ESPnet-Codec

Jiatong Shi

Carnegie Mellon University, Langauge Technologies Institute

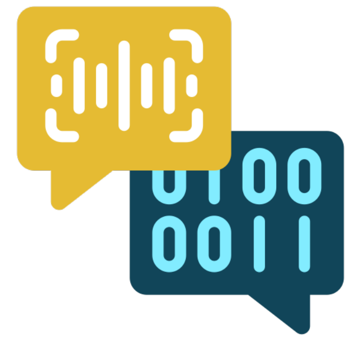
jiatongs@cs.cmu.edu

10/27/2024

Content

- Motivation
- Platform Details
  - Methods
  - Evaluation
- Selected Experimental Findings

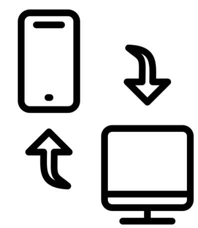
Motivation of ESPnet-Codec



- Expanded usage of neural codecs

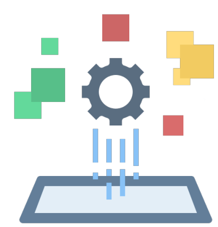
- In the past:

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



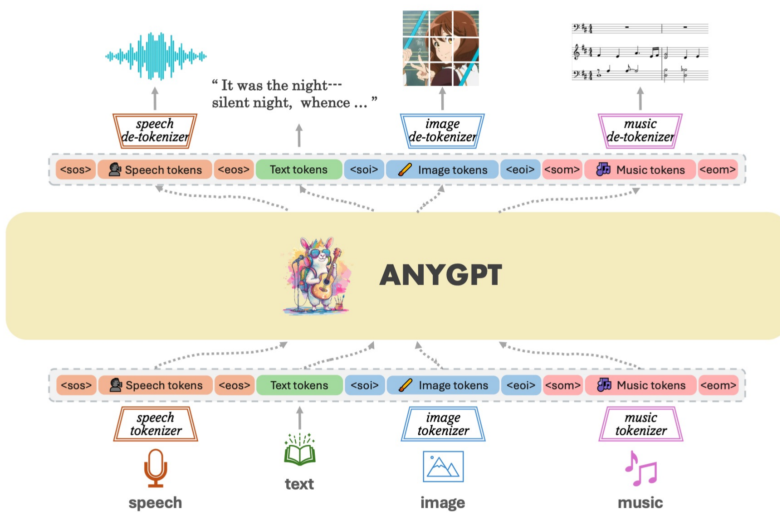
- Now:

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



Potential of going discrete

- The power of connecting to different modalities



AnyGPT (Zhan et al. 2024)

General Formulation of Speech Coding

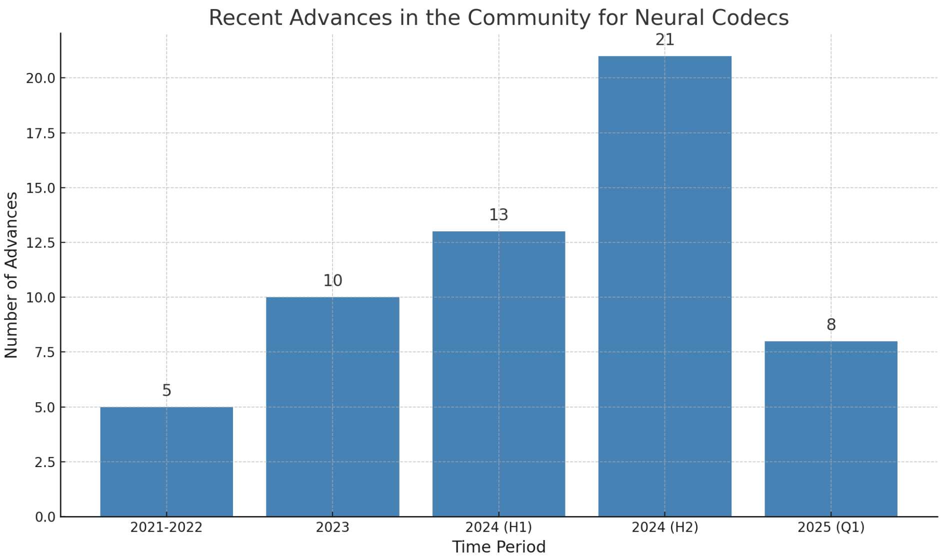
$h = E(x)$ $\quad$ $c = Q(h)$ $\quad$ $\hat{x} = D(c)$

