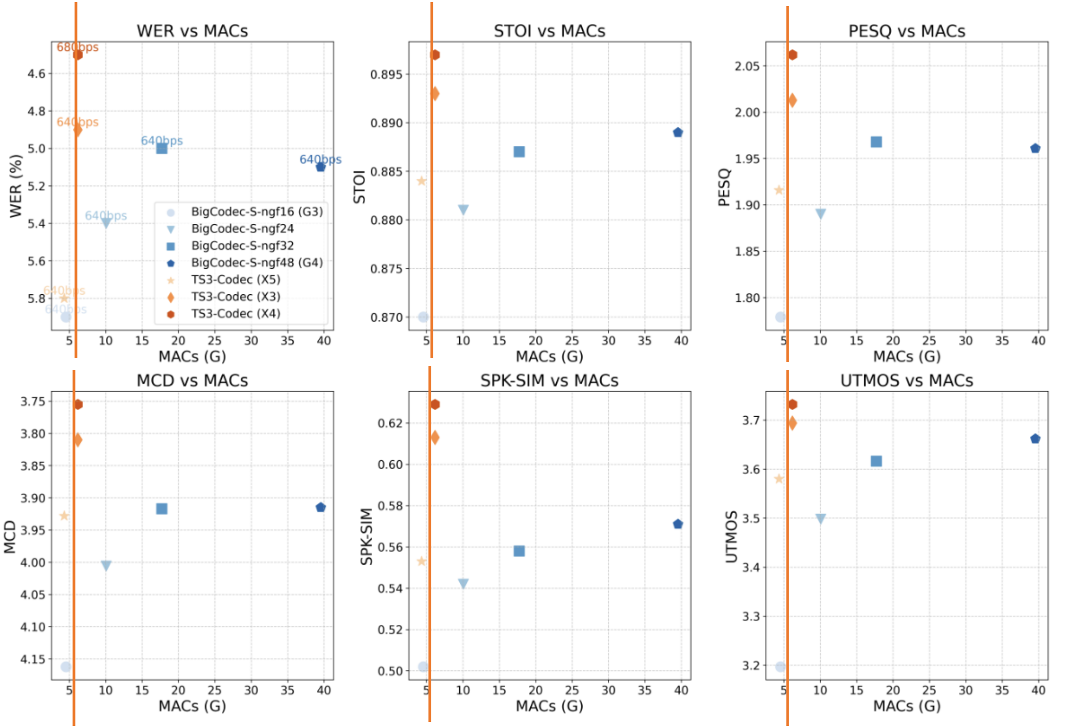


Figure 3: *Comparison between BigCodec-S and TS3-Codec (Bitrate ≈ 600 bps). To enhance visualization, the y-axes for WER and MCD are inverted, so that model points in the upper-left corner exhibit the best performance with the least computational cost.*

- Under similar computational budgets, TS3-Codec always outperforms BigCodec-S significantly across all metrics

Experiments - TS3-Codec vs BigCodec

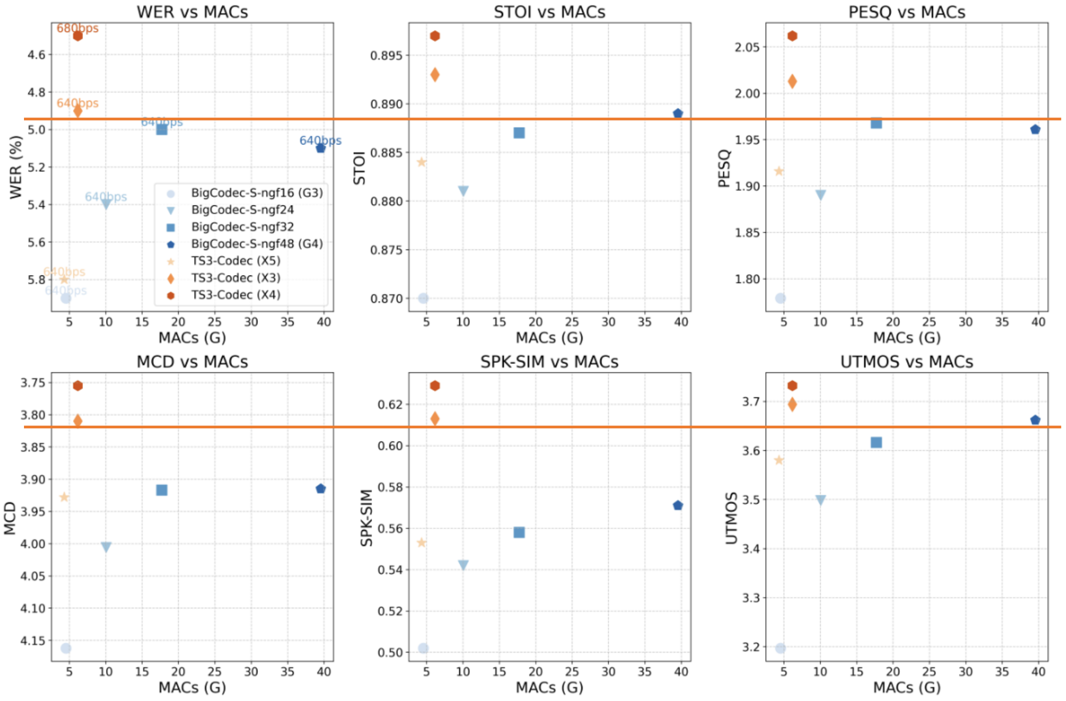
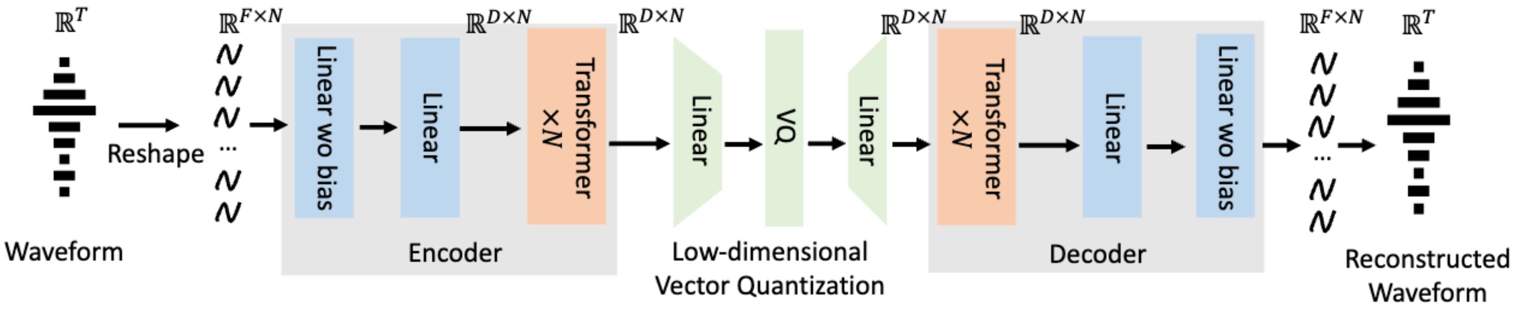


