ESPnet-Codec and VERSA

Jiatong Shi

Carnegie Mellon University, Langauge Technologies Institute

jiatongs@cs.cmu.edu

10/27/2024

Content

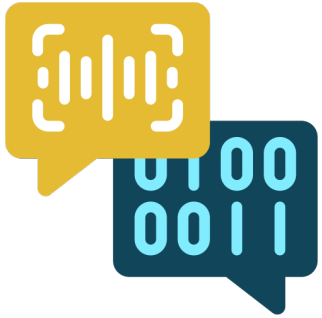
- ESPnet-Codec

  - Platform
  - Modeling
  - Experimental Findings

- VERSA toolkit

  - An integration of various evaluation metrics and more

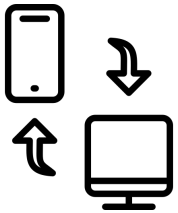
Motivation of ESPnet-Codec



- Expanded usage of neural codecs

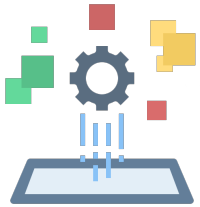
- In the past:

  - Transmission
    → the most widely used technique in speech/audio technologies



- Now:

  - Downstream modeling (speech/audio generation, etc.)



Potential of going discrete

- The power of scaling up

Table 1: A comparison between VALL-E and current cascaded TTS systems.

	Current Systems	**VALL-E**
Intermediate representation	mel spectrogram	audio codec code
Objective function	continuous signal regression	language model
Training data	$\leq$ 600 hours	60K hours
In-context learning	✗	✓

Vall-E (Wang et al. 2023)

Potential of going discrete

- The power of efficiency (storage)

Table 1: *Data size (bit) comparison of a T-second utterance among: 1) 16 kHz raw waveform in 16-bit wav format; 2) acoustic features in D-dimensional float vector with 100 frames/sec.; 3) SSL-based features using WavLM; 4) discrete tokens in 12-bit with 50 frames/sec.*

Data format	Data size (bits)
Raw waveform	$16 \times 16000 \times T$
Acoustic features	$32 \times D \times 100 \times T$
SSL-based features	$32 \times 1024 \times 50 \times T$
Discrete tokens	$12 \times 50 \times T$

Discrete ASR (Chang et al. 2023)

Potential of going discrete

- The power of efficiency (training)

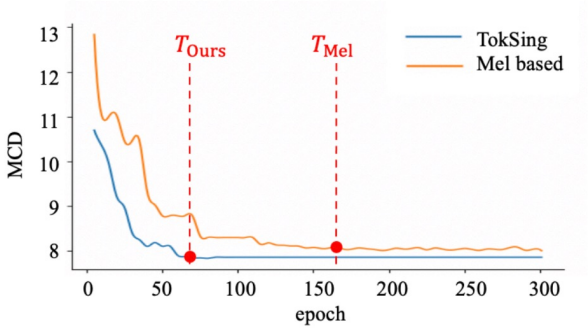
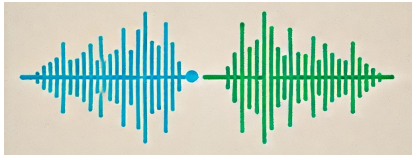


Figure 4: *Convergence speed comparison of TokSing and the Mel spectrogram-based system.*

TokSing (Wu et al. 2024)

Potential of going discrete



- The power of efficiency (Computation needs)

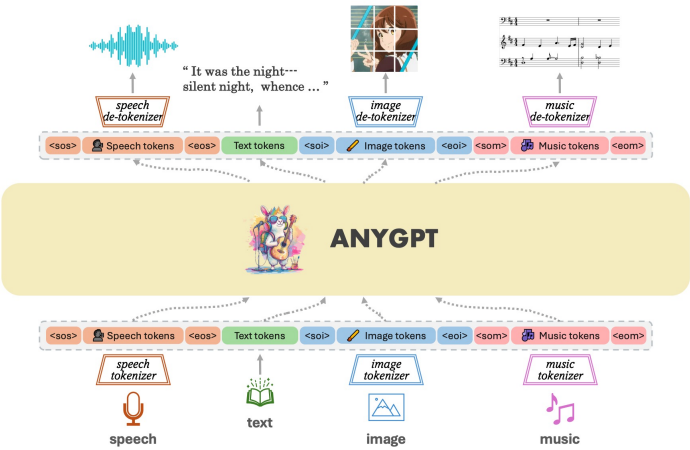
Table 2: *Computation and traning speech of Codecformer compared against Sepformer*