x: input speech
h: hidden representation
c: codec tokens
$\hat{x}$: reconstructed speech

E: Encoder
Q: Quantizer
D: Decoder

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. -> Dec.) [21]
  - Super-Codec, WavTokenizer, X-Codec, BigCodec, SoCodec, NDVQ, Mimi, DM-Codec, TAAE, DC-Spin, VChangeCodec, LSCodec, SNAC, APCodec, MDCTCodec, SimVQ, UniCodec, PyamidCodec, TAAE, FreeCodec, TS3Codec
- 2025 (Jan. -> Mar.) [8]
  - ComplexDec, X-Codec2, FocalCodec, Baichuan-Audio Tokenizer, UniCodec, FlowDec, BiCodec



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

However…

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

ESPnet-Codec

Platform Supports

Encoder

- Sound Representation **(3 choices)**
  - Time domain
  - Spectral domain
  - Pre-trained features (e.g., from self-supervised learning representation)
- Encoder Architecture **(3 choices)**
  - Conv. encoder
  - Conv. + RNN encoder (e.g., SEAnets)
  - Conv. + Transformer encoder

Quantizer

- Quantization Algorithm **(3 choices)**
  - Residual vector quantization (RVQ)
  - Group RVQ (GRVQ)
  - Finite scalar quantization (FSQ)

Decoder

- Decoder Architecture **(3 choices)**
  - Conv. decoder
  - Conv. + RNN decoder
  - Conv. + Transformer decoder

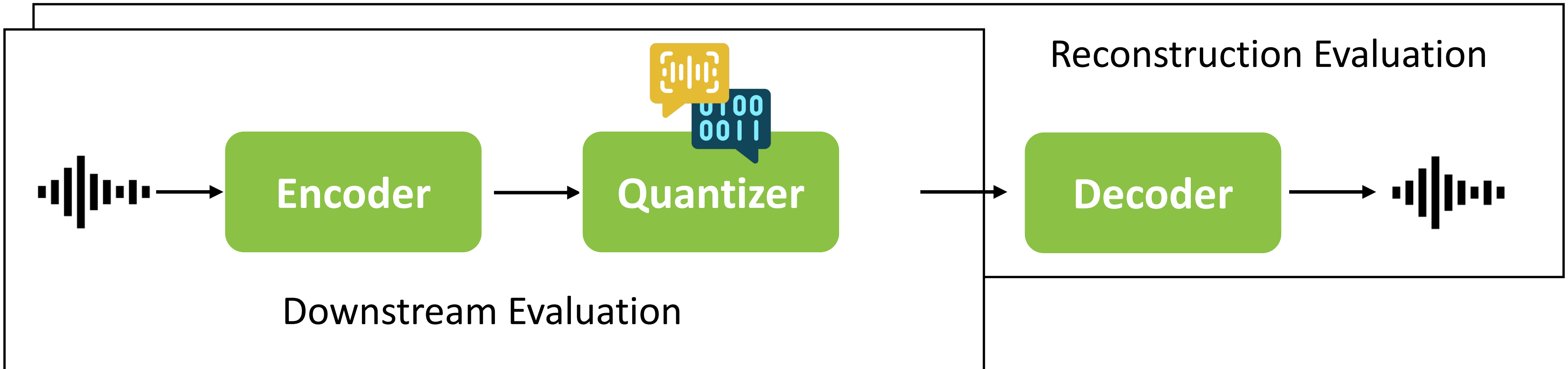
Discriminators

- Discriminator Architecture **(7 choices)**
  - Multi-resolution STFT (MS-STFT)
  - Filter-bank random window (FB-RWD)
  - Multi-scale (MSD)
  - Multi-period (MPD)
  - Simple STFT
  - Multi-scale complex STFT
  - Multi-scale multi-period multi-band (MSMPMB)