Figure 3: *Comparison between BigCodec-S and TS3-Codec (Bitrate ≈ 600 bps). To enhance visualization, the y-axes for WER and MCD are inverted, so that model points in the upper-left corner exhibit the best performance with the least computational cost.*

- TS3-Codec achieves comparable or better performance to BigCodec with significantly less computation.

Summary: TS3-Codec



Properties

1. Streaming[1]
2. Low computation[2]
3. Single codebook
4. Low token rate (bitrate) [3]

Advantages

Full duplex speech LMs
Save computation for speech LMs
Avoid complicated speech LM decoding structures
Easy the speech LM training

[1]Fixed left context window for self-attention
[2]TS3-Codec (**1.6B** paras): 60.52G MACs, while convolutional based BigCodec (**160M** paras): 61.1G MACs
[3]Bitrate=0.6k, token rate= 40

Research roadmap and open questions

Evaluation

[**Codec-SUPERB** ACL’ 24;
Codec-SUPERB@SLT SLT’ 24;
Espnet-codec SLT’ 24; VERSA Preprint]

? How to comprehensively evaluate codec models?

Codec

First- or corresponding-author paper Co-author paper

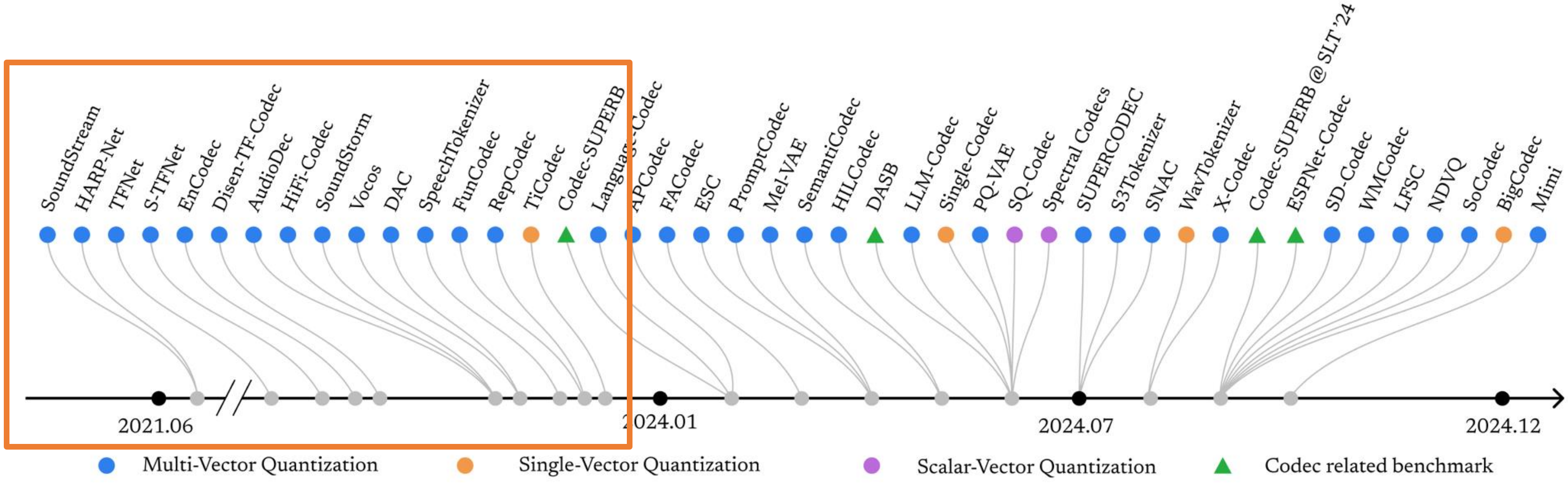
Codec-SUPERB: An In-Depth Analysis of Sound Codec Models

Haibin Wu[1] *, Ho-Lam Chung[1,3] *, Yi-Cheng Lin[1] †, Yuan-Kuei Wu[1] †, Xuanjun Chen[1] †, Yu-Chi Pai[1], Hsiu-Hsuan Wang[1], Kai-Wei Chang[1], Alexander H. Liu[2], Hung-yi Lee[1]

[1]National Taiwan University
[2]Massachusetts Institute of Technology
[3]ASUS Intelligent Cloud Services
hungyilee@ntu.edu.tw

TL;DR: The first benchmark to evaluate codec models from both signal- and application-level perspectives.

