
# Conversational AI Reading Group

Every Thursday 11AM- 12PM EST

Webpage

Youtube

Slack

# Discrete Audio Tokens: More Than a Survey!

Pooneh Mousavi

Concordia - Mila

Sep 18, 2025

**Conversational AI Reading Group**

# Meet the Team

## Core Contributors


**Pooneh Mousavi**
Concordia University, Mila

**Gallil Maimon**
The Hebrew University of Jerusalem

**Adel Moumen**
University of Cambridge

**Darius Petermann**
Indiana University

**Jiatong Shi**
Carnegie Mellon University

**Haibin Wu**
Microsoft

**Haici Yang**
Indiana University

**Anastasia Kuznetsova**
Indiana University


## Collaborators


**Artem Ploujnikov**
Université de Montréal, Mila

**Ricard Marxer**
Université de Toulon

**Bhuvana Ramabhadran**
Google

**Benjamin Elizalde**
Apple

**Loren Lugosch**
Apple

**Jinyu Li**
Microsoft

**Cem Subakan**
Laval University, Mila

**Phil Woodland**
University of Cambridge

**Minje Kim**
University of Illinois at Urbana-Champaign

**Hung-yi Lee**
National Taiwan University

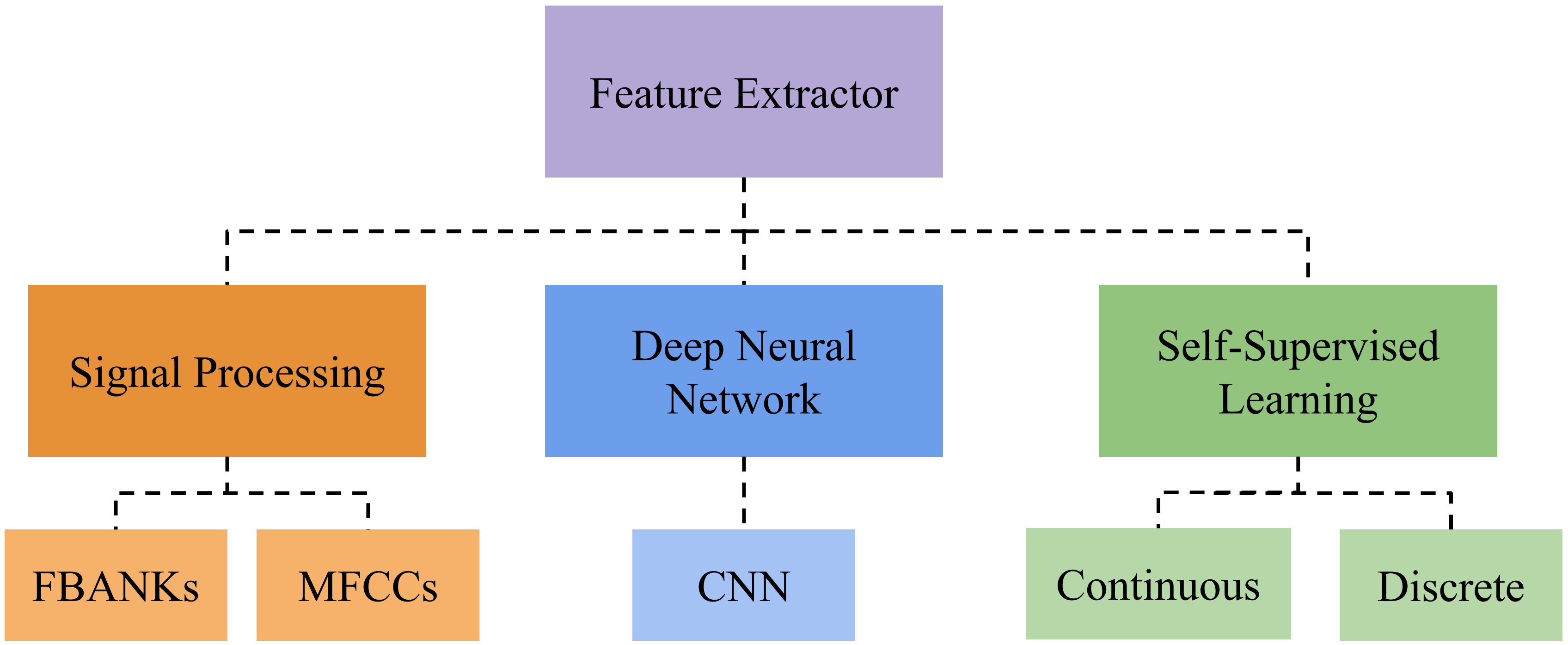
**Shinji Watanabe**
Carnegie Mellon University

**Yossi Adi**
The Hebrew University of Jerusalem

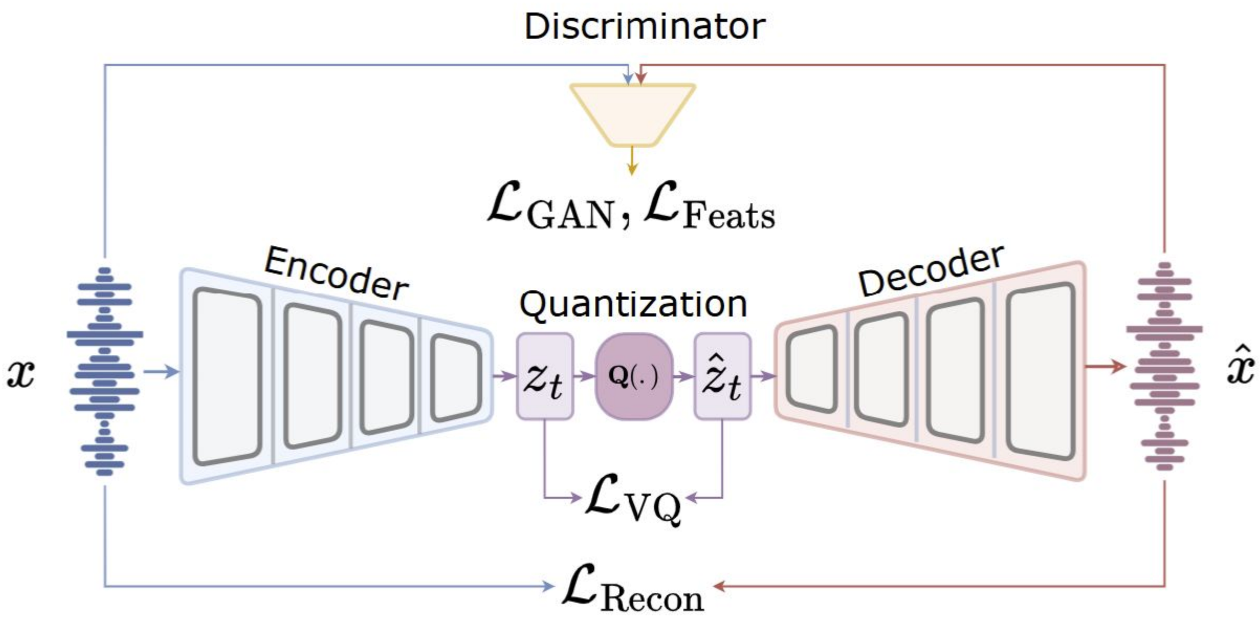
**Mirco Ravanelli**
Concordia University, Mila

# Background

# Taxonomy of Speech Feature Extractor



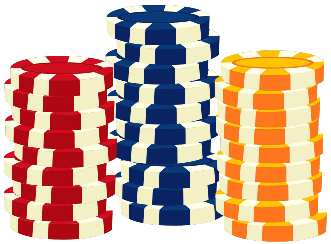
# Overall Architecture of a Standard Audio Tokenizer.



# Audio Tokens

- Ideal audio tokens must preserve content, paralinguistic elements, speaker identity, and many other audio details.

- Benefits of discrete audio tokens:
  - Storage benefits
  - Efficient transmission
  - Simplify audio generation task
  - Faster inference
  - Easier integration to LLMs and multimodal models



# Motivation

- Traditional audio codecs rely heavily on domain knowledge, combining signal processing pipelines with hand-crafted components to achieve efficient but lossy compression.

# Motivation

- Traditional audio codecs rely heavily on domain knowledge, combining signal processing pipelines with hand-crafted components to achieve efficient but lossy compression.

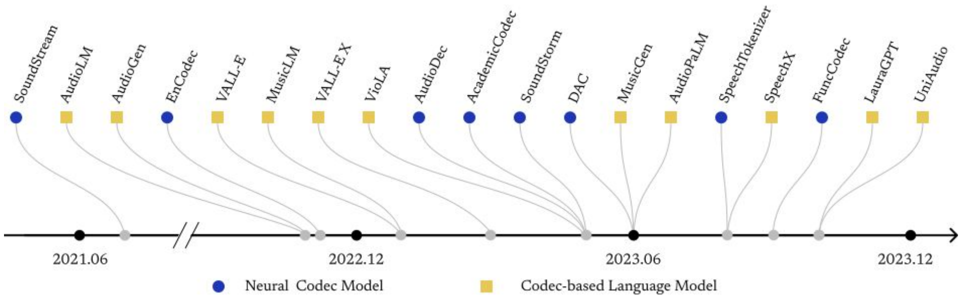
- This has motivated a shift toward data-driven approaches with deep learning, known as neural codecs.

# Motivation

- Traditional audio codecs rely heavily on domain knowledge, combining signal processing pipelines with hand-crafted components to achieve efficient but lossy compression.

- This has motivated a shift toward data-driven approaches with deep learning, known as neural codecs.