Model	GMACs	Training Time (h/epoch)
Sepformer [15]	77.3	2.7
Codecformer (Ours)	**1.5**	**1.0**

CodecFormer (Yip et al. 2024)

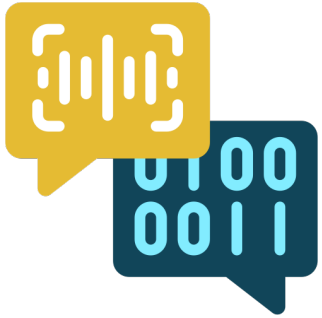
Potential of going discrete

- The power of connecting to different modalities



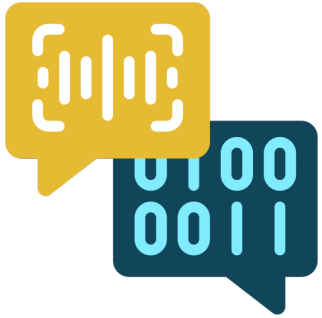
AnyGPT (Zhan et al. 2024)

Codecs



- Definition in the context
  - Discrete tokens
  - Optimized with speech/audio reconstruction

Recent advances in the community



- 2021-2022 [5]
  - Soundstream, TFNet, S-TFNet, Encodec, Disen-TF-Codec
- 2023 [10]
  - LMCodec, HiFi-Codec, AudioDec, DAC, VOCOS, Speechtokenizer, Funcodec, RepCodec, TiCodec, HierSpeech++
- 2024 (Jan. -> Jun.) [13]
  - ScoreDec, Language-Codec, AP-Codec, FACodec, ESC, PromptCodec, SemantiCodec, HILCodec, LLM-Codec, PQ-VAE, SQ-Codec, Single-Codec, CodecFake
- 2024 (Jul. -> Oct.) [14]
  - Super-Codec, WavTokenizer, X-Codec, BigCodec, SoCodec, NDVQ, Mimi, DM-Codec, TAAE, DC-Spin, VChangeCodec, LSCodec, SNAC, APCodec

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

However…

- The concern in fair comparison and comprehensive evaluation
  - Same dataset
  - Controlled experiment
  - Connection to downstream tasks

ESPnet-Codec

Platform

- General Neural Codec Framework
  - Sound Representation:
    - Time-domain / Spectral domain / Pre-trained features (SSL)
  - Encoder-Decoder Architecture
    - SEAnets and its variants (Conv. encoder/decoder)
    - SSL-based encoder (Transformer/Conformer/Branchformer encoder)
    - Vocoder-sourced decoder (A range of GAN-based vocoder)
  - Quantization Algorithm
    - Residual vector quantization (RVQ), group-RVQ (GRVQ)
    - [Supported after the paper] Multi-head VQ, Random projection quantizer, Finite state quantizer, look-up free quantizer, latent quantization (powered by https://github.com/lucidrains/vector-quantize-pytorch)
  - Training and Loss Functions
    - GAN-based training, time-domain/spectral-domain reconstruction loss, quantization loss

Example Neural Codec Models

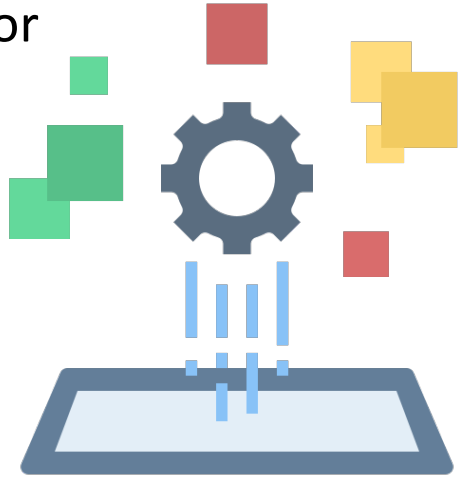
- SoundStream (Zeghidour et al. 2021)
  - The base RVQ-style neural tokens
- Encodec (Defossez et al. 2023)
  - Enhanced SoundStream with very initial efforts of open-source
- DAC (Kumar et al. 2024)
  - Enhanced Encodec (SOTA performance)
- FunCodec (Du et al. 2024)
  - Spectral-based codec
- HiFi-Codec (Yang et al. 2024)
  - GRVQ-style quantization method