Codec-SUPERB - Motivation



- Great developments of codecs (transmission; codec-based LMs)
- Codec models are only evaluated on authors' selected settings

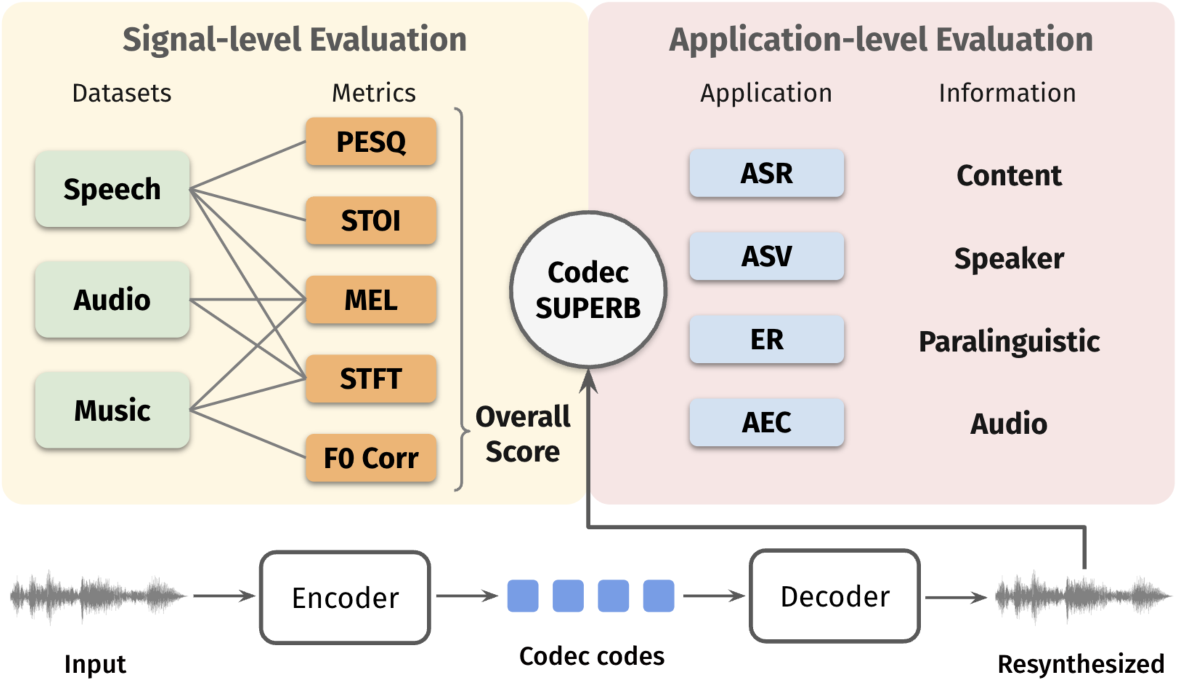
Codec-SUPERB - Motivation

- Only signal-level evaluation is conducted for codecs in previous papers

Signal-level evaluation	**Not enough for** →	Information preservation
STOI (Short-Time Objective Intelligibility)		Speaker information
PESQ (Perceptual Evaluation of Speech Quality)		Content information
Signal-to-noise ratio		Paralinguistic information
…		…

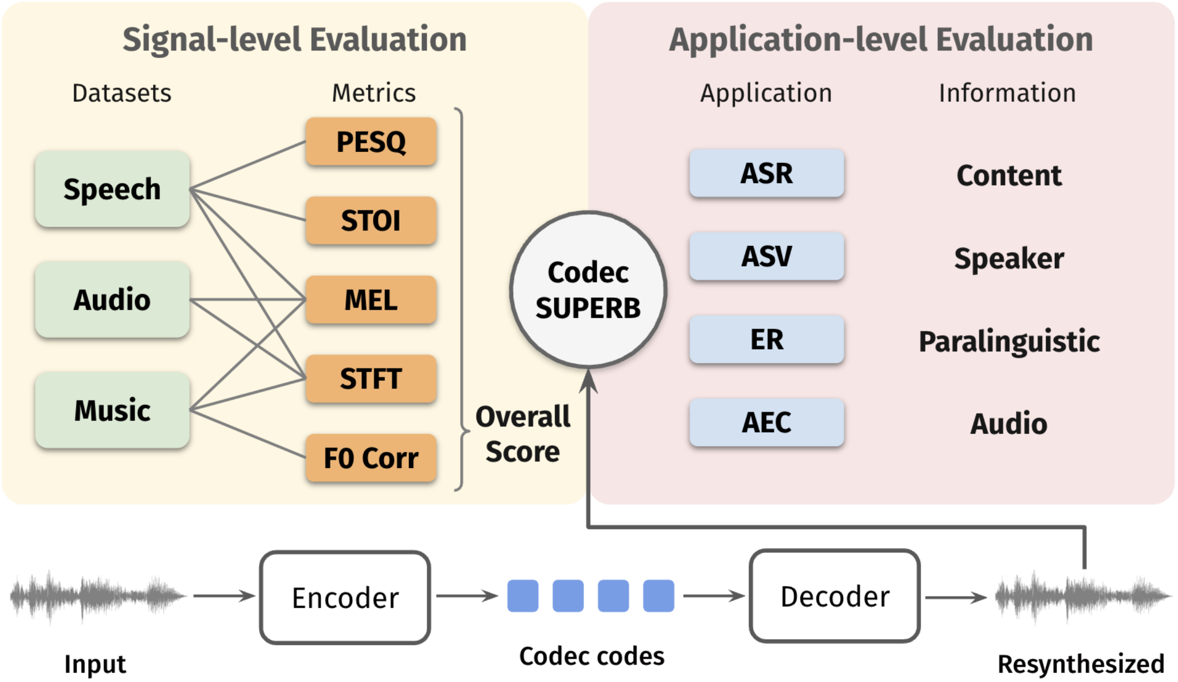
We need application-level evaluation

Unified evaluation framework



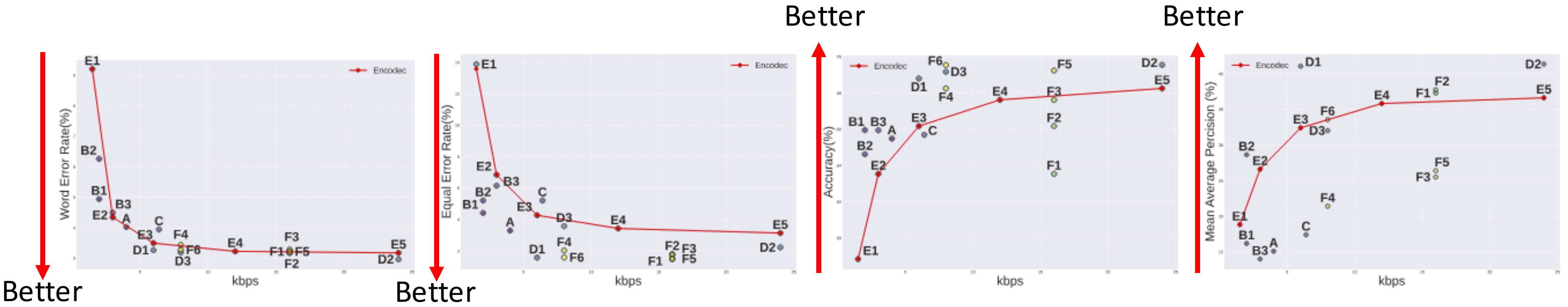
1. Input audio will undergo re-synthesis procedure
2. Both signal- and application-level evaluations are conducted