- Many audio tokenizers are proposed in last 3 years.




Adopted from Codec SUPERB

# Our contribution is organized into three core studies!

Taxonomy

# Our contribution is organized into three core studies!

Taxonomy

Benchmarking

# Our contribution is organized into three core studies!

Taxonomy

Benchmarking

Ablation Study

# Our contribution is organized into three core studies!

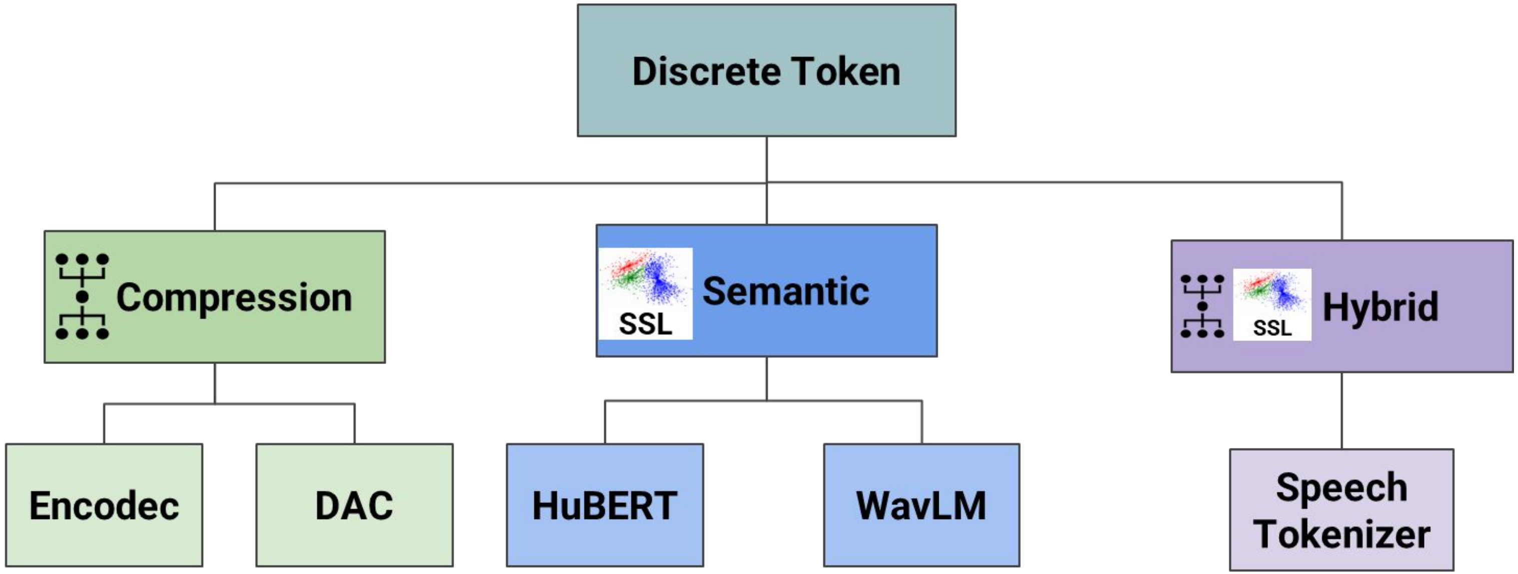
Taxonomy

Benchmarking

Ablation Study

# Study 1:
# Audio Tokenizer Taxonomy

# Traditional Taxonomy of Speech Discrete Tokenizer



# What is the Problem?

- We argue the common division of discrete tokens into acoustic and semantic categories has notable limitations.

# What is the Problem?

- We argue the common division of discrete tokens into acoustic and semantic categories has notable limitations.

- Acoustic tokenizers can capture semantic information , while semantic tokenizers have been effectively used in generative task.

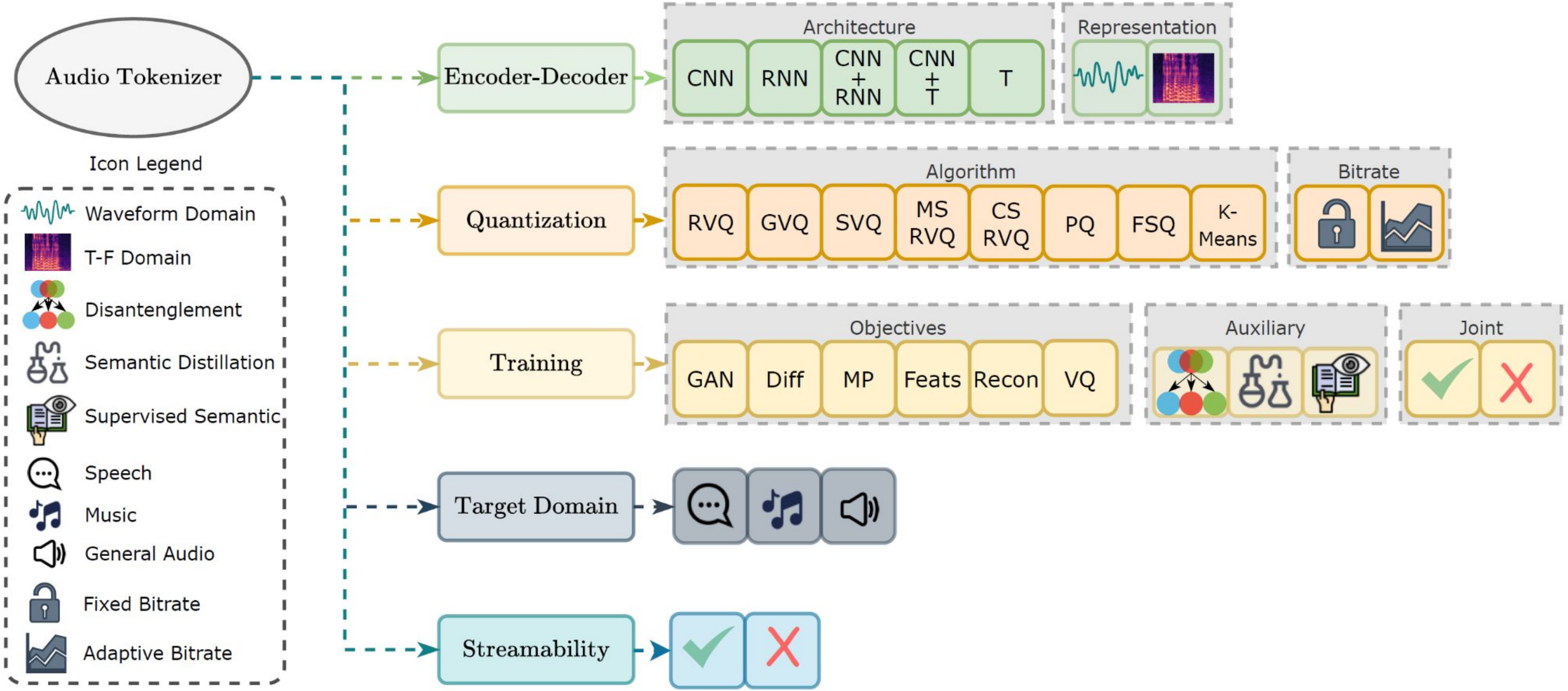
# What is the Problem?

- We argue the common division of discrete tokens into acoustic and semantic categories has notable limitations.
- Acoustic tokenizers can capture semantic information , while semantic tokenizers have been effectively used in generative task.
- This overlap blurs the boundary between the two categories

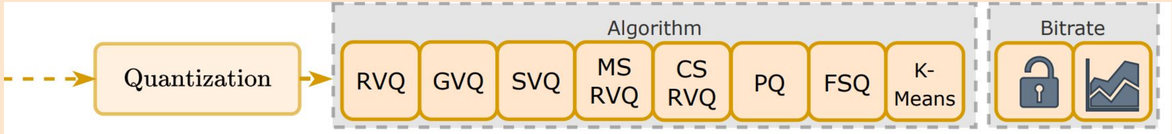
# What is the Problem?

- We argue the common division of discrete tokens into acoustic and semantic categories has notable limitations.
- Acoustic tokenizers can capture semantic information , while semantic tokenizers have been effectively used in generative task.
- This overlap blurs the boundary between the two categories
- It fail to capture key architectural differences and practical tradeoffs.

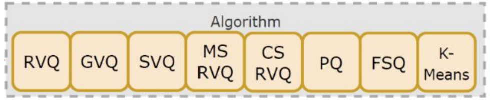
# Refined Taxonomy of Speech Discrete Tokenizer



# Quantization



# Quantization Algorithm

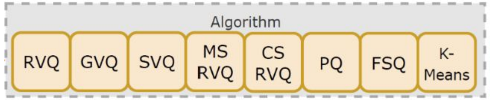



K-means :

- It is frequently used for post-training quantization.
- Select one or more layers from a pretrained SSL model → apply offline k-means clustering → assign cluster IDs as discrete tokens.

$$q_t = \arg \min_{k \in \{1,\ldots,K\}} \|z_t - c_k\|^2$$

# Quantization Algorithm



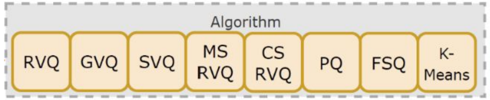

## Residual Vector Quantization (RVQ)

- RVQ maps each frame-wise feature to the closest entry in a codebook and then refines this process by computing the residual after quantization.