Evaluation

- Reconstruction Evaluation
  - Reconstruction Quality Evaluation with VERSA (more later)
  - Reconstruction Quality Evaluation with Codec-SUPERB
- Downstream Evaluation
  - Various downstream evaluation supported

Downstream Applications

- Strong backbone downstream models supported for
  - Speech recognition
  - Text-to-speech
    - Non-autoregressive TTS
    - SpeechLM-style TTS (more in next section)
  - Speaker recognition
  - Speech separation and enhancement
  - Singing voice synthesis



ESPnet

Experiments

- Two settings
  - LibriTTS in 16kHz and 24kHz
  - AMUSE dataset (a combination of speech/audio/music) in 16kHz and 44.1kHz
    - Speech (Train: 24.1khrs)
    - Audio (Train 4.8khrs)
    - Music (Train: 1.8khrs)

Check our paper for detailed info
https://arxiv.org/pdf/2409.15897

A glance of results in the paper

Table 3: LibriTTS codec performance comparison at {16 / 24}kHz. * indicates models trained using existing open-source codebases.

Model	Intrusive										Non-Intrusive			Perceptual	
	Non-learning						Learning								
	MCD ↓	F0-RMSE ↓	F0-CORR ↑	STOI ↑	PESQ ↑	CI-SDR ↑	D-BLEU ↑	D-Distance ↑	S-BERT ↑	VISQOL ↑	UTMOS ↑	PLCMOS ↑	DNSMOS ↑	WER ↓	SPK-SIM ↑
DAC* [24]	13.85 / **3.17**	61.03 / 52.09	**0.61 / 0.72**	**0.98 / 0.99**	**3.81** / 3.45	15.48 / 14.02	**0.88 / 0.93**	**0.91 / 0.94**	**0.98 / 0.99**	2.90 / 4.60	3.81 / 3.95	4.01 / 4.19	3.05 / 3.02	2.9 / 2.7	0.85 / **0.97**
Encodec* [23]	4.84 / 3.94	**42.04 / 38.00**	0.59 / 0.65	0.97 / 0.98	3.45 / **4.09**	9.89 / 11.38	0.85 / 0.89	0.90 / 0.92	**0.98 / 0.99**	4.49 / **4.73**	3.86 / **4.08**	4.16 / **4.27**	3.06 / 3.09	2.4 / 2.4	0.86 / 0.95
FunCodec* [35]	5.40 / 5.52	43.19 / 43.81	0.55 / 0.55	0.90 / 0.93	3.23 / 3.11	11.46 / 36.29	0.78 / 0.79	0.87 / 0.87	0.95 / 0.96	4.57 / 4.52	3.84 / 3.84	**4.48** / 4.22	**3.15 / 3.14**	2.2 / **2.1**	**0.93** / 0.92
SpeechTokenizer* [18]	5.46 / -	43.32 / -	0.55 / -	0.93 / -	2.62 / -	**72.63** / -	0.80 / -	0.87 / -	0.95 / -	4.36 / -	3.80 / -	4.08 / -	3.06 / -	2.2 / -	0.62 / -
SoundStream	4.64 / 4.28	43.02 / 41.08	0.57 / 0.57	0.95 / 0.93	2.83 / 2.54	50.75 / 93.19	0.83 / 0.78	0.89 / 0.86	0.96 / 0.94	4.45 / 4.21	3.76 / 3.58	4.38 / 4.03	3.10 / 2.97	**2.1 / 2.1**	0.84 / 0.76
Encodec	**3.73** / 4.26	43.57 / 46.12	0.55 / 0.53	0.97 / 0.97	3.37 / 3.25	51.31 / 92.68	0.87 / 0.87	0.91 / 0.91	**0.98** / 0.97	**4.65** / 4.55	**3.92** / 3.93	4.45 / 4.17	3.11 / 3.12	**2.1 / 2.1**	0.90 / 0.86
DAC	5.16 / 4.98	65.20 / 62.92	0.56 / 0.53	0.95 / 0.97	3.00 / 3.43	48.84 / 93.51	0.85 / 0.87	0.89 / 0.90	0.97 / 0.97	4.48 / 4.21	3.78 / 3.93	4.28 / 4.07	3.04 / 3.05	2.9 / 2.8	0.76 / 0.83
HiFi-Codec	6.03 / 6.37	42.33 / 42.58	0.56 / 0.56	0.94 / 0.95	2.48 / 2.42	71.33 / **96.81**	0.78 / 0.77	0.86 / 0.85	0.94 / 0.93	4.01 / 3.96	3.61 / 3.58	4.23 / 3.88	2.89 / 2.89	3.1 / 3.7	0.66 / 0.65
Fun-Codec	5.87 / 5.76	67.72 / 69.90	0.54 / 0.51	0.83 / 0.86	2.43 / 2.40	50.23 / 89.73	0.71 / 0.73	0.82 / 0.83	0.90 / 0.91	4.26 / 4.24	3.01 / 3.09	4.13 / 3.90	3.04 / 3.03	3.7 / 3.6	0.72 / 0.70
Ground Truth	-	-	-	-	-	-	-	-	-	-	4.05	4.42	3.08	2.1	-