An overall score for clear comparison



1. Overall score: take the Harmonic mean for all normalized metrics
2. The overall score is with strong correlation to all signal-level metrics

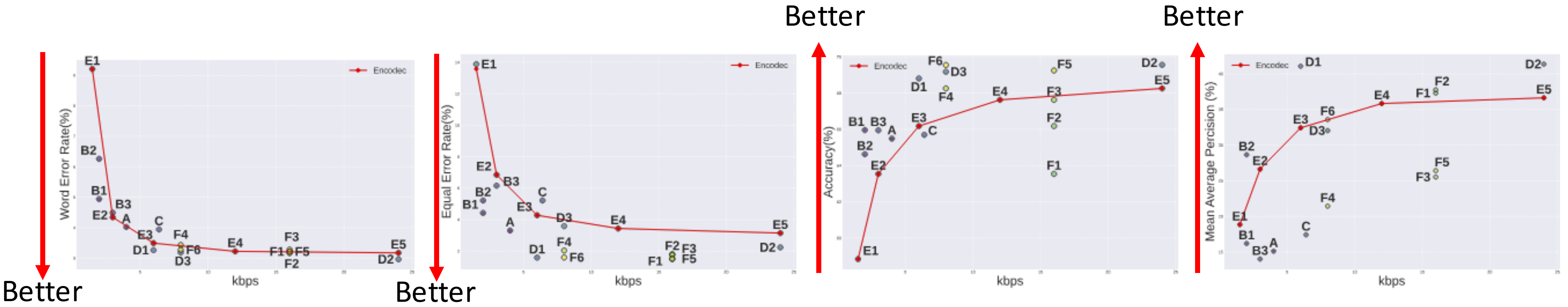
Results (in 2023) on application-level evaluation



(a) ASR WER (%) vs bitrate. (b) ASV EER (%) vs bitrate. (c) ER ACC (%) vs bitrate. (d) AEC mAP (%) vs bitrate.

- x-axis is the bitrate, and y-axis is the application performance
- Encodec (E) with different bitrates serves as the baseline
- Four applications are involved for content, speaker, emotion and audio information

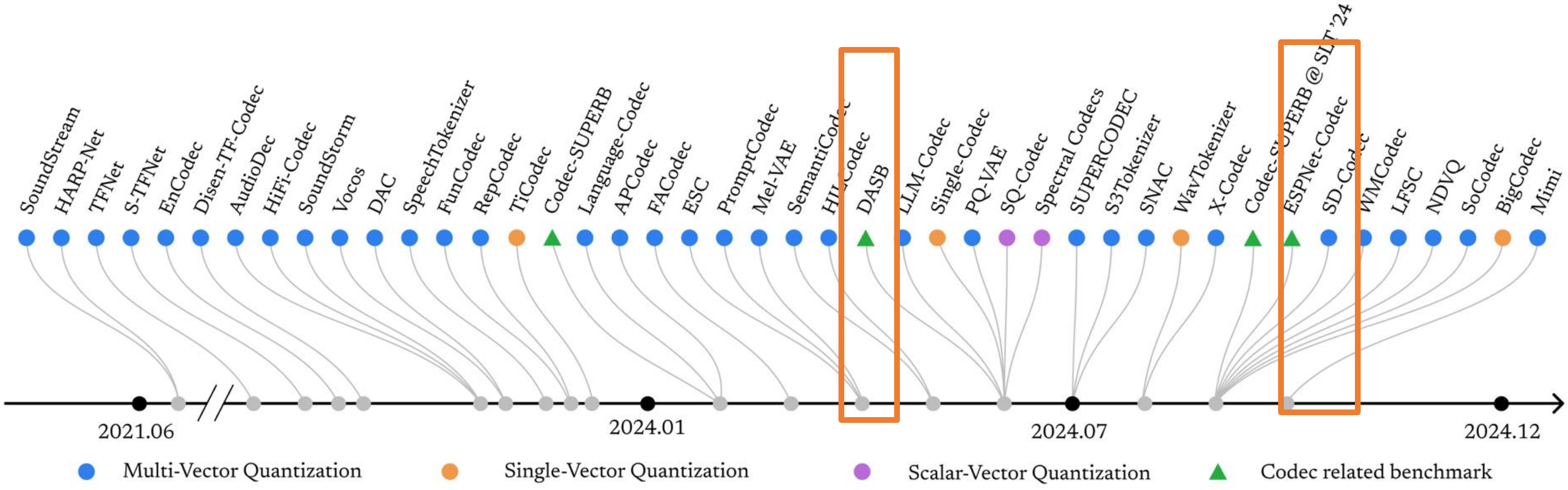
Results (in 2023) on application-level evaluation



(a) ASR WER (%) vs bitrate. (b) ASV EER (%) vs bitrate. (c) ER ACC (%) vs bitrate. (d) AEC mAP (%) vs bitrate.