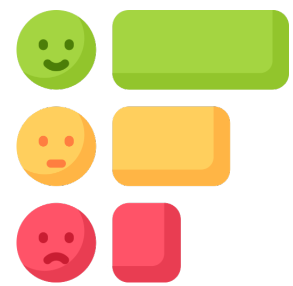
Example Neural Codec Models

Model	Sound Representation	Encoder & Decoder	Quantizer	Discriminator
SoundStream (2021)	Time domain	Conv.	RVQ	STFT Discriminator
Encodec (2023)	Time domain	Conv. + RNN	RVQ	MS-STFT Discriminator
DAC (2024)	Time domain	Conv.	RVQ (factorized + L2-normalized)	MSMPMBD
FunCodec (2024)	Spectral domain	Conv.	RVQ	MSD, MPD, MS-STFTD
HiFi-Codec (2024)	Time domain	Conv. + RNN	GRVQ	MS-STFTD

Evaluation



- Reconstruction Evaluation
  - Reconstruction Quality Evaluation with Quality Metrics (VERSA)
  - Reconstruction Quality Evaluation with Codec-SUPERB



- Downstream Evaluation
  - Evaluation conducted on various downstream speech processing tasks

Reconstruction Quality Evaluation with 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
    - Up to **65 metrics with 729 variants supported**
  - Highly integrated to speech processing toolkit ESPnet

Metrics in VERSA

Dependency	Example Evaluation Metrics
Dependent	MCD, F0-RMSE, F0-CORR, SI-SNR, CI-SDR, PESQ, STOI, WARPQ, VISQOL, SpeechDiscreteBLEU, SpeechDiscreteDistance, SpeechBERT, etc.
Independent	DNSMOS, UTMOS, PLCMOS, SingMOS, Torch-squim (PESQ, STOI, SI-SNR), Sheet-SSQA, PAM, SpoofS, SWR, etc.
Non-matching	CER/WER (ESPnet/ESPnet-OWSM, Whisper),, CLAP-score, Log-WMSE, EMO-SIM, APA, LLR, NOMAD, LLR, etc.
Distributional	FAD, KID, Coverage, Density, KLD

Check full list and details about metrics in our website at https://github.com/wavlab-speech/versa

Codec-SUPERB

Codec SUPERB Challenge @ SLT 2024

Codec Speech processing Universal PERformance Benchmark Challenge

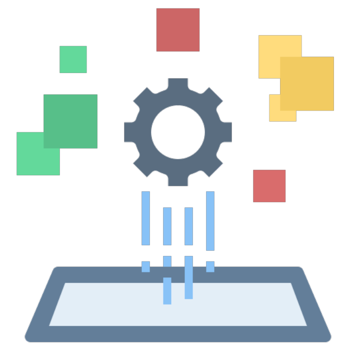
Evaluating the Codec Models on Reconstruction Quality

- Speech recognition (word error rate)
- Emotion recognition (recognition accuracy)
- Speaker verification (equal error rate)
- Audio event classification (classification accuracy)

Downstream Applications

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

- Strong backbone downstream models supported for
  - Speech recognition → WER
  - Text-to-speech → WER, UTMOS, speaker similarity (SPK-SIM)
    - Non-autoregressive TTS
    - SpeechLM-style TTS
  - Speaker recognition → EER
  - Speech separation and enhancement → PESQ, STOI, DNSMOS
  - Singing voice synthesis → MCD, Semitone Accuracy (SACC), SingMOS