**Algorithm 1** Residual Vector Quantization (RVQ)

**Input:** Embedding $z_t$, Codebooks $\{\mathcal{C}^{(m)}\}_{m=1}^{M}$
Initialize residual: $r_t^{(1)} \leftarrow z_t$
**for** $m = 1$ to $M$ **do**
    $q_t^{(m)} \leftarrow \arg\min_k \left\| r_t^{(m)} - c_k^{(m)} \right\|^2$
    $\hat{z}_t^{(m)} \leftarrow c_{q_t^{(m)}}^{(m)}$
    $r_t^{(m+1)} \leftarrow r_t^{(m)} - \hat{z}_t^{(m)}$
**end for**
**Output:** $\hat{z}_t \leftarrow \sum_{m=1}^{M} \hat{z}_t^{(m)}$

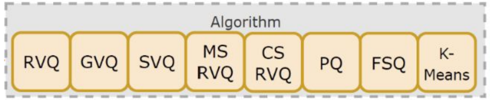
# Quantization Algorithm




Single Vector Quantization (SVQ) :

- Use a single codebook for quantization
- is simpler and particularly useful for training acoustic language models.
- To compensate for the potential loss of information → adopt larger codebook sizes.

# Quantization Algorithm




Group Vector Quantization (GVQ).

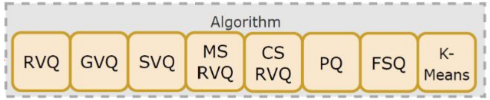
- Increases capacity at the first quantization stage by dividing the latent feature.

$$z_t = \left[ z_t^{(1)} \parallel z_t^{(2)} \parallel \ldots \parallel z_t^{(G)} \right],$$

- Quantized independently using a separate RVQ module.

$$\hat{z}_t = \left[ \hat{z}_t^{(1)} \parallel \hat{z}_t^{(2)} \parallel \ldots \parallel \hat{z}_t^{(G)} \right].$$

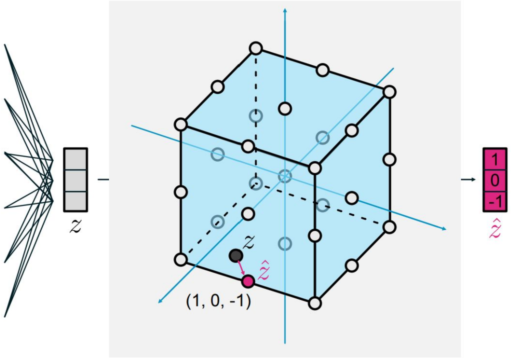
# Quantization Algorithm




Finite Scalar Quantization (FSQ).

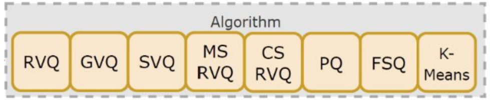
- FSQ maps each dimension of a feature vector to a fixed set of scalar values.
- No embedding saved for codebooks.

FSQ



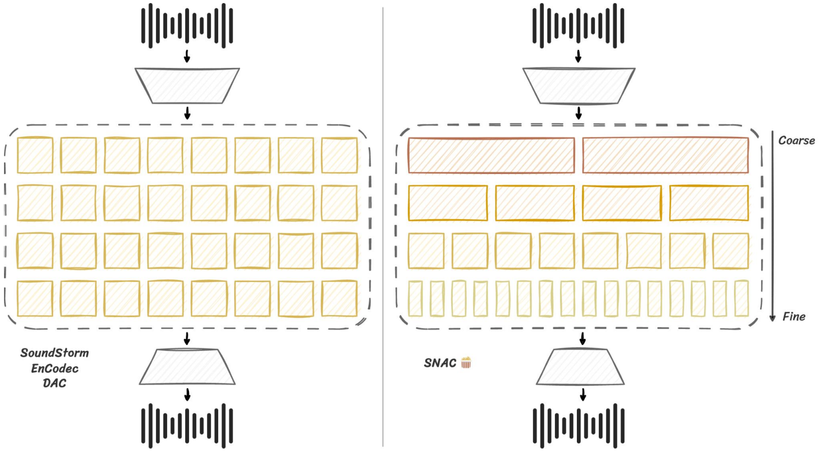

Adopted from Finite Scalar Quantization: VQ-VAE Made Simple

# Quantization Algorithm



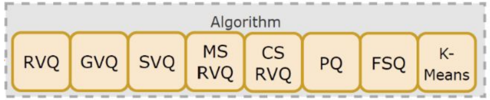

Multi-Scale RVQ (MSRVQ).

- Extends standard RVQ by applying quantizers at different temporal resolutions.



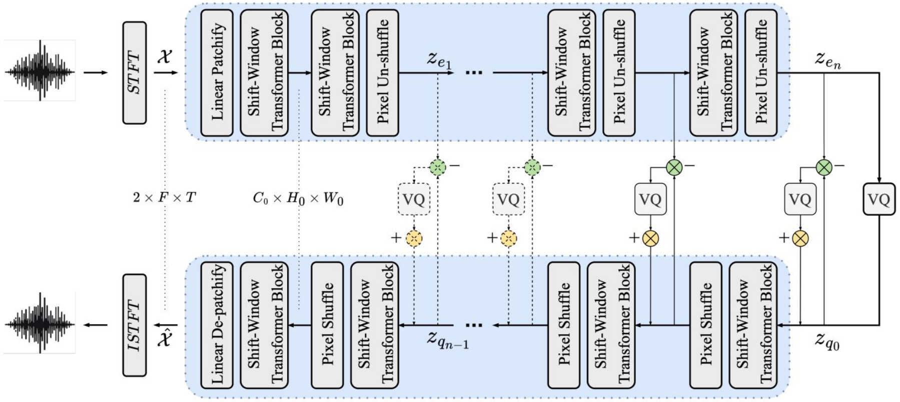

Adopted from SNAC: Multi-Scale Neural Audio Codec

# Quantization Algorithm



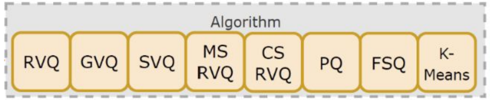

Cross-Scale RVQ (CSRVQ).

- Encode residuals between encoder and decoder features at multiple hierarchical levels.




Adopted from ESC: Efficient Speech Coding with Cross-Scale Residual Vector Quantized Transformers

# Quantization Algorithm




Product Quantization (PQ).

- Commonly used in self-supervised learning (SSL)
- Partition embeddings into smaller subvectors and using Random-Projection Quantization.

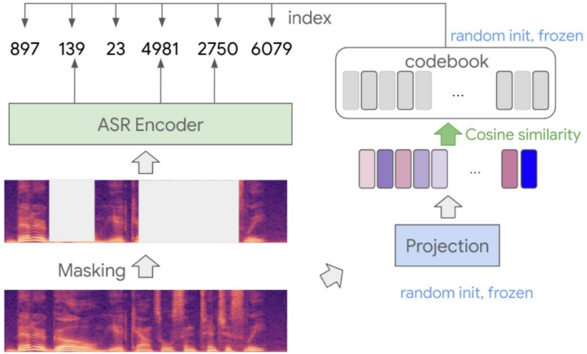
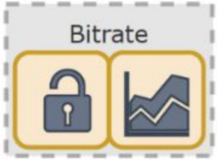


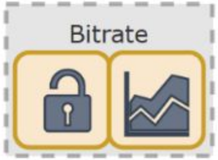

Figure 3: BEST-RQ based pre-training with conformer encoder.

# Fixed vs. Adaptive Bitrate




- **Fixed bitrate**, such as those based on codebooks, the bitrate is determined by the number of bits required to represent each code index, irrespective of the actual token distribution.

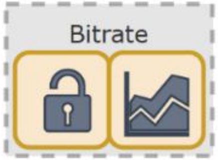
# Fixed vs. Adaptive Bitrate




- **Fixed bitrate** is determined by the number of bits required to represent each code index, irrespective of the actual token distribution.

- **Adaptive bitrate** refers to entropy-based coding schemes that assign variable-length codes based on the statistical frequency of tokens.

# Fixed vs. Adaptive Bitrate

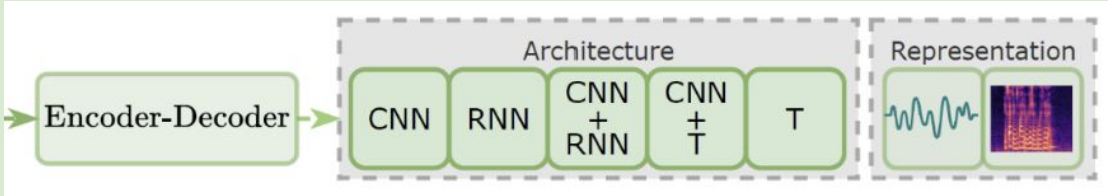



- **Fixed bitrate** is determined by the number of bits required to represent each code index, irrespective of the actual token distribution.

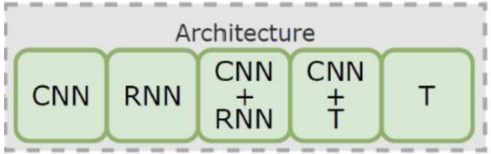
- **Adaptive bitrate** refers to entropy-based coding schemes that assign variable-length codes based on the statistical frequency of tokens.

- It is also important to distinguish between **adaptive** bitrate and **scalable** bitrate.

# Encoder-Decoder

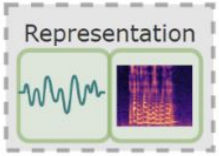


# Architecture



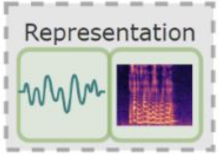

- Convolutional (CNN)
- Convolutional + RNN (CNN+RNN).
- Transformer (T).
- Convolutional + Transformer (CNN+T).