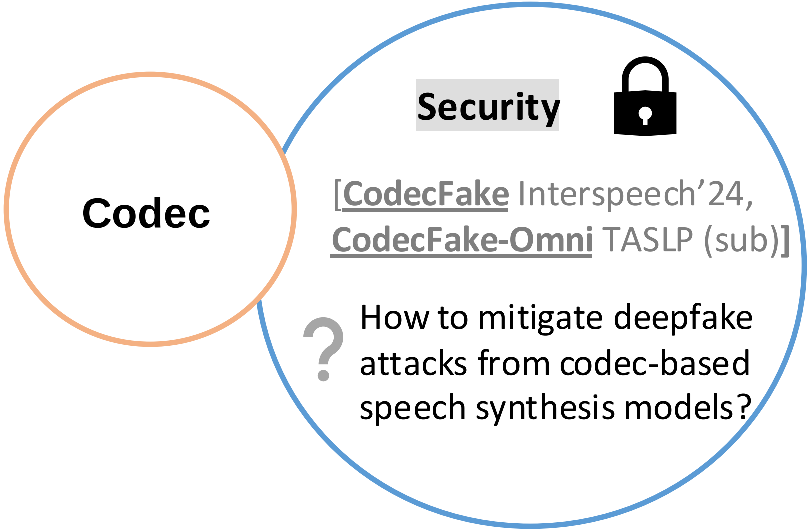
- Under low bitrate, B (AcamediCodec) is preferable
- Under mid bitrate, D (DAC) is preferable

Other codec benchmarks



- DASB extracts codec discrete codes for discrete representation learning
- ESPNet-Codec unifies the codec training setting for various codec models

Research roadmap and open questions



First- or corresponding-author paper Co-author paper

CodecFake: Enhancing Anti-Spoofing Models Against Deepfake Audios from Codec-Based Speech Synthesis Systems

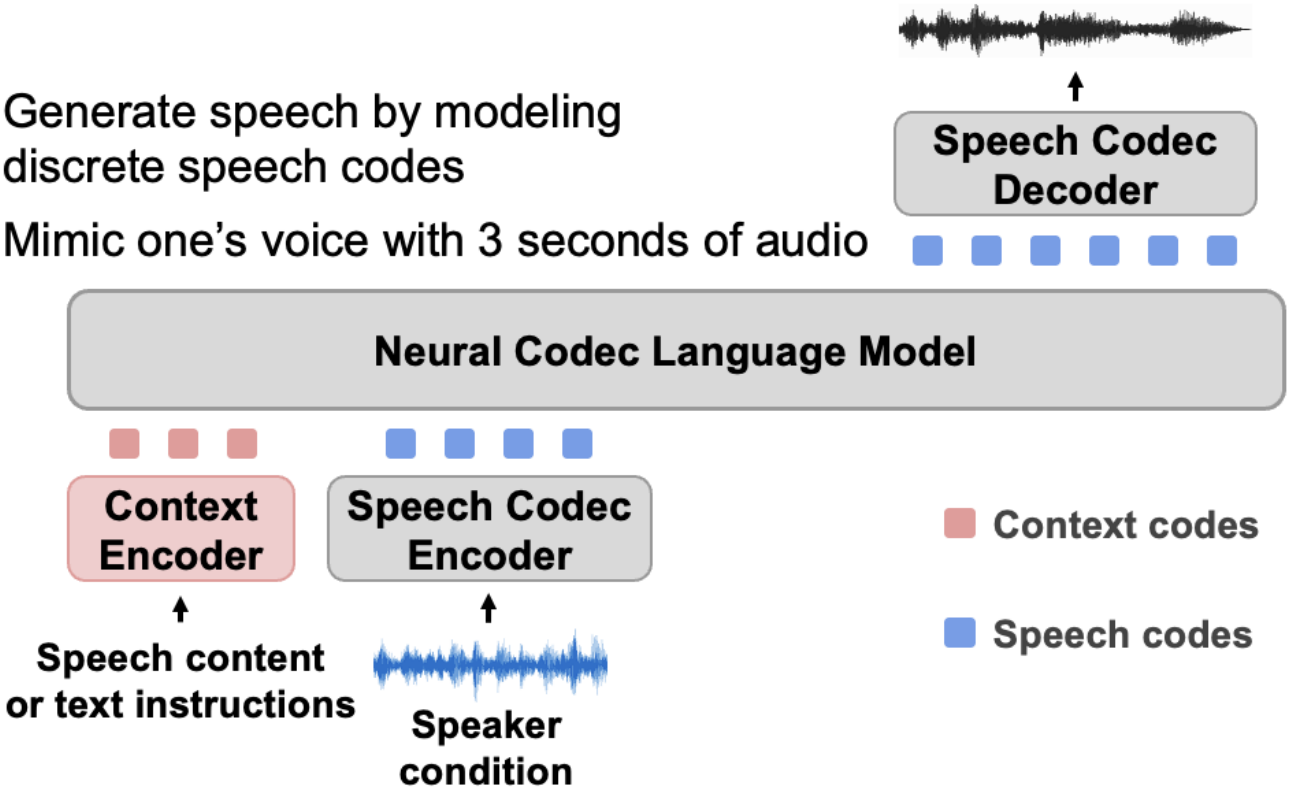
Haibin Wu[1,2], *Yuan Tseng*[1,2], *Hung-yi Lee*[1,2]

[1]Speech Processing and Machine Learning Laboratory, National Taiwan University
[2]Graduate Institute of Communication Engineering, National Taiwan University
f07921092@ntu.edu.tw

TL;DR: The first anti-spoofing dataset to counter codec-based speech synthesis deepfake attacks.

Codec-based speech generation systems

- Generate speech by modeling discrete speech codes
- Mimic one's voice with 3 seconds of audio



Can previous anti-spoofing models counter such attacks?

Use the state-of-the-art anti-spoofing model AASIST

Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks
J Jung, HS Heo, H Tak, H Shim… - ICASSP 2022-2022 …, 2022 - ieeexplore.ieee.org
… approach, named AASIST, outperforms the current state-of-the-art by 20% relative. Even a lightweight variant, AASIST-L, with only 85k parameters, outperforms all competing systems. …
☆ 儲存 99 引用 被引用 306 次 相關文章 全部共 12 個版本

Use the mainstream anti-spoofing dataset ASVspoof

ASVspoof 2019: Future horizons in spoofed and fake audio detection
M Todisco , X Wang , V Vestman , M Sahidullah … - arXiv preprint arXiv…, 2019 - arxiv.org
ASVspoof , now in its third edition, is a series of community-led challenges which promote the development of countermeasures to protect automatic speaker verification (ASV) from the …
☆ Stored 99 citations, cited 712 times, all related articles, 24 versions in total »