ESPnet

Experiments

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

- Two data settings

Dataset	Size (Hours)	Domain	Supporting Sampling Rate (Hz)
LibriTTS	560hrs	Speech	16k/24k
AMUSE	30.7khrs	Speech/Audio/Music (24.1k/4.8k/1.8k hrs)	16k/44.1k

- Three models (most widely used models in downstream applications)

Model	Sound Representation	Encoder & Decoder	Quantizer	Discriminator
SoundStream (2021)	Time domain	Conv.	RVQ	STFT Discriminator
Encodec (2023)	Time domain	Conv. + RNN	RVQ	MS-STFT Discriminator
DAC (2024)	Time domain	Conv.	RVQ (factorized + L2-normalized)	MSMPMBD

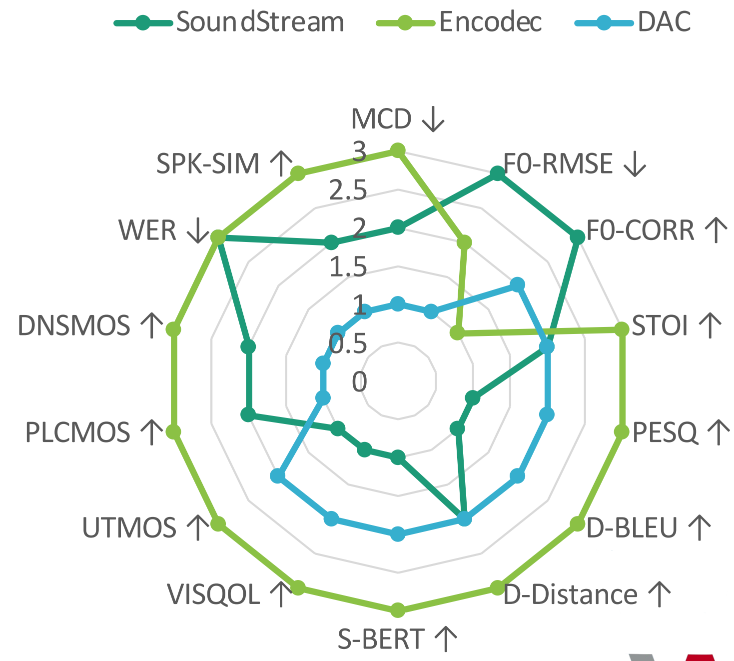
Reconstruction Experiments

In LibriTTS 16kHz setup:

- Encodec has good performance on 13 metrics
- Soundstream has good performance on 3 metrics (all F0-related)
- DAC has no best performing metrics

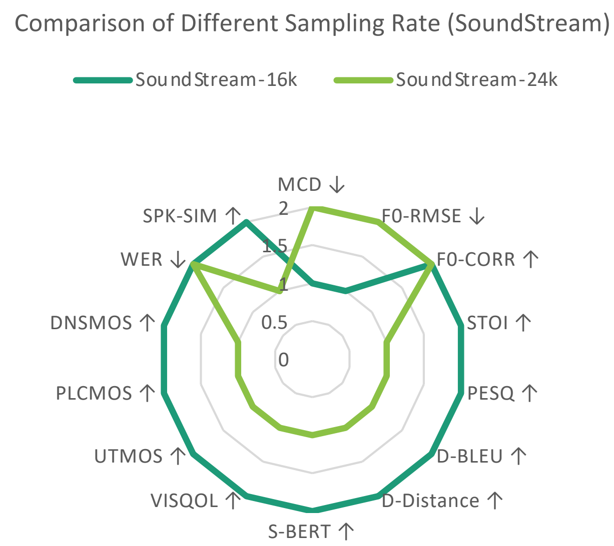
In matching experimental setups (for speech modeling), DAC **does not** demonstrate benefits over Soundstream and Encodec