# Input and Output Representations




- Encoders can process audio inputs in either the time or the frequency domain.

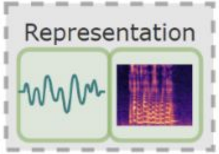
# Input and Output Representations




- Encoders can process audio inputs in either the time or the frequency domain.

- The output representation can follow two Approaches:

# Input and Output Representations

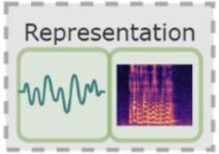



- Encoders can process audio inputs in either the time or the frequency domain.

- The output representation can follow two Approaches:

    1. **Time domain waveforms**, where the decoder directly upsamples the discrete representation into waveforms.

# Input and Output Representations

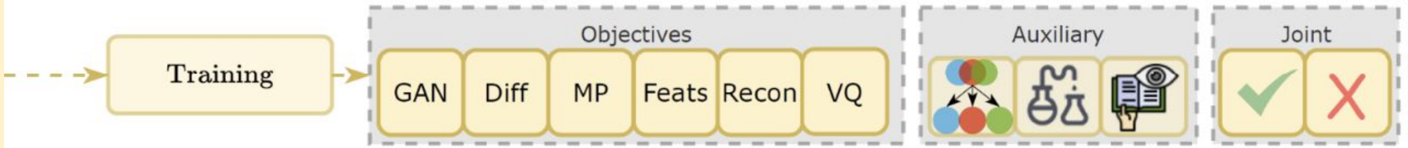



- Encoders can process audio inputs in either the time or the frequency domain.

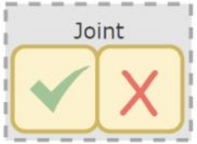
- The output representation can follow two Approaches:

  1. **Time domain waveforms**, where the decoder directly upsamples the discrete representation into waveforms.
  2. **Time-frequency** domain features, where the decoder outputs time-frequency domain features and the Inverse Short-Time Fourier Transform (ISTFT) is applied for upsampling.

# Training Paradigm

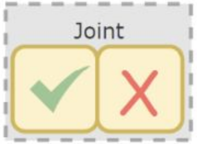


# Training Strategies




- **Separate (Post-Training)**. The encoder and decoder are optimized independently from the quantization module.

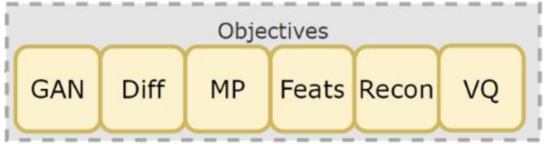
# Training Strategies




- **Separate (Post-Training)**. The encoder and decoder are optimized independently from the quantization module.

- **Joint (End-to-End Training)**. The encoder, quantizer, and decoder are optimized simultaneously within a unified end-to-end framework.

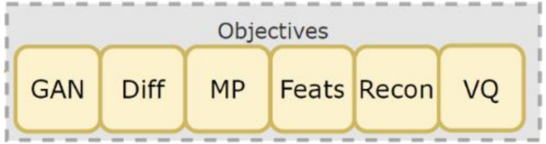
# Main Training Objectives:




- Reconstruction (Recon).

$$\mathcal{L}_{\text{Recon}} = \sum_{t=1}^{T} \left\| x_t - \hat{x}_t \right\|^2 .$$

# Main Training Objectives:




- Reconstruction (Recon).

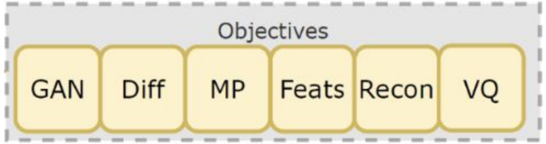
$$\mathcal{L}_{\text{Recon}} = \sum_{t=1}^{T} \|x_t - \hat{x}_t\|^2 .$$

- Vector Quantization (VQ).

$$\mathcal{L}_{VQ} = \|z - \hat{z}\|, \ \hat{z} = \sum_{m=1}^{M} \alpha_m * c_m,$$

$$\mathcal{L}_{\text{VQ}} = \sum_{t=1}^{T} \sum_{m=1}^{M} \left\| z_t^{(m)} - \text{sg}\left[ \hat{z}_t^{(m)} \right] \right\|^2 ,$$

# Main Training Objectives:




- Reconstruction (Recon).

$$\mathcal{L}_{\text{Recon}} = \sum_{t=1}^{T} \|x_t - \hat{x}_t\|^2 .$$

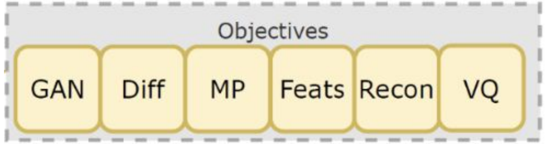
- Vector Quantization (VQ).

$$\mathcal{L}_{VQ} = \|z - \hat{z}\|, \ \hat{z} = \sum_{m=1}^{M} \alpha_m * c_m, \qquad \mathcal{L}_{\text{VQ}} = \sum_{t=1}^{T} \sum_{m=1}^{M} \left\| z_t^{(m)} - \text{sg}\left[\hat{z}_t^{(m)}\right] \right\|^2 ,$$

- Adversarial (GAN)

$$\mathcal{L}_G = \frac{1}{K} \sum_{k=1}^{K} \max(1 - D_k(\hat{x}), 0) \qquad \mathcal{L}_D = \frac{1}{K} \sum_{k=1}^{K} [\max(1 - D_k(x), 0) + \max(1 + D_k(\hat{x}), 0)]$$

# Main Training Objectives:




- Reconstruction (Recon).

$$\mathcal{L}_{\text{Recon}} = \sum_{t=1}^{T} \|x_t - \hat{x}_t\|^2 .$$

- Vector Quantization (VQ).

$$\mathcal{L}_{\text{VQ}} = \sum_{t=1}^{T} \sum_{m=1}^{M} \left\| z_t^{(m)} - \text{sg}\left[\hat{z}_t^{(m)}\right] \right\|^2 ,$$

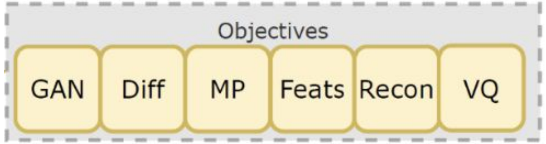
- Adversarial (GAN)

$$\mathcal{L}_G = \frac{1}{K} \sum_{k=1}^{K} \max(1 - D_k(\hat{x}), 0) \qquad \mathcal{L}_D = \frac{1}{K} \sum_{k=1}^{K} [\max(1 - D_k(x), 0) + \max(1 + D_k(\hat{x}), 0)]$$

- Feature Matching (Feat).

$$\mathcal{L}_{\text{Feats}} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\|D_k^l(x) - D_k^l(\hat{x})\|_1}{\text{mean}(\|D_k^l(x)\|_1)},$$

# Main Training Objectives:




- Reconstruction (Recon).

$$\mathcal{L}_{\text{Recon}} = \sum_{t=1}^{T} \left\| x_t - \hat{x}_t \right\|^2 .$$

- Vector Quantization (VQ).

$$\mathcal{L}_{\text{VQ}} = \sum_{t=1}^{T} \sum_{m=1}^{M} \left\| z_t^{(m)} - \text{sg}\left[ \hat{z}_t^{(m)} \right] \right\|^2 ,$$

- Adversarial (GAN)

$$\mathcal{L}_G = \frac{1}{K} \sum_{k=1}^{K} \max(1 - D_k(\hat{x}), 0) \qquad \mathcal{L}_D = \frac{1}{K} \sum_{k=1}^{K} [\max(1 - D_k(x), 0) + \max(1 + D_k(\hat{x}), 0)]$$

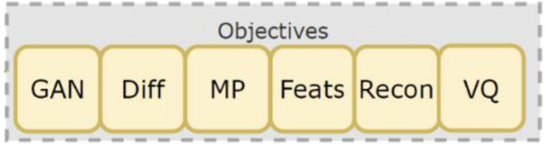
- Feature Matching (Feat).

$$\mathcal{L}_{\text{Feats}} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\| D_k^l(x) - D_k^l(\hat{x}) \|_1}{\text{mean}(\| D_k^l(x) \|_1)},$$

- Diffusion (Diff).