Equal error rate for codec-based TTS systems

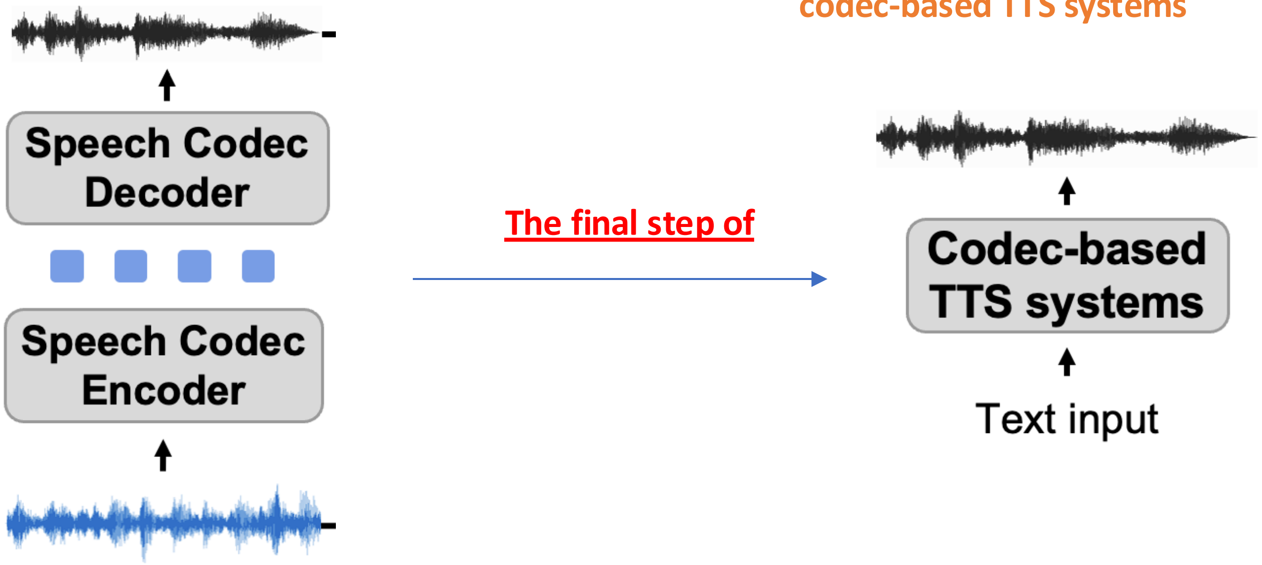
	ASVspoof (All)	
	AASIST	AASIST-L
VALL-E*	9.93	13.91
VALL-E X†	36.11	33.33
SpeechX†	33.33	25.00

Equal error rate for traditional TTS systems: less than 2%

Solution: Train on speech re-synthesized data by codecs

Codec resynthesized data

CodecFake: fake data generated by codec-based TTS systems



We assume codec resynthesized data shares similarities with CodecFake data.

→ Gather codec resynthesized data to train anti-spoofing models.

CodecFake: Dataset Creation

1. Divide VCTK corpus into 3 speaker-disjoint subsets

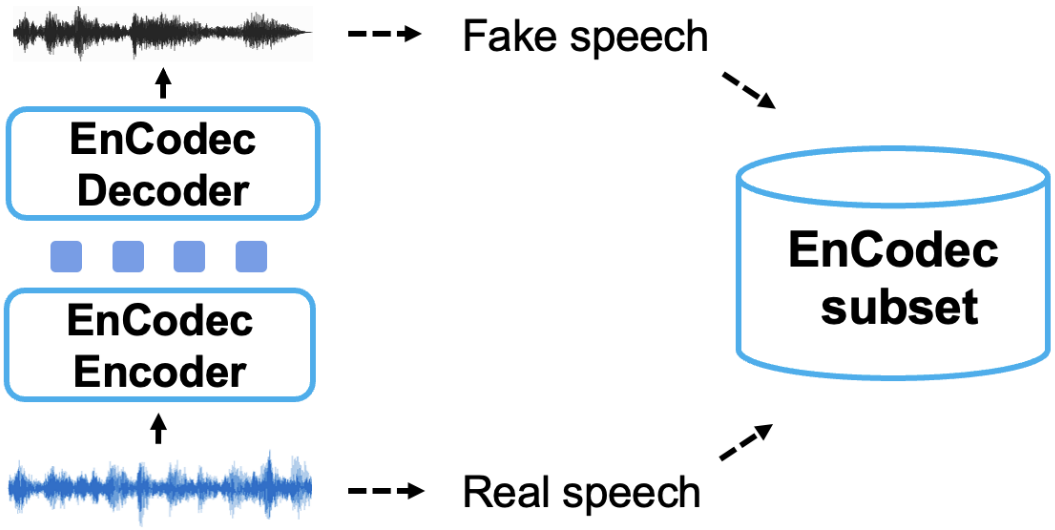
	# utterances/ # speakers		
	Training	Validation	Testing
Dataset Generation			
CodecFake	42752 / 103	735 / 2	755 / 2
VCTK	42752 / 103	735 / 2	755 / 2

2. Re-synthesize speech with 15 audio codecs from 6 frameworks

- **SpeechTokenizer**
- **AcademiaCodec**
- **AudioDec**
- **Descript Audio Codec**
- **EnCodec**
- **FunCodec**