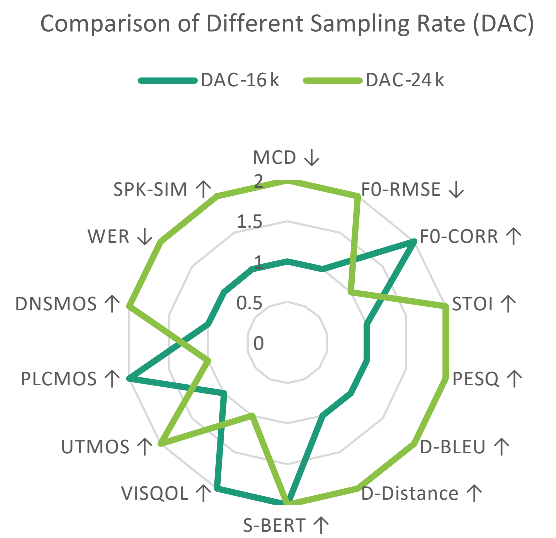
Comparison of Reconstruction Performance



Effect of Different Sampling Rate

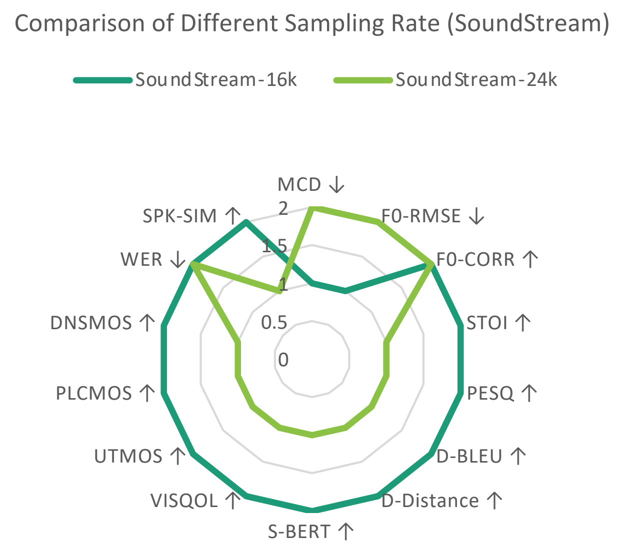
- The behavior of higher sampling rate can be dependent to models

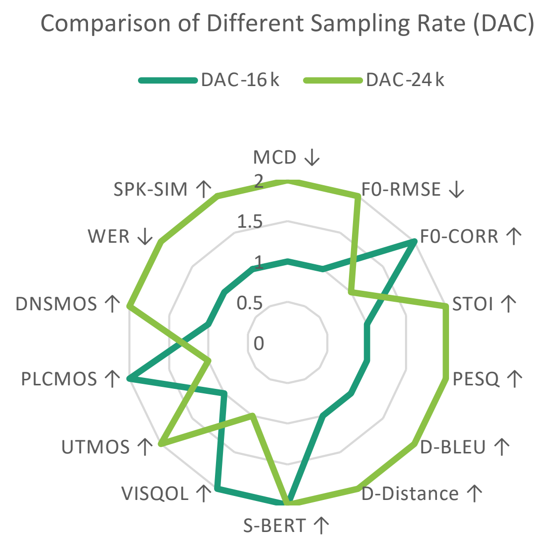




Effect of Different Sampling Rate

- The behavior of higher sampling rate can be dependent to models





Simple STFT discriminator are **difficult to expand for high sampling rate** modeling than MSMPMB discrimiantor

Reconstruction Experiments (Codec-SUPERB)

Model	Training Data	ASR (WER ↓)	SPK (EER↓)	ER (ACC ↑)	AEC (ACC ↑)
SoundStream	LibriTTS	5.99	2.8	49.49	51.13
SoundStream+	AMUSE	5.88	2.4	53.54	**68.11**
Encodec	LibriTTS	**5.58**	**2.2**	52.53	61.58
Encodec+	AMUSE	5.98	2.4	50.51	66.50
DAC	LibriTTS	6.33	3.4	**55.56**	52.65
DAC+	AMUSE	6.08	3.2	54.55	61.89
Original Audio		5.28	1.6	59.60	78.01