Table 4: AMUSE codec performance comparison on the ***speech*** test set at {16 / 44.1}kHz.

Model	Intrusive										Non-Intrusive			Perceptual
	Non-learning						Learning							
	MCD ↓	F0-RMSE ↓	F0-CORR ↑	STOI ↑	PESQ ↑	CI-SDR ↑	D-BLEU ↑	D-Distance ↑	S-BERT ↑	VISQOL ↑	UTMOS ↑	PLCMOS ↑	DNSMOS ↑	SPK-SIM ↑
SoundStream	**4.77** / 5.87	**42.95** / 46.67	**0.51** / 0.49	**0.92** / 0.94	2.71 / 2.79	63.12 / **60.64**	**0.71** / 0.73	**0.88** / 0.88	**0.95** / 0.96	4.51 / 4.49	1.95 / 1.93	**3.78 / 3.65**	2.66 / 2.59	**0.87** / 0.82
Encodec	4.91 / 5.14	46.73 / 57.64	0.46 / 0.37	0.91 / 0.91	2.28 / 1.81	60.57 / 58.57	0.70 / 0.73	0.87 / 0.88	0.94 / 0.95	**4.52** / 4.45	1.55 / 1.59	3.64 / 2.17	2.54 / 2.44	0.85 / 0.85
DAC	5.76 / **4.99**	44.36 / **43.19**	**0.51 / 0.55**	0.91 / **0.97**	**2.75 / 4.02**	**66.08** / 60.20	0.68 / **0.79**	0.86 / **0.90**	0.94 / **0.98**	4.29 / **4.67**	**2.14 / 2.37**	3.65 / 2.34	**2.77 / 2.84**	0.78 / **0.93**
Ground Truth	-	-	-	-	-	-	-	-	-	-	2.45	3.85	2.79	-

Table 5: AMUSE codec performance comparison on the ***audio*** test set at {16 / 44.1}kHz.

Model	MCD ↓	CI-SDR ↑	VISQOL ↑
SoundStream	**3.57** / 5.05	2.47 / **82.01**	**4.38** / 3.92
Encodec	3.59 / **4.90**	-4.64 / 73.14	4.32 / **4.03**
DAC	3.59 / 4.97	**7.72** / 75.84	4.23 / 3.90

Table 6: AMUSE codec performance comparison on the ***music*** test set at {16 / 44.1}kHz.

Model	MCD ↓	CI-SDR ↑	VISQOL ↑
SoundStream	3.86 / 5.60	5.41 / **70.51**	3.77 / **3.93**
Encodec	4.20 / 5.60	0.00 / 65.79	**3.81** / 3.81
DAC	**3.66 / 5.14**	**14.61** / 65.49	3.52 / 3.67

Table 7: CodecSUPERB benchmark performance (hidden set).+ indicate models trained on large-scale AMUSE dataset.