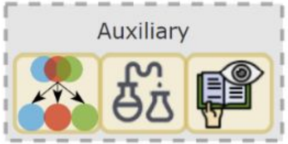
$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z_0, t, z_q} \left[ \| \epsilon_t - \epsilon_\theta(z_t, t, z_q) \| \right] ,$$

# Main Training Objectives:



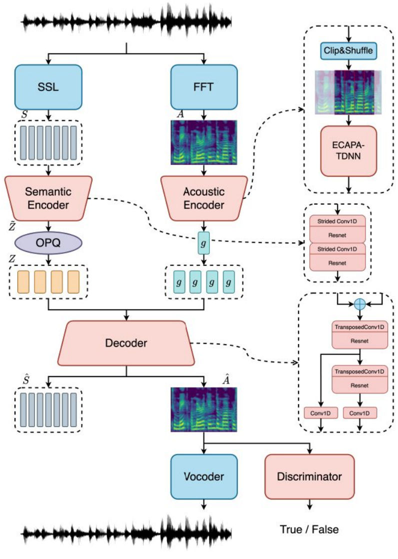

- Reconstruction (Recon). $$\mathcal{L}_{\text{Recon}} = \sum_{t=1}^{T} \left\| x_t - \hat{x}_t \right\|^2 .$$
- Vector Quantization (VQ). $$\mathcal{L}_{\text{VQ}} = \sum_{t=1}^{T} \sum_{m=1}^{M} \left\| z_t^{(m)} - \text{sg}\left[ \hat{z}_t^{(m)} \right] \right\|^2 ,$$
- Adversarial (GAN) $$\mathcal{L}_G = \frac{1}{K} \sum_{k=1}^{K} \max(1 - D_k(\hat{x}), 0) \qquad \mathcal{L}_D = \frac{1}{K} \sum_{k=1}^{K} [\max(1 - D_k(x), 0) + \max(1 + D_k(\hat{x}), 0)]$$
- Feature Matching (Feat). $$\mathcal{L}_{\text{Feats}} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\| D_k^l(x) - D_k^l(\hat{x}) \|_1}{\text{mean}(\| D_k^l(x) \|_1)},$$
- Diffusion (Diff). $$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z_0, t, z_q} \left[ \| \epsilon_t - \epsilon_\theta(z_t, t, z_q) \| \right],$$
- Masked Prediction (MP). $$\mathcal{L}_{\text{MP}} = \sum_{t=1}^{T} M_t \cdot \ell(Z_t, x_t)$$

# Auxiliary Components



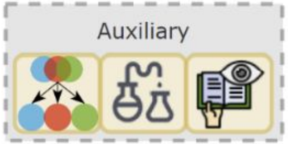

Disentanglement.

- Separate different speech attributes into distinct representations.
- Reduce redundancy while allowing independent control over acoustic properties and simplifying downstream tasks.




The model architecture of SoCodec

# Auxiliary Components




Semantic Distillation.

- Enhance codec representations by incorporating phonetic information into specific codebooks.

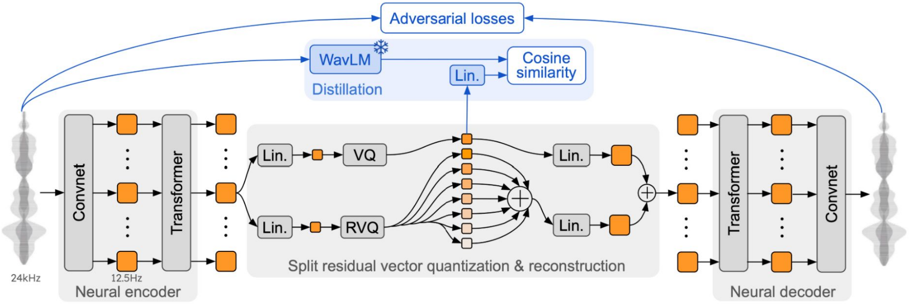
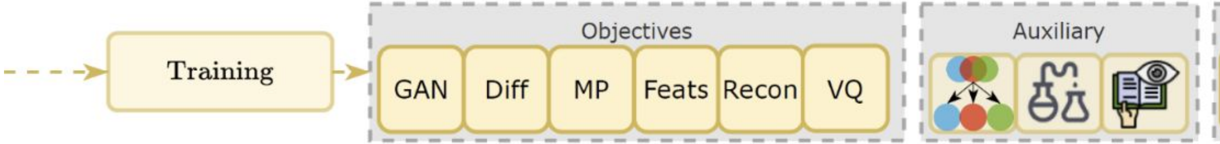



Figure 2: **Architecture and training of Mimi, our neural audio codec, with its split residual vector quantization.**

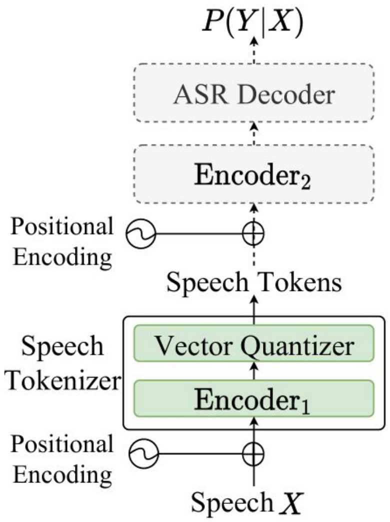
The model architecture of SoCodec

# Training Paradigm

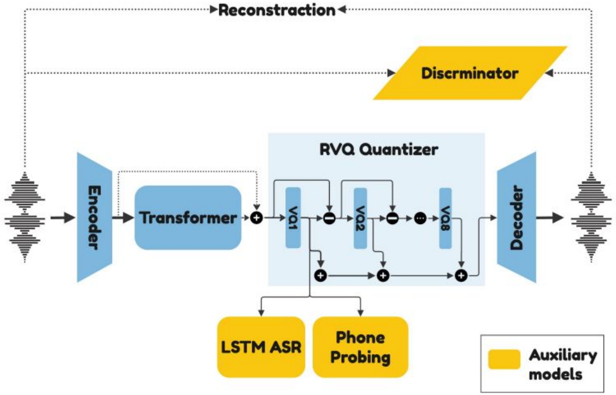



## Supervised Semantic Tokenization

- Some tokenizers explicitly capture phonetic detail through supervised training.

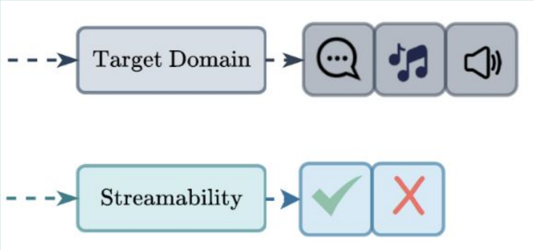



S3 Tokenizer - CosyVoice




PAST: Phonetic-Acoustic Speech Tokenizer

# Streamability and Domain Categorization

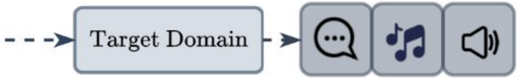


# Streamability and Domain Categorization


- **Streamability** refers to the ability of a tokenizer to process and generate audio in real-time with minimal latency, using little or no future context.

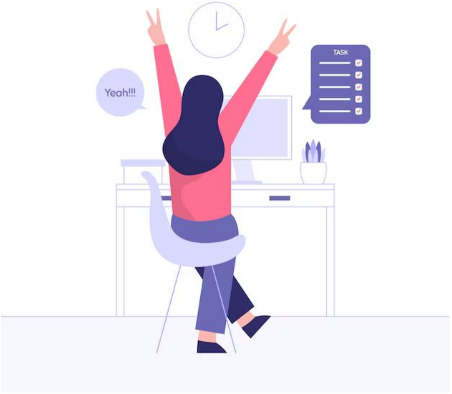
# Streamability and Domain Categorization


- **Streamability** refers to the ability of a tokenizer to process and generate audio in real-time with minimal latency, using little or no future context.




- **Target Domain.** The type of data the tokenizer is specifically trained on  e.g., speech, music, general audio or multiple domains.

# We are done with our Taxonomy Section !




We provide the database of around 70 tokenizers in our website:

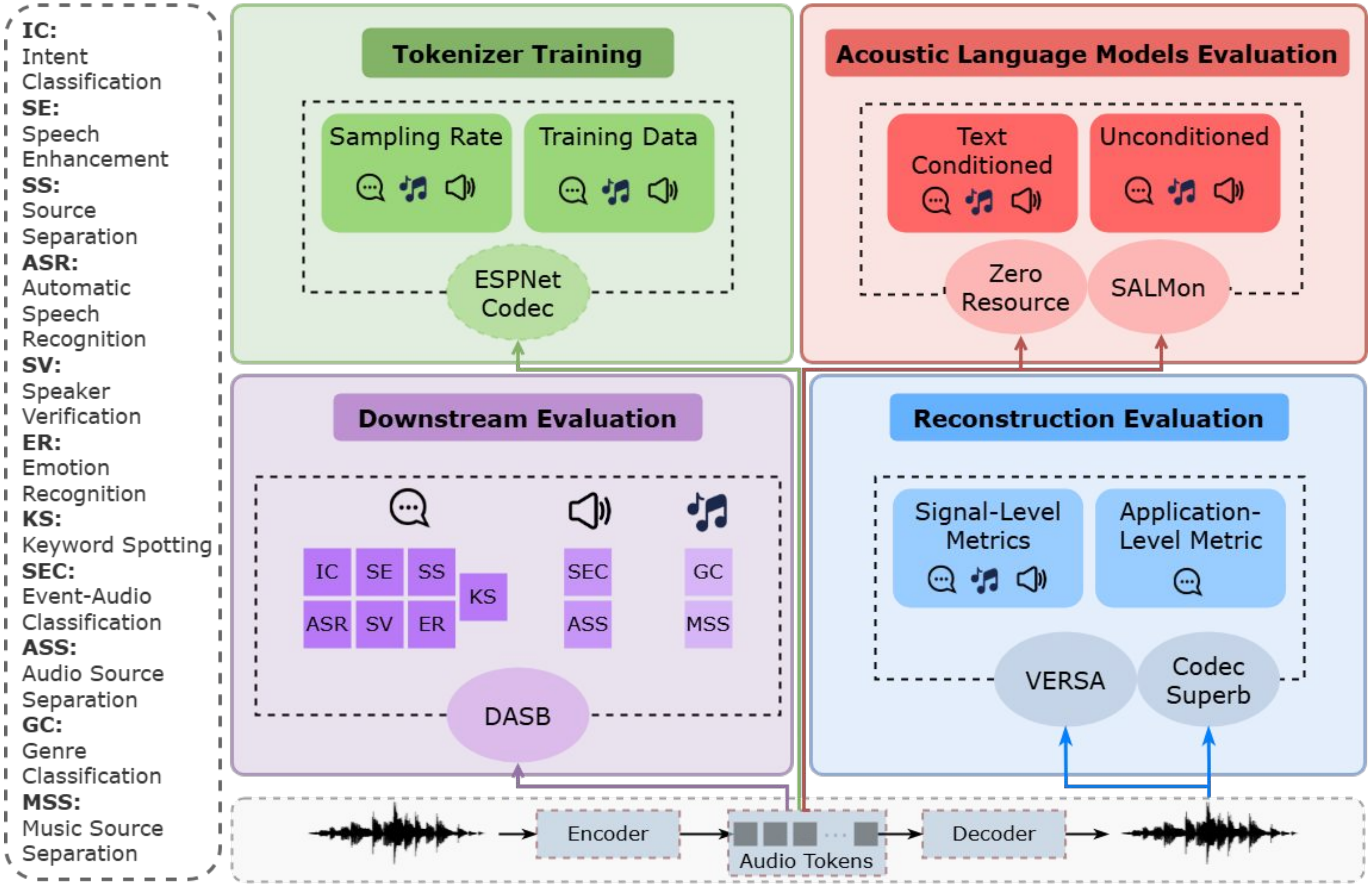
Check out our tokenizer database!