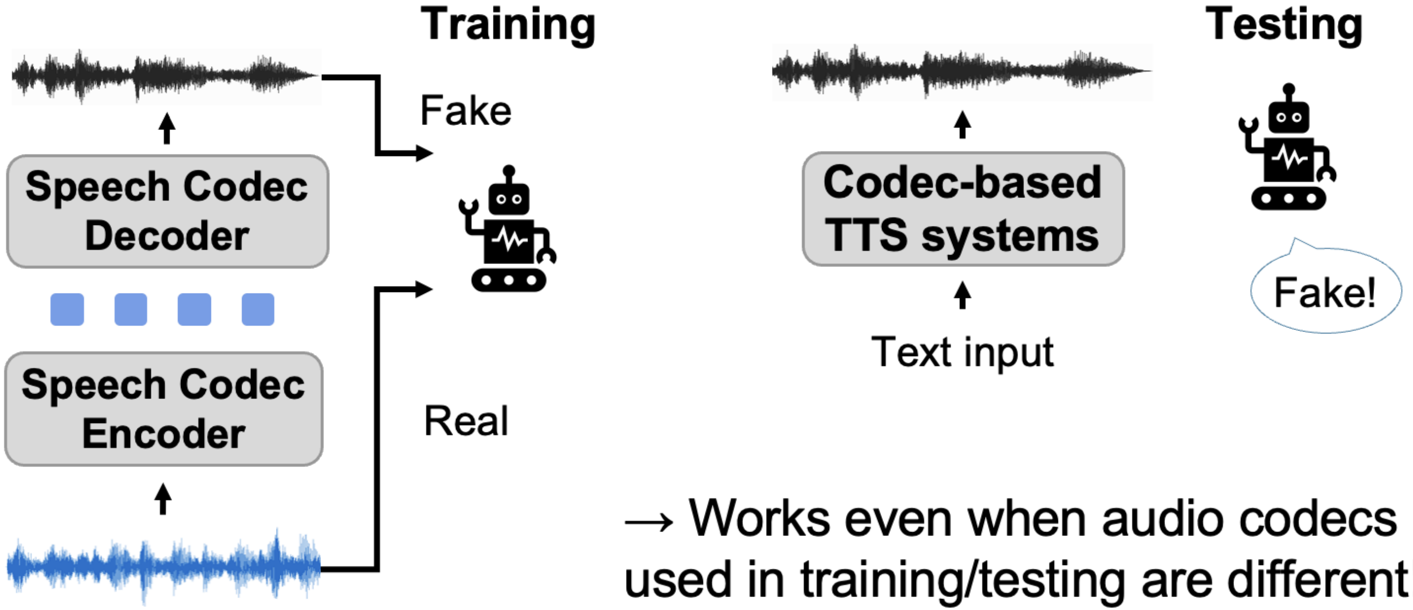
CodecFake: Dataset Creation (cont.)

3. Use each codec to encode VCTK utterances into discrete codes, then re-synthesize codes back into speech.



→ Train anti-spoofing models.

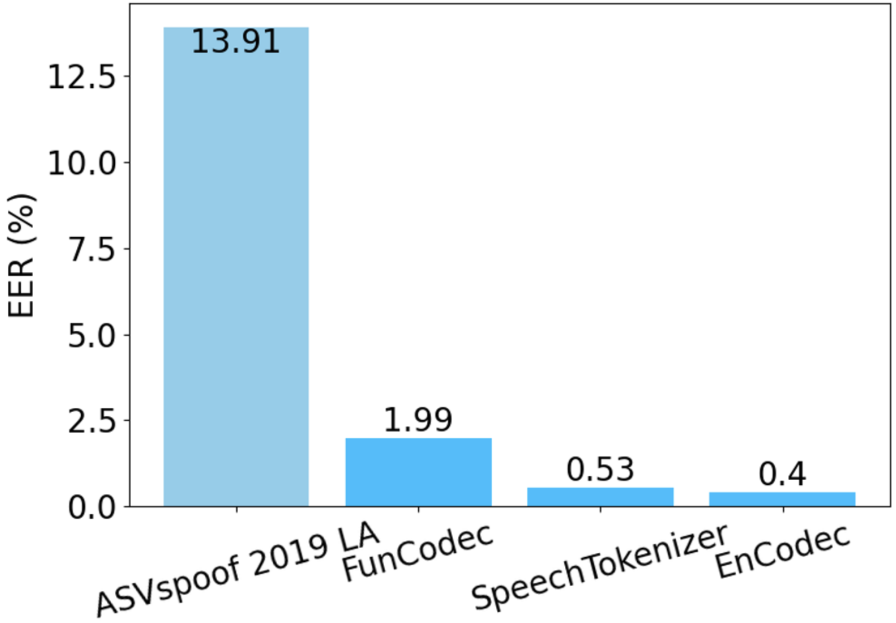
Solution: Train on speech re-synthesized by codecs



→ Works even when audio codecs used in training/testing are different!

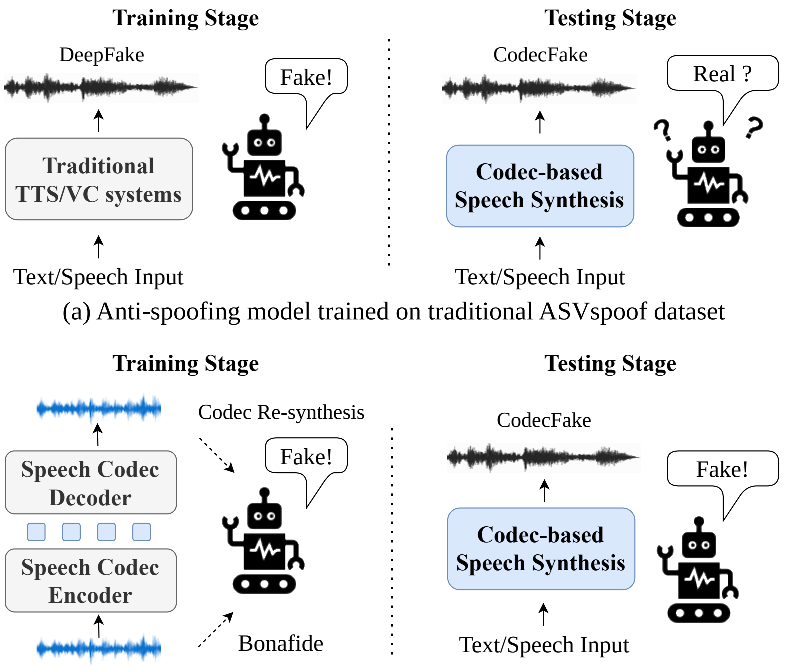
Detecting Speech from Codec-based TTS Systems

Comparison of AASIST models trained on different datasets



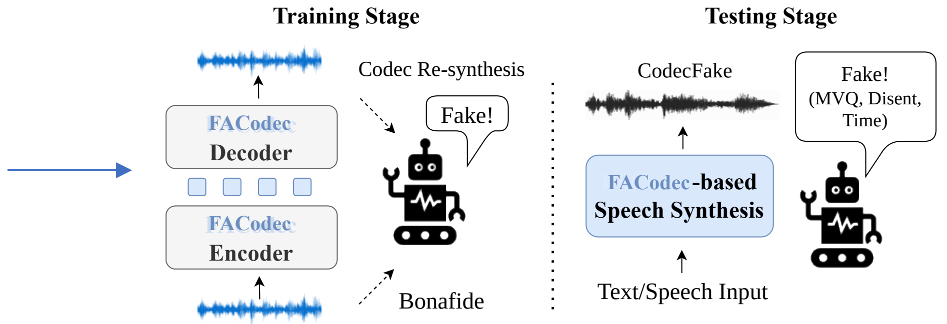
Y-axis: detection equal error rate for VALL-E generated data
X-axis: training subset

The next step - source tracing



(a) Anti-spoofing model trained on traditional ASVspoof dataset

(b) Anti-spoofing model trained on CodecFake dataset



(c) Anti-spoofing model trained on CodecFake-Omni dataset

Source tracing offers potential to improve generalization to spoofing attacks that are unseen during training but are composed of blocks encountered in training.