(+ indicated model trained with AMUSE)

Reconstruction Experiments (Codec-SUPERB)

Model	Training Data	ASR (WER ↓)	SPK (EER↓)	ER (ACC ↑)	AEC (ACC ↑)
SoundStream	LibriTTS	5.99	2.8	49.49	51.13
SoundStream+	AMUSE	5.88	2.4	53.54	**68.11**
Encodec	LibriTTS	**5.58**	**2.2**	52.53	61.58
Encodec+	AMUSE	5.98	2.4	50.51	66.50
DAC	LibriTTS	6.33	3.4	**55.56**	52.65
DAC+	AMUSE	6.08	3.2	54.55	61.89
Original Audio		5.28	1.6	59.60	78.01

(+ indicated model trained with AMUSE)

Inclusion of in-domain data could be helpful for related in-domain tasks (e.g., data in general audio domain → AEC task)

Reconstruction Experiments (Codec-SUPERB)

Model	Training Data	ASR (WER ↓)	SPK (EER↓)	ER (ACC ↑)	AEC (ACC ↑)
SoundStream	LibriTTS	5.99	2.8	49.49	51.13
SoundStream+	AMUSE	5.88	2.4	53.54	**68.11**
Encodec	LibriTTS	**5.58**	**2.2**	52.53	61.58
Encodec+	AMUSE	5.98	2.4	50.51	66.50
DAC	LibriTTS	6.33	3.4	**55.56**	52.65
DAC+	AMUSE	6.08	3.2	54.55	61.89
Original Audio		5.28	1.6	59.60	78.01

(+ indicated model trained with AMUSE)

More data (domains) can help SoundStream and DAC to perform better reconstruction → but not always for Encodec

Reconstruction Experiments (Codec-SUPERB)

Model	Training Data	ASR (WER ↓)	SPK (EER↓)	ER (ACC ↑)	AEC (ACC ↑)
SoundStream	LibriTTS	5.99	2.8	49.49	51.13
SoundStream+	AMUSE	5.88	2.4	53.54	**68.11**
Encodec	LibriTTS	**5.58**	**2.2**	52.53	61.58
Encodec+	AMUSE	5.98	2.4	50.51	66.50
DAC	LibriTTS	6.33	3.4	**55.56**	52.65
DAC+	AMUSE	6.08	3.2	54.55	61.89
Original Audio		5.28	1.6	59.60	78.01

(+ indicated model trained with AMUSE)

Potential causes:
- The use of RNN in its encoder (difficulty for modeling general audio/music with RNN)

Downstream Application (ASR + TTS)

	ASR	NAR-TTS			AR-TTS		
Model	**WER ↓**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**
SoundStream	3.7	3.4	2.34	0.58	6.7	3.70	0.63
Encodec	3.6	4.3	2.35	0.59	7.7	3.85	0.63
DAC	3.6	5.2	2.12	0.54	10.2	3.75	0.66
SoundStream+	3.7	4.7	1.84	0.57	9.8	2.97	0.61
Encodec+	3.9	5.4	1.92	0.58	8.6	2.32	0.62
DAC+	4.1	6.2	1.82	0.56	15.9	2.94	0.64

(+ indicated model trained with AMUSE)

Downstream Application (ASR + TTS)

	ASR	NAR-TTS			AR-TTS		
Model	**WER ↓**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**
SoundStream	3.7	3.4	2.34	0.58	6.7	3.70	0.63
Encodec	3.6	4.3	2.35	0.59	7.7	3.85	0.63
DAC	3.6	5.2	2.12	0.54	10.2	3.75	0.66
SoundStream+	3.7	4.7	1.84	0.57	9.8	2.97	0.61
Encodec+	3.9	5.4	1.92	0.58	8.6	2.32	0.62
DAC+	4.1	6.2	1.82	0.56	15.9	2.94	0.64