Model	Application				Signal Metrics			
	ASR	SPK	ER	AEC	Speech			Audio
	WER ↓	EER	ACC ↑	ACC ↑	PESQ ↑	STOI ↑	Mel Loss ↓	Mel Loss ↓
SoundStream	5.99	2.80	49.49	51.13	2.54	0.91	0.87	2.21
Encodec	**5.58**	**2.20**	52.53	61.58	**2.99**	**0.95**	**0.74**	1.02
DAC	6.33	3.40	**55.56**	52.65	2.51	0.91	0.90	1.25
SoundStream+	5.88	2.40	53.54	**68.11**	2.95	0.93	**0.74**	**0.96**
Encodec+	5.98	2.40	50.51	66.50	2.48	0.92	0.80	0.97
DAC+	6.08	3.20	54.55	61.89	2.94	0.93	0.86	1.08
Original Audio	5.28	1.60	59.60	78.01	-	-	-	-

scores but still has comparable VISQOL scores to other codec mod-

Table 8: Comparison of codecs on ASR, TTS, SPK, SSE, SVS, and SSL (evaluated on ASR and SLU). + stands for codecs trained on AMUSE. Test-{clean / other} sets are reported for ASR. Models that failed to train are marked with ✗.

Model	ASR	NAR-TTS			AR-TTS			SPK	SSE			SVS			SSL (ASR)	SSL (SLU)
	WER ↓	WER ↓	UTMOS ↑	SPK-SIM ↑	WER ↓	UTMOS ↑	SPK-SIM ↑	EER ↓	PESQ ↑	STOI ↑	DNSMOS ↑	MCD ↓	SACC ↑	SingMOS ↑	WER ↓	ACC ↑
SoundStream	3.7 / 11.1	**3.4**	2.34	0.58	**6.7**	3.70	0.63	27.5	1.79	0.73	0.93	-	-	-	✗	✗
Encodec	**3.6 / 10.0**	4.3	**2.35**	**0.59**	7.7	**3.85**	0.63	15.7	1.85	0.76	**0.95**	-	-	-	✗	1.5
DAC	**3.6** / 10.3	5.2	2.12	0.54	10.2	3.75	**0.66**	29.5	1.73	0.74	0.87	-	-	-	✗	✗
SoundStream+	**3.7 / 10.2**	4.7	1.84	0.57	9.8	2.97	0.61	15.9	1.76	0.73	0.81	8.82	**0.60**	**2.92**	✗	✗
Encodec+	3.9 / 10.3	5.4	1.92	0.58	8.6	2.32	0.62	**14.1**	1.24	0.51	0.62	**8.51**	0.58	2.86	28.1	**68.9**
DAC+	4.1 / 10.9	6.2	1.82	0.56	15.9	2.94	0.64	24.6	**2.00**	**0.78**	0.87	9.26	0.49	2.58	✗	3.6

Selected Findings (VERSA)

Check our paper for comprehensive results
https://arxiv.org/pdf/2409.15897

- There is no agreements on winners over each evaluation metric in different settings

- In the LibriTTS 16/24kHz setting:

  - SoundStream (Best in 4/13 metrics (F0-related metrics))
  - Encodec (Best in 11/13 metrics, tied WER with Soundstream)
  - DAC (Better than Soundstream in most metrics, but worse than Encodec)

- 15 metrics in total

Selected Findings (VERSA)

Check our paper for comprehensive results
https://arxiv.org/pdf/2409.15897

- There is no agreements on winners over each evaluation metric in different settings

- In the AMUSE 16/44.1kHz setting (speech-test):
  - SoundStream (Best in 9/2 metrics)
  - Encodec (Best in 0/0 metrics)
  - DAC (Best in 5/12 metrics)

- 14 metrics in total

Selected Findings (VERSA)

Check our paper for comprehensive results
https://arxiv.org/pdf/2409.15897

- There is no agreements on winners over each evaluation metric in different settings
- Different behaviors in audio/music tests

Table 5: AMUSE codec performance comparison on the ***audio*** test set at {16 / 44.1}kHz.