Contribute your tokenizer ➜ Fill out the form at the bottom of the page.

Want to contribute your tokenizer?

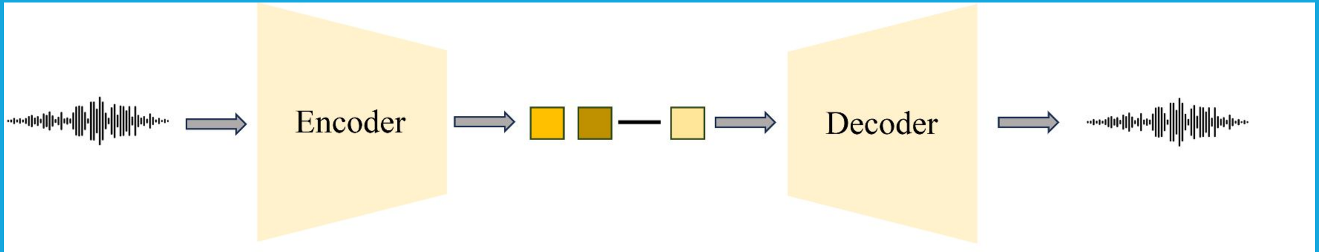
✍️ Submit a New Tokenizer

# Study 2:
# Benchmark Evaluation

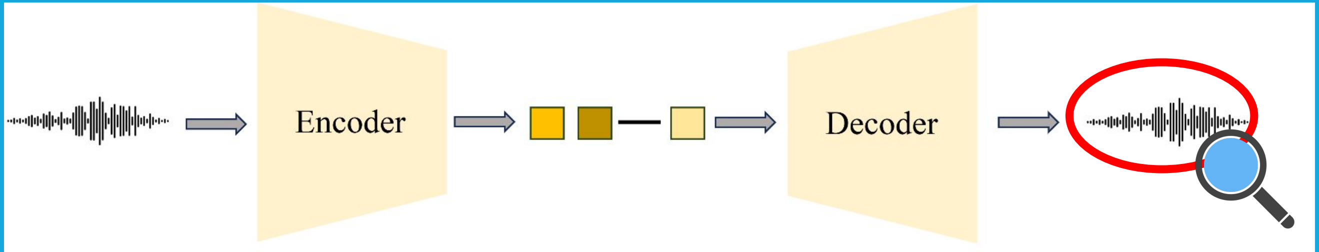

IC: Intent Classification
SE: Speech Enhancement
SS: Source Separation
ASR: Automatic Speech Recognition
SV: Speaker Verification
ER: Emotion Recognition
KS: Keyword Spotting
SEC: Event-Audio Classification
ASS: Audio Source Separation
GC: Genre Classification
MSS: Music Source Separation
Tokenizer Training
Sampling Rate
Training Data
ESPNet Codec
Acoustic Language Models Evaluation
Text Conditioned
Unconditioned
Zero Resource
SALMon
Downstream Evaluation
IC
SE
SS
KS
ASR
SV
ER
SEC
ASS
GC
MSS
DASB
Reconstruction Evaluation
Signal-Level Metrics
Application-Level Metric
VERSA
Codec Superb
Encoder
Audio Tokens
Decoder

# Audio tokenizers used throughout the study.

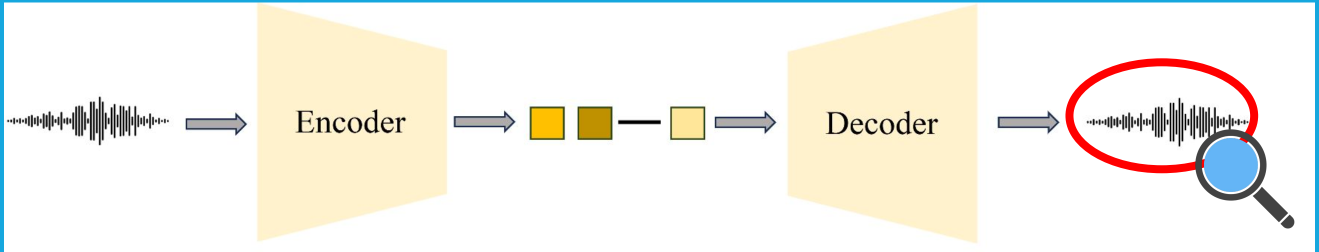
| **Tokenizer** | **Abbreviations** | **Domain** | | | **SR** | **Frame** | **#Codes** | **Params** | **MACs** | **Link** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Speech | Music | Audio | **(kHz)** | **Rate** | | **(Mil)** | **(G)** | |
| EnCodec | Enc-SMA-24 | ✓ | ✓ | ✓ | 24 | 75 | 1024 | 14.9 | 6.1 | Link |
| | Enc-M-32 | | ✓ | | 32 | 50 | 2048 | 56.9 | 14.4 | Link |
| | Enc-A-16 | | | ✓ | 16 | 50 | 2048 | 56.8 | 14.0 | Link |
| DAC | DAC-SMA-44 | ✓ | ✓ | ✓ | 44 | 86 | 1024 | 76.7 | 147.0 | Link |
| | DAC-SMA-24 | ✓ | ✓ | ✓ | 24 | 75 | 1024 | 74.7 | 83.4 | Link |
| | DAC-SMA-16 | ✓ | ✓ | ✓ | 16 | 50 | 1024 | 74.1 | 55.6 | Link |
| SpeechTokenizer | ST-S-16 | ✓ | | | 16 | 50 | 1024 | 103.7 | 17.1 | Link |
| Mimi | Mimi-S-24 | ✓ | | | 24 | 12.5 | 2048 | 79.3 | 8.1 | Link |
| Discrete-WavLM | DWavL-S-16 | ✓ | | | 16 | 50 | 1000 | 331.9 | 21.1 | Link |
| SQ-Codec | SQ-SMA-16 | ✓ | ✓ | ✓ | 16 | 50 | 19683 | 23.5 | 14.7 | Link |
| WavTokenizer | WT-SMA-24 | ✓ | ✓ | ✓ | 24 | 75 | 4096 | 80.6 | 6.3 | Link |
| | WT-S-24 | ✓ | | | 24 | 40 | 4096 | 80.9 | 3.4 | Link |




# Reconstruction Evaluation




# Reconstruction Evaluation




# Reconstruction Evaluation

Codec Superb

Versa

# Evaluation Metrics on Resynthesized Audio

Table 3: Summary of evaluation metrics on resynthesized audio.

| Metric | Functionality | Range | Domain | | |
|---|---|---|---|---|---|
| | | | Speech | Music | Audio |
| *Signal-level* | | | | | |
| SDR | Signal-to-distortion Ratio | (-inf, inf) | ✓ | ✓ | ✓ |
| SI-SNR | Scale-invariant signal-to-noise ratio | (-inf, inf) | ✓ | ✓ | ✓ |
| PESQ | Perceptual Evaluation of Speech Quality | [1, 5] | ✓ | | |
| UTMOS | UTokyo-SaruLab System for VoiceMOS 2022 | [1, 5] | ✓ | | |
| DNSMOS P808 | Deep Noise Suppression MOS Score of P.808 | [1, 5] | ✓ | | |
| DNSMOS P835 | Deep Noise Suppression MOS Score of P.835 | [1, 5] | ✓ | | |
| PLCMOS | Packet Loss Concealment-focus MOS | [1, 5] | ✓ | | |
| STOI | Short-Time Objective Intelligibility | [0, 1] | ✓ | | |
| VISQOL | Virtual Speech Quality Objective Listener | [1, 5] | | ✓ | ✓ |
| SingMOS | Singing voice MOS | [1, 5] | | ✓ | ✓ |
| *Application-level* | | | | | |
| WER | Word Error Rate (beam=5) | [0, inf) | ✓ | | |
| Spk Sim | Speaker Similarity | [-1, 1] | ✓ | | |