(+ indicated model trained with AMUSE)

Compared to model trained on AMUSE, models trained on LibriTTS suggest better performance on almost all metrics

Downstream Application (ASR + TTS)

	ASR	NAR-TTS			AR-TTS		
Model	**WER ↓**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**
SoundStream	3.7	3.4	2.34	0.58	6.7	3.70	0.63
Encodec	3.6	4.3	2.35	0.59	7.7	3.85	0.63
DAC	3.6	5.2	2.12	0.54	10.2	3.75	0.66
SoundStream+	3.7	4.7	1.84	0.57	9.8	2.97	0.61
Encodec+	3.9	5.4	1.92	0.58	8.6	2.32	0.62
DAC+	4.1	6.2	1.82	0.56	15.9	2.94	0.64

(+ indicated model trained with AMUSE)

Inclusion of data from other domains can **degrade** the performance on ASR, TTS → reasonable due to more information to compress with limited bandwidth

Downstream Application (ASR + TTS)

	ASR	NAR-TTS			AR-TTS		
Model	WER ↓	WER ↓	UTMOS ↑	SPK-SIM ↑	WER ↓	UTMOS ↑	SPK-SIM ↑
SoundStream	3.7	3.4	2.34	0.58	6.7	3.70	0.63
Encodec	3.6	4.3	2.35	0.59	7.7	3.85	0.63
DAC	3.6	5.2	2.12	0.54	10.2	3.75	0.66
SoundStream+	3.7	4.7	1.84	0.57	9.8	2.97	0.61
Encodec+	3.9	5.4	1.92	0.58	8.6	2.32	0.62
DAC+	4.1	6.2	1.82	0.56	15.9	2.94	0.64

(+ indicated model trained with AMUSE)

Comparison between NAR-TTS and AR-TTS:
- Better intelligibility from NAR-TTS
- Better naturalness from AR-TTS

Downstream Application (ASR + TTS)

	ASR	NAR-TTS			AR-TTS		
Model	**WER ↓**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**
SoundStream	3.7	3.4	2.34	0.58	6.7	3.70	0.63
Encodec	3.6	4.3	2.35	0.59	7.7	3.85	0.63
DAC	3.6	5.2	2.12	0.54	10.2	3.75	0.66
SoundStream+	3.7	4.7	1.84	0.57	9.8	2.97	0.61
Encodec+	3.9	5.4	1.92	0.58	8.6	2.32	0.62
DAC+	4.1	6.2	1.82	0.56	15.9	2.94	0.64

(+ indicated model trained with AMUSE)

Non-autoregressive modeling -> easy control for content, difficult control for style
Autoregressive modeling -> difficult control for content, better control for speech progression

Downstream Application (ASR + TTS)

	ASR	NAR-TTS			AR-TTS		
Model	**WER ↓**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**
SoundStream	3.7	3.4	2.34	0.58	6.7	3.70	0.63
Encodec	3.6	4.3	2.35	0.59	7.7	3.85	0.63
DAC	3.6	5.2	2.12	0.54	10.2	3.75	0.66
SoundStream+	3.7	4.7	1.84	0.57	9.8	2.97	0.61
Encodec+	3.9	5.4	1.92	0.58	8.6	2.32	0.62
DAC+	4.1	6.2	1.82	0.56	15.9	2.94	0.64

(+ indicated model trained with AMUSE)

Compared to SoundStream, Encodec can offer better quality for speech, but at the risk of making it less clear for content modeling

Downstream Application (ASR + TTS)