Model	MCD ↓	CI-SDR ↑	VISQOL ↑
SoundStream	**3.57** / 5.05	2.47 / **82.01**	**4.38** / 3.92
Encodec	3.59 / **4.90**	-4.64 / 73.14	4.32 / **4.03**
DAC	3.59 / 4.97	**7.72** / 75.84	4.23 / 3.90

Table 6: AMUSE codec performance comparison on the ***music*** test set at {16 / 44.1}kHz.

Model	MCD ↓	CI-SDR ↑	VISQOL ↑
SoundStream	3.86 / 5.60	5.41 / **70.51**	3.77 / **3.93**
Encodec	4.20 / 5.60	0.00 / 65.79	**3.81** / 3.81
DAC	**3.66 / 5.14**	**14.61** / 65.49	3.52 / 3.67

Selected Findings (Codec-SUPERB)

- Codec-SUPERB evaluation (Wu et al. 2024)
  - Experiments on reconstructed audio signals on a benchmark hidden set, including
    - Signal metrics
      - PESQ/STOI/Mel Distortion
    - Application metrics
      - Speech Recognition/Speaker recognition/Emotion Recognition/Audio Event Classification

Selected Findings (Codec-SUPERB)

- Codec-SUPERB evaluation (Wu et al. 2024)

Model	Application				Signal Metrics			
	ASR	SPK	ER	AEC	Speech			Audio
	WER ↓	EER	ACC ↑	ACC ↑	PESQ ↑	STOI ↑	Mel Loss ↓	Mel Loss ↓
SoundStream	5.99	2.80	49.49	51.13	2.54	0.91	0.87	2.21
Encodec	**5.58**	**2.20**	52.53	61.58	**2.99**	**0.95**	**0.74**	1.02
DAC	6.33	3.40	**55.56**	52.65	2.51	0.91	0.90	1.25
SoundStream+	5.88	2.40	53.54	**68.11**	2.95	0.93	**0.74**	**0.96**
Encodec+	5.98	2.40	50.51	66.50	2.48	0.92	0.80	0.97
DAC+	6.08	3.20	54.55	61.89	2.94	0.93	0.86	1.08
Original Audio	5.28	1.60	59.60	78.01	-	-	-	-

Selected Findings (Codec-SUPERB)

- Codec-SUPERB evaluation (Wu et al. 2024)

Model	Application				Signal Metrics			
	ASR	SPK	ER	AEC	Speech			Audio
	WER ↓	EER	ACC ↑	ACC ↑	PESQ ↑	STOI ↑	Mel Loss ↓	Mel Loss ↓
SoundStream	5.99	2.80	49.49	51.13	2.54	0.91	0.87	2.21
Encodec	**5.58**	**2.20**	52.53	61.58	**2.99**	**0.95**	**0.74**	1.02
DAC	6.33	3.40	**55.56**	52.65	2.51	0.91	0.90	1.25
SoundStream+	5.88	2.40	53.54	**68.11**	2.95	0.93	**0.74**	**0.96**
Encodec+	5.98	2.40	50.51	66.50	2.48	0.92	0.80	0.97
DAC+	6.08	3.20	54.55	61.89	2.94	0.93	0.86	1.08
Original Audio	5.28	1.60	59.60	78.01	-	-	-	-

Selected Findings (Codec-SUPERB)

- Codec-SUPERB evaluation (Wu et al. 2024)

Model	Application				Signal Metrics			
	ASR	SPK	ER	AEC	Speech			Audio
	WER ↓	EER	ACC ↑	ACC ↑	PESQ ↑	STOI ↑	Mel Loss ↓	Mel Loss ↓
SoundStream	5.99	2.80	49.49	51.13	2.54	0.91	0.87	2.21
Encodec	**5.58**	**2.20**	52.53	61.58	**2.99**	**0.95**	**0.74**	1.02
DAC	6.33	3.40	**55.56**	52.65	2.51	0.91	0.90	1.25
SoundStream+	5.88	2.40	53.54	**68.11**	2.95	0.93	**0.74**	**0.96**
Encodec+	5.98	2.40	50.51	66.50	2.48	0.92	0.80	0.97
DAC+	6.08	3.20	54.55	61.89	2.94	0.93	0.86	1.08
Original Audio	5.28	1.60	59.60	78.01	-	-	-	-