# Reconstruction Performance of speech.

| Tokenizer | #Q | kbps | Token rate | SDR ↑ | SI-SNR↑ | PESQ ↑ | UTMOS ↑ | DNSMOS P808↑ | DNSMOS P835↑ | PLCMOS ↑ | STOI ↑ | WER ↓ | Spk Sim↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ground truth | - | - | - | **290.16** | **55.92** | **4.64** | **4.09** | **3.84** | 3.18 | 4.16 | **1.00** | 2.83 | **1.00** |
| Enc-SMA-24 | 2 | 1.5 | 150 | 0.82 | −1.53 | 1.56 | 1.58 | 3.21 | 2.39 | 3.44 | 0.85 | 5.44 | 0.42 |
| | 8 | 6 | 600 | 6.50 | 4.83 | 2.77 | 3.09 | 3.57 | 2.96 | 4.08 | 0.94 | 2.78 | 0.72 |
| | 32 | 24 | 2400 | **9.75** | **7.90** | 3.71 | 3.74 | 3.74 | 3.19 | 4.29 | 0.97 | 2.77 | 0.78 |
| DAC-SMA-24 | 2 | 1.5 | 150 | -0.57 | −8.40 | 1.48 | 1.68 | 3.24 | 2.61 | 3.27 | 0.83 | 9.59 | 0.45 |
| | 8 | 6 | 600 | 1.79 | −9.51 | 3.40 | 3.60 | 3.69 | 3.16 | 4.15 | 0.95 | 3.53 | 0.73 |
| | 32 | 24 | 2400 | 2.20 | −9.47 | **4.45** | **4.05** | 3.78 | 3.20 | 4.40 | **0.99** | 2.72 | 0.80 |
| ST-S-16 | 2 | 1 | 100 | -7.10 | −14.46 | 1.21 | 2.32 | 3.37 | 2.78 | 2.96 | 0.77 | 4.20 | 0.35 |
| | 8 | 4 | 400 | 3.01 | 0.53 | 2.62 | 3.84 | 3.77 | 3.17 | 4.00 | 0.92 | 2.41 | 0.86 |
| Mimi-S-24 | 8 | 1.1 | 100 | 3.43 | 1.19 | 2.22 | 3.60 | 3.68 | 3.17 | 4.27 | 0.90 | 3.72 | 0.70 |
| | 32 | 4.4 | 400 | 9.32 | 7.45 | 3.38 | 3.92 | 3.74 | 3.18 | 4.40 | 0.96 | 2.96 | 0.85 |
| DWavL-S-16 | 2 | 1 | 100 | -13.96 | −37.23 | 1.13 | 3.32 | 3.68 | 3.13 | 3.86 | 0.75 | 4.97 | 0.33 |
| | 6 | 3 | 300 | -12.69 | −35.43 | 1.19 | 3.32 | 3.72 | 3.13 | 4.05 | 0.75 | 4.34 | 0.35 |
| SQ-SMA-16 | 4 | 3 | 200 | 1.91 | −8.61 | 3.31 | 3.90 | **3.83** | **3.28** | 4.13 | 0.96 | **2.37** | **0.87** |
| WT-SMA-24 | 1 | .98 | 75 | 2.02 | −0.79 | 1.88 | 3.77 | 3.76 | 3.18 | **4.41** | 0.87 | 8.10 | 0.60 |
| WT-S-24 | 1 | .52 | 40 | 0.17 | −3.16 | 2.05 | 3.89 | 3.82 | 3.27 | 4.38 | 0.89 | 8.91 | 0.61 |

# Reconstruction Performance for both General Audio and Music

| | | | | Audio | | | | | Music | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tokenizer | #Q | kbps | Token rate | SDR ↑ | CI-SDR↑ | SI-SNR↑ | VISQOL ↑ | Sing MOS↑ | SDR ↑ | CI-SDR↑ | SI-SNR↑ | VISQOL ↑ | Sing MOS↑ |
| Ground truth | - | - | - | **252.75** | **84.90** | **57.96** | **4.73** | **2.70** | **254.24** | **87.26** | **60.26** | **4.73** | **2.79** |
| Enc-SMA-24 | 2 | 1.5 | 150 | −1.29 | −1.28 | −4.31 | 3.94 | 2.59 | 2.16 | 2.13 | 0.46 | 4.05 | 2.67 |
| | 8 | 6 | 600 | 4.28 | 4.10 | 2.33 | 4.25 | 2.60 | 7.32 | 7.17 | 5.87 | 4.38 | 2.66 |
| | 32 | 24 | 2400 | **7.72** | **7.33** | 5.64 | 4.36 | 2.60 | **11.04** | **10.75** | **9.19** | 4.50 | 2.66 |
| DAC-SMA-24 | 2 | 1.5 | 150 | −2.60 | −2.55 | −11.55 | 3.99 | 2.59 | 1.75 | 1.71 | −2.21 | 3.94 | **2.70** |
| | 8 | 6 | 600 | 1.35 | 1.22 | −10.28 | 4.35 | 2.61 | 4.82 | 4.67 | −1.25 | 4.30 | 2.68 |
| | 32 | 24 | 2400 | 2.45 | 2.22 | −9.91 | **4.59** | 2.60 | 5.56 | 5.37 | −1.16 | **4.56** | 2.66 |
| SQ-SMA-16 | 4 | 3 | 200 | −2.33 | −2.33 | −10.50 | 4.32 | **2.62** | 3.44 | 3.39 | −0.38 | 4.34 | 2.68 |
| WT-SMA-24 | 1 | .98 | 75 | −4.55 | −4.45 | **9.78** | 3.96 | 2.56 | −14.30 | −14.28 | −23.09 | 3.64 | 2.60 |
| WT-S-24 | 1 | .52 | 40 | −11.00 | −10.85 | −20.91 | 3.85 | 2.53 | −19.91 | −19.89 | −45.55 | 3.33 | 2.42 |

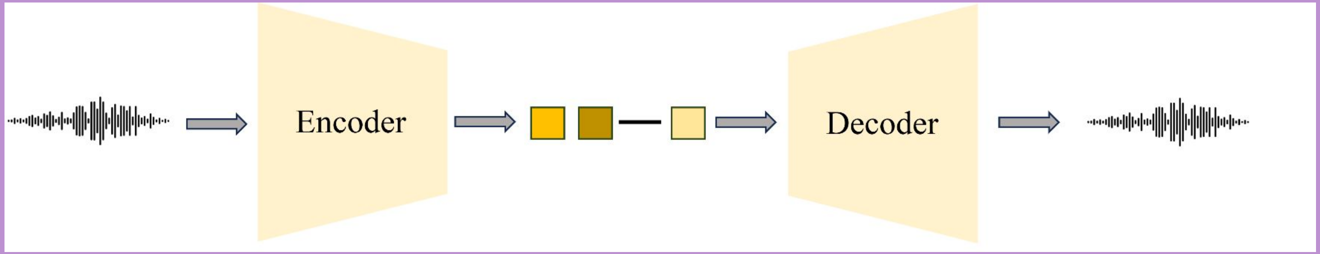
# Main Takeaways

- Overall, these results underscore the importance of evaluating audio tokenizers beyond traditional waveform fidelity measures.

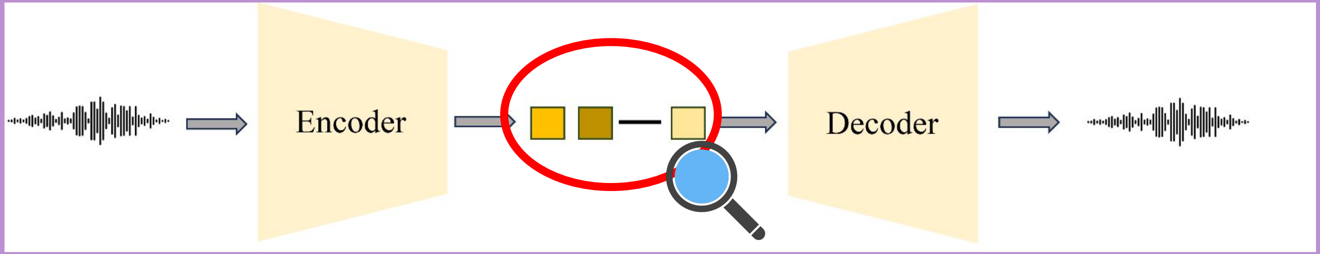
# Main Takeaways

- Overall, these results underscore the importance of evaluating audio tokenizers beyond traditional waveform fidelity measures.

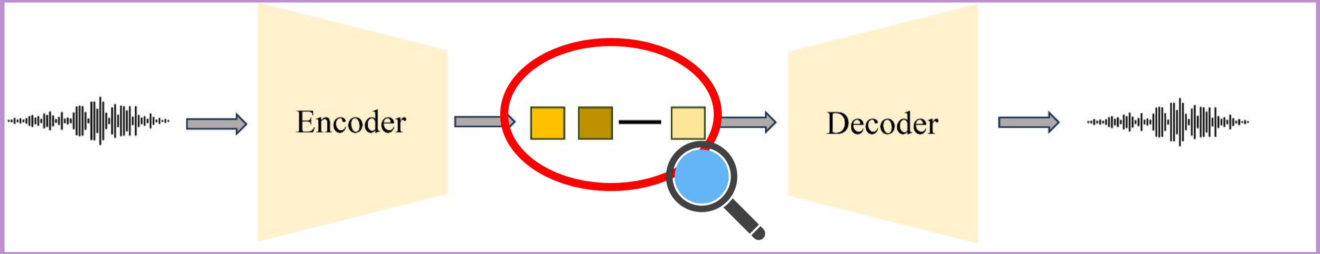
- Models optimized for perceptual or downstream tasks may exhibit low signal reconstruction performance, yet still produce subjectively high-quality audio reconstructions.