	ASR	NAR-TTS			AR-TTS		
Model	**WER ↓**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**	**WER ↓**	**UTMOS ↑**	**SPK-SIM ↑**
SoundStream	3.7	3.4	2.34	0.58	6.7	3.70	0.63
Encodec	3.6	4.3	2.35	0.59	7.7	3.85	0.63
DAC	3.6	5.2	2.12	0.54	10.2	3.75	0.66
SoundStream+	3.7	4.7	1.84	0.57	9.8	2.97	0.61
Encodec+	3.9	5.4	1.92	0.58	8.6	2.32	0.62
DAC+	4.1	6.2	1.82	0.56	15.9	2.94	0.64

(+ indicated model trained with AMUSE)

Dilemma for Codec:
Better reconstruction quality vs. More focuses on fundamental content information

Downstream Application (Speaker, Enhancement, and Singing Synthesis)

	SPK	SSE			SVS		
Model	EER ↓	PESQ ↑	STOI ↑	DNSMOS ↑	MCD ↓	SACC ↑	SingMOS ↑
SoundStream	27.5	1.79	0.73	0.93			
Encodec	15.7	1.85	0.76	0.95			
DAC	29.5	1.73	0.74	0.87			
SoundStream+	15.9	1.76	0.73	0.81	8.82	0.60	2.92
Encodec+	14.1	1.24	0.51	0.62	8.51	0.58	2.86
DAC+	24.6	2.00	0.78	0.87	9.26	0.49	2.58

Cannot work well from reconstruction

(+ indicated model trained with AMUSE)

Downstream Application (Speaker, Enhancement, and Singing Synthesis)

	SPK	SSE			SVS		
Model	EER ↓	PESQ ↑	STOI ↑	DNSMOS ↑	MCD ↓	SACC ↑	SingMOS ↑
SoundStream	27.5	1.79	0.73	0.93	-	-	-
Encodec	15.7	1.85	0.76	0.95	-	-	-
DAC	29.5	1.73	0.74	0.87	-	-	-
SoundStream+	15.9	1.76	0.73	0.81	8.82	0.60	2.92
Encodec+	14.1	1.24	0.51	0.62	8.51	0.58	2.86
DAC+	24.6	2.00	0.78	0.87	9.26	0.49	2.58

(+ indicated model trained with AMUSE)

Inclusion of data from other domains can **improve** the performance on SPK, SVS → comes from the more diverse information (a trade-off with the finding of ASR/TTS)

Summary

- The comparison of audio codecs presents **multiple complexities** due to:
  - Diverse downstream applications (understanding vs. generation)
  - Various domains in data (speech, music, general audio)
  - Numerous training hyperparameters (many significantly impacting model performance)

- ESPnet-Codec addresses these challenges by providing:
  - Reproducible modeling frameworks
  - Seamless integration with state-of-the-art downstream systems
  - Ongoing development of comprehensively tuned models for optimal performance

Note: We are continuously adding more models with extensive tuning to achieve best-in-class performance. Please stay tuned for updates!

Summary

- The comparison of audio codecs presents **multiple complexities** due to:
  - Diverse downstream applications (understanding vs. generation)
  - Various domains in data (speech, music, general audio)
  - Numerous training hyperparameters (many significantly impacting model performance)

- ESPnet-Codec addresses these ch
  - Reproducible modeling framework
  - Seamless integration with state-of-
  - Ongoing development of comprehe

ESPnet-Codec has **been continuously providing support** to various applications in ESPnet, especially the recent SpeechLM project

Note: We are continuously adding more models with extensive tuning to achieve best-in-class performance. Please stay tuned for updates!

Future Works

- Cross-domain adaptation techniques for codec modeling
  - Developing methods to adapt codecs trained on one audio domain (e.g., speech) to perform well on others (e.g., music, environmental sounds)
- End-to-end optimization frameworks
  - Designing systems that jointly optimize codecs with downstream tasks (ASR, TTS, etc.) rather than treating them as separate components
- Community benchmarking platform
  - Establishing standardized evaluation frameworks and datasets to enable fair comparison between different codec approaches

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!