Selected Findings (Downstream Application)

- Codecs trained on LibriTTS vs. AMUSE

Model	ASR	NAR-TTS			AR-TTS			SPK	SSE			SVS			SSL (ASR)	SSL (SLU)
	WER ↓	WER ↓	UTMOS ↑	SPK-SIM ↑	WER ↓	UTMOS ↑	SPK-SIM ↑	EER ↓	PESQ ↑	STOI ↑	DNSMOS ↑	MCD ↓	SACC ↑	SingMOS ↑	WER ↓	ACC ↑
SoundStream	3.7 / 11.1	**3.4**	2.34	0.58	**6.7**	3.70	0.63	27.5	1.79	0.73	0.93	-	-	-	✗	✗
Encodec	**3.6 / 10.0**	4.3	**2.35**	**0.59**	7.7	**3.85**	0.63	15.7	1.85	0.76	**0.95**	-	-	-	✗	1.5
DAC	**3.6** / 10.3	5.2	2.12	0.54	10.2	3.75	**0.66**	29.5	1.73	0.74	0.87	-	-	-	✗	✗
SoundStream+	3.7 / 10.2	4.7	1.84	0.57	9.8	2.97	0.61	15.9	1.76	0.73	0.81	8.82	**0.60**	**2.92**	✗	✗
Encodec+	3.9 / 10.3	5.4	1.92	0.58	8.6	2.32	0.62	**14.1**	1.24	0.51	0.62	**8.51**	0.58	2.86	28.1	**68.9**
DAC+	4.1 / 10.9	6.2	1.82	0.56	15.9	2.94	0.64	24.6	**2.00**	**0.78**	0.87	9.26	0.49	2.58	✗	3.6

Summary

- The comparison of codecs are complex due to
  - Diverse downstream applications (understanding vs. generation)
  - Various domains in data (speech, music, general audio)
  - Tremendous hyper-parameters in training (many of them have significant impact to the model performance)

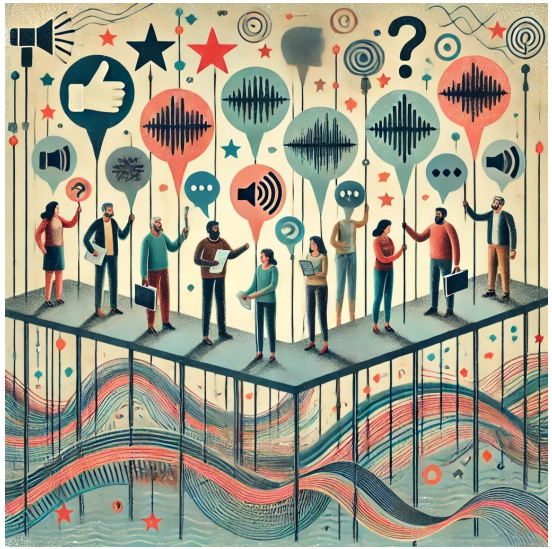
- ESPnet-Codec has proposed an attempt to provide
  - Reproducible modeling
  - Easy connection to strong downstream modeling
  - → we are adding more models with comprehensive tuning for the best performance. Please stay tuned!

Motivation of a new evaluation toolkit

- Plenty of difficulties exist in evaluation for generated speech/audio quality

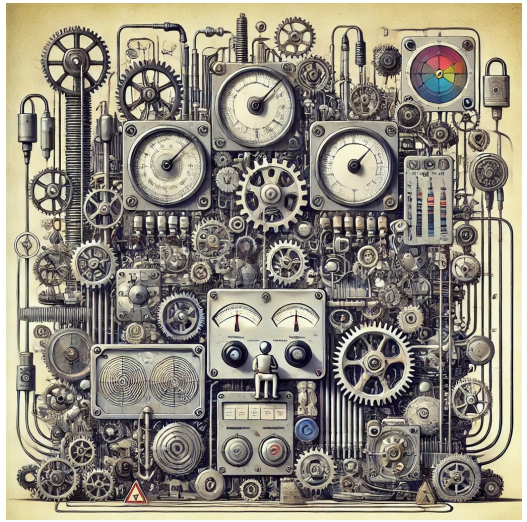
Motivation of a new evaluation toolkit

- Plenty of difficulties exist in evaluation for generated speech/audio quality
  - Collecting subjective evaluations is not easy



Motivation of a new evaluation toolkit

- Plenty of difficulties exist in evaluation for generated speech/audio quality
  - Collecting subjective evaluations is not easy
  - Implementing objective evaluations is not easy



Motivation of a new evaluation toolkit

- Plenty of difficulties exist in evaluation for generated speech/audio quality:
  - Collecting subjective evaluations is not easy
  - Implementing objective evaluations is not easy
  - Increasing need to support multi-domain larger-scale evaluation with unified speech modeling

Introducing VERSA

- VERSA (Versatile Evaluation for Speech and Audio)
  - Targets a general interface for speech and audio evaluation
  - A collection of conventional/recent automatic quality evaluation metrics
  - Highly integrated to speech processing toolkit ESPnet

Metrics in VERSA

- Up to **30+ metrics** in speech and audio evaluation supported currently

- Working in progress to add more recent techniques/evaluation algorithms
  - Let us know if you want anything more supported!

Metrics in VERSA

Dependency	Base	Example Evaluation Metrics
Intrusive	Non-Learning	MCD, F0-RMSE, F0-CORR, SI-SNR
		CI-SDR, PESQ, STOI, WARPQ
	Learning	VISQOL, SpeechDiscreteBLEU, SpeechDiscreteDistance, SpeechBERT, FAD(10+ embeddings), KLD(10+ embeddings)
Non-intrusive		DNSMOS, UTMOS, PLCMOS, SingMOS, Torch-squim (PESQ, STOI, SI-SNR), Sheet
Other Perceptual		CER/WER (ESPnet/ESPnet-OWSM, Whisper), SPK-SIM (10+ pre-trained speaker models), CLAP-score, Log-WMSE

Check full list and details about metrics at https://github.com/shinjiwlab/versa

Comparison to existing open-source toolkits

Toolkit	Github Repo	Evaluation-Focus	Supported Metrics
ESPnet	espnet/espnet	❌	16
SpeechBrain	speechbrain/speechbrain	❌	0
Amphion	open-mmlab/Amphion	❌	15
Stable-Audio-Metrics	Stability-AI/stable-audio-metrics	✅	3
AudioLDM-Eval	haoheliu/audioldm_eval	✅	9
VERSA	shinjiwlab/versa	✅	**30 (100+ variants)**

Compared to other open-source toolkits, VERSA is dedicated to speech and audio evaluation with **extensive integration** of both conventional and recent evaluation metrics.

VERSA Next Move

Github Codebase

Colab Demo

- Towards audio-profiling

- We are targeting for a speech/music/audio profiling toolkit that
  - integrates extensive pre-trained models to achieve comprehensive profiling
  - includes descriptive metrics such as:
    - Silence information and speaker counting (with pre-trained VAD model and speaker diarization models)
    - Estimated SNR (with pre-trained speech enhancement model)
    - Audio event type (with pre-trained audio classification model)
    - Music information such as genre, mood, intrumental, chord, singing technique, pitch range, transcription, beat tracking (with pre-trained music information retrieval models

Acknowledgements

- Thanks to all the collaborators to these projects, including

  Jinchuan Tian (CMU), Yihan Wu (RUC), Jee-weon Jung (CMU->Apple), Jia Qi Yip (NTU - Singapore), Yoshiki Masuyama (TMU), William Chen (CMU), Yuning Wu (RUC), Yuxun Tang (RUC), Massa Baali (CMU), Dareen Alharhi (CMU), Dong Zhang (FDU), Ruifan Deng (FDU), Tejes Srivastava (Uchicago), Haibin Wu (NTU – Taiwan),

  Hye-Jin Shim (CMU), Alexander H. Liu (MIT), Bhiksha Raj (CMU), Qin Jin (RUC), Ruihua Song (RUC), Shinji Watanabe (CMU)

Icons and images are either from flaticon.com or generated from Dall-E.

Thank you for listening!