# Downstream Evaluation




# Downstream Evaluation




# Downstream Evaluation

DASB

# Datasets, metrics, and downstream models for the DASB

| Task | Dataset | Architecture | Metric(s) |
|---|---|---|---|
| | *Speech (Discriminative)* | | |
| ASR (En) | LibriSpeech (Korvas et al., 2014) | Branchformer | WER |
| ASR (Low-resource) | CommonVoice 17.0 (Ardila et al., 2020) | BiLSTM | WER |
| Speaker ID / Verification | VoxCeleb1 (Nagrani et al., 2017) | ECAPA-TDNN | Accuracy / EER |
| Emotion Recognition | IEMOCAP (Busso et al., 2008) | ECAPA-TDNN | Accuracy |
| Keyword Spotting | Speech Commands (Warden, 2018) | ECAPA-TDNN | Accuracy |
| Intent Classification | SLURP (Bastianelli et al., 2020) | BiLSTM+Linear | Accuracy |
| | *Speech (Generative)* | | |
| Speech Enhancement | VoiceBank (Valentini-Botinhao et al., 2016) | Conformer | DNSMOS / dWER |
| Speech Separation | Libri2Mix (Cosentino et al., 2020) | Conformer | DNSMOS / dWER / SpkSim |
| | *Music* | | |
| Music Genre Classification | GTZAN (Tzanetakis & Cook, 2002) | ECAPA-TDNN | Accuracy |
| Music Source Separation | MUSDB (Rafii et al., 2017) | Conformer | SDR / SIR / SAR |
| | *General Audio* | | |
| Sound Event Classification | ESC-50 (Piczak, 2015) | ECAPA-TDNN | Accuracy |
| Audio Separation | FUSS (Wisdom et al., 2021) | Conformer | SDR |

# DASB Results for Discriminative Tasks (speech)

| Tokenizer | #Q | ASR-En WER↓ Clean | ASR-En WER↓ Other | ASR-LR WER↓ Welsh | ASR-LR WER↓ Basque | ER ACC↑ | IC ACC↑ | KS ACC↑ | SI ACC↑ | SV EER↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| Continuous | – | **4.07** | **6.81** | **41.77** | **14.32** | **63.10** | **86.10** | **99.00** | **99.70** | **2.10** |
| Enc-SMA-24 | 2 | 12.70±0.37 | 29.09±0.13 | 90.90±0.32 | 51.00±0.98 | 45.50±0.02 | 42.90±0.16 | 77.73±3.12 | 89.81±5.46 | 18.33±0.26 |
| | 8 | 8.43±0.13 | 21.77± 0.36 | 84.53±1.90 | 45.36±0.57 | 44.73±0.02 | 40.03±0.29 | 74.30±1.69 | 94.26±3.99 | 13.54±0.57 |
| | 32 | 9.95±1.17 | 23.24± 1.22 | 97.39±1.19 | 58.21±0.92 | 42.96±0.02 | 33.66±2.65 | 69.10±3.42 | 91.12±1.92 | 10.12±6.66 |
| DAC-SMA-24 | 2 | 14.84±0.25 | 33.88±0.20 | 95.21±0.84 | 68.93±0.42 | 45.20±0.01 | 29.83±0.19 | 67.27±1.56 | **97.88±0.79** | 21.80±1.00 |
| | 8 | 10.73± 0.10 | 25.39± 0.20 | 97.20±0.14 | 62.45±1.40 | 44.73±0.02 | 23.97±0.41 | 65.27±2.82 | 87.33±10.98 | 15.86±5.26 |
| | 32 | 13.13±0.16 | 28.47±0.19 | 98.96±0.18 | 73.57±1.56 | 43.20±0.02 | 44.60±39.19 | 68.67±2.91 | 87.69±4.99 | 17.12 ± 0.76 |
| ST-S-16 | 2 | 9.48±0.10 | 22.68±0.10 | 71.36±0.32 | 42.17±0.05 | 54.86±0.01 | 56.80±0.08 | 94.11±0.63 | 73.16±0.37 | 24.23±0.29 |
| | 8 | 9.06± 0.45 | 21.72±0.23 | 68.36±0.44 | 35.35±0.22 | 55.00±0.01 | 53.83±0.05 | 94.11±0.07 | 96.78±0.45 | 10.45±0.43 |
| Mimi-S-24 | 8 | 9.73±0.61 | 22.65±0.41 | 91.59±0.15 | 59.18±8.52 | 51.13±0.02 | 53.83±0.19 | 92.18±0.20 | 79.50±0.43 | 18.68±0.35 |
| | 32 | 10.84±0.56 | 24.10±0.36 | 96.89±0.07 | 58.15±6.90 | 46.76±0.01 | 50.73±0.50 | 91.31±0.19 | 63.93±13.64 | 23.91±4.60 |
| DWavL-S-16 | 2 | **4.78±0.25** | 10.58±0.17 | 58.98±0.15 | 22.02±0.17 | 61.53±0.02 | 76.33±0.17 | **96.82±0.92** | 76.57±0.33 | 22.41±0.19 |
| | 6 | 5.07±0.17 | **9.57±0.20** | **48.94±0.38** | **19.66±0.33** | **63.20±0.01** | **78.73±0.12** | 95.89±0.50 | 92.31±0.09 | 13.47±0.22 |
| SQ-SMA-16 | 4 | 91.57±0.49 | 92.90±0.41 | 94.80±0.88 | 94.24±1.24 | 41.30±0.06 | 58.13±0.26 | 92.74±0.42 | 97.38±0.03 | **9.69±0.25** |
| SQ-SMA-16* | 4 | 11.63±0.08 | 30.91±0.17 | – | – | – | – | – | – | – |
| WT-SMA-24 | 1 | 16.11±0.18 | 35.48±0.35 | 97.41±0.08 | 75.82±0.20 | 43.43±0.02 | 15.25±0.15 | 59.13±2.10 | 85.90±2.48 | 19.38±0.36 |

# DASB Results for Generative Tasks (speech).

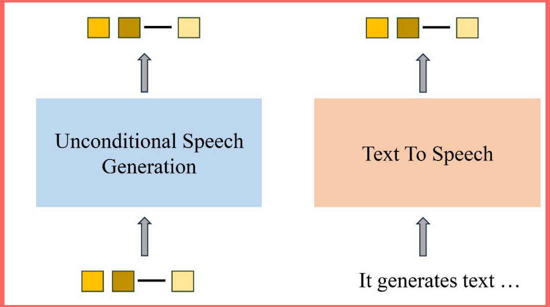
| Models\Tasks | #Q | SE | | | SS - Speech | | | |
|---|---|---|---|---|---|---|---|---|
| | | DNSMOS ↑ | dWER ↓ | Spk Sim↑ | DNSMOS Rec↑ | DNSMOS Sep↑ | dWER ↓ | Spk Sim↑ |
| Continuous | – | 3.49 | **4.92** | **0.93** | – | 3.68 | **9.97** | **0.94** |
| Enc-SMA-24 | 2 | 3.15±0.01 | 34.95±0.64 | 0.86±0.00 | 3.19 | 3.13±0.00 | 80.33±1.77 | 0.88±0.00 |
| | 8 | 3.08±0.01 | 22.70±1.84 | 0.88±0.00 | 3.54 | 3.08±0.00 | 53.37±0.65 | 0.90±0.00 |
| | 32 | 2.78±0.01 | 65.70±6.09 | 0.80±0.01 | 3.72 | 2.97±0.01 | 92.42±0.97 | 0.85±0.00 |
| DAC-SMA-24 | 2 | 3.26±0.01 | 54.85±1.82 | 0.86±0.00 | 3.16 | 3.01±0.00 | 101.19±1.99 | 0.85±0.00 |
| | 8 | 3.51±0.01 | 29.44±3.93 | **0.90±0.01** | 3.67 | 3.30±0.00 | 52.77±2.48 | **0.93±0.00** |
| | 32 | 2.93±0.01 | 30.66±0.97 | 0.88±0.00 | 3.76 | 2.67±0.01 | 92.07±0.05 | 0.88±0.01 |
| ST-S-16 | 2 | 3.19±0.02 | 29.98±0.58 | 0.86±0.00 | 3.20 | 3.13±0.00 | 84.94±0.63 | 0.87±0.00 |
| | 8 | 3.49±0.01 | 21.65±0.57 | 0.87±0.00 | 3.72 | 3.43±0.01 | 60.90±0.77 | 0.91±0.00 |
| Mimi-S-24 | 8 | 3.25±0.01 | 67.56±2.21 | 0.85±0.00 | 3.65 | 3.29±0.00 | 109.30±3.30 | 0.87±0.00 |
| | 32 | 3.18±0.01 | 102.61±2.40 | 0.82±0.00 | 3.72 | 3.00±0.00 | 137.00±2.16 | 0.82±0.00 |
| DWavL-S-16 | 2 | 3.56±0.01 | 25.88±2.15 | 0.88±0.00 | 3.57 | 3.56±0.00 | 49.57±0.64 | 0.85±0.00 |
| | 6 | **3.57±0.01** | **9.43±0.33** | 0.89±0.00 | 3.75 | **3.75±0.01** | **30.39±0.45** | 0.91±0.00 |
| SQ-SMA-16 | 4 | 3.28±0.01 | 122.33±8.74 | 0.83±0.00 | **3.77** | 3.19±0.00 | 136.00±3.58 | 0.83±0.00 |
| WT-SMA-24 | 1 | 3.33±0.01 | 67.53±10.65 | 0.85±0.00 | 3.57 | 3.42±0.00 | 118.33±4.50 | 0.86±0.00 |
| Mixture | – | – | – | – | – | 3.43 | – | – |

# DASB Results  Music and General Audio Tasks.