Summary with papers

Evaluation

The first neural audio codec benchmark

[1] **Wu, H.**, Chung, H. L., Lin, Y. C., Wu, Y. K., Chen, X., Pai, Y. C., ... & Lee, H. Y. (2024). Codec-superb: An in-depth analysis of sound codec models. ACL findings 2024

[2] Shi, J., Tian, J., Wu, Y., Jung, J. W., Yip, J. Q., Masuyama, Y., ... & Watanabe, S. (2024). Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. SLT 2024

[3] **Wu, H.**, Chen, X., Lin, Y. C., Chang, K., Du, J., Lu, K. H., ... & Lee, H. Y. (2024). Codec-SUPERB@ SLT 2024: A lightweight benchmark for neural audio codec models. SLT 2024

[4] Shi, J., et al. "VERSA: A Versatile Evaluation Toolkit for Speech, Audio, and Music." arXiv preprint arXiv:2412.17667 (2024).

Modeling

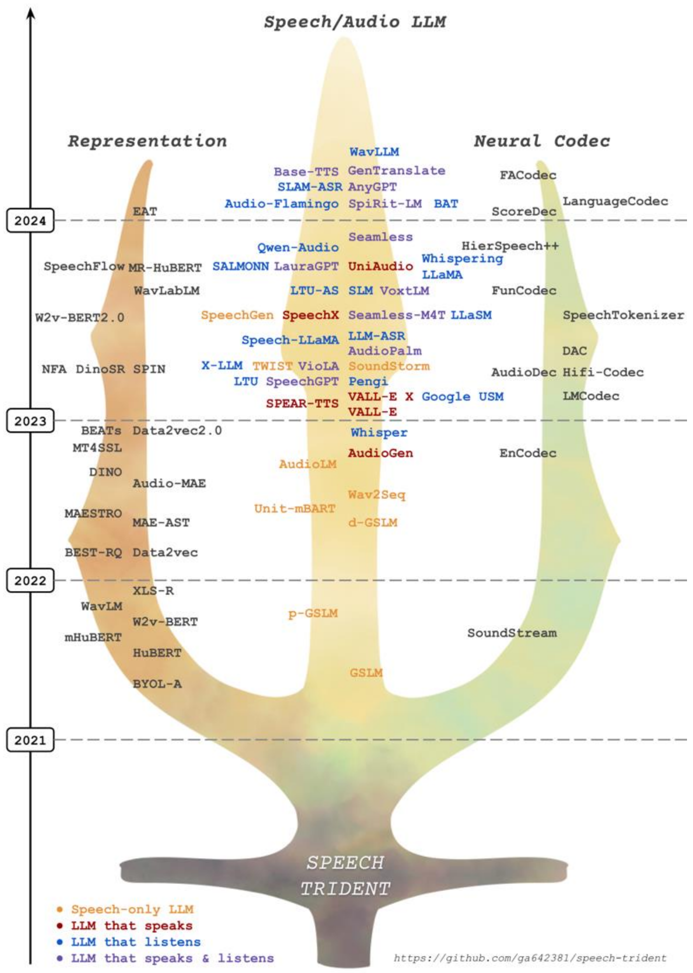
The first Transformer-only codec

[5] **Wu, H.**, Kanda, N., Eskimez, S. E., & Li, J. (2024). TS3-Codec: Transformer-Based Simple Streaming Single Codec. arXiv preprint arXiv:2411.18803.

Security

The first CodecFake dataset

[6] **Wu, H.**, Tseng, Y., & Lee, H. Y. (2024). CodecFake: Enhancing Anti-Spoofing Models Against Deepfake Audios from Codec-Based Speech Synthesis Systems. Interspeech 2024



Codec-SUPERB SLT' 24 special session

- The challenge covers nowday's neural audio codecs and speech / audio language models.
  - Time: December 3 15:00-18:30
  - Detailed agenda: https://codecsuperb.github.io/
- Keynote speakers
  - Neil Zeghidour (Moshi): 15:15-16:00
    - slides | recording
    - Title: Audio Language Models
  - Dongchao Yang (CUHK): 16:00-16:35
    - slides | recording
    - Title: Challenges in Developing Universal Audio Foundation Model
  - Shang-Wen Li (Meta): 16:35-17:10
    - slides | recording
    - Title: VoiceCraft: Zero-Shot Speech Editing and TTS in the Wild
  - Wenwu Wang (University of Surrey): 17:40-18:15
    - slides | recording
    - Title: Neural Audio Codecs: Recent Progress and a Case Study with SemantiCodec
  - Minje Kim (UIUC): 18:15-18:50
    - slides | recording
    - Title: Future Directions in Neural Speech Communication Codecs
- Host
  - Hung-yi Lee (NTU)
  - Haibin Wu (Microsoft)
- Accepted papers
  - ESPnet-Codec: Comprehensive Training and Evaluation of Neural Codecs for Audio, Music, and Speech
  - Codec-SUPERB @ SLT 2024: A lightweight benchmark for neural audio codec models
  - Investigating neural audio codecs for speech language model-based speech generation
  - Addressing Index Collapse of Large-Codebook Speech Tokenizer with Dual-Decoding Product-Quantized Variational Auto-Encoder
  - MDCTCodec: A Lightweight MDCT-based Neural Audio Codec towards High Sampling Rate and Low Bitrate Scenarios

https://github.com/ga642381/speech-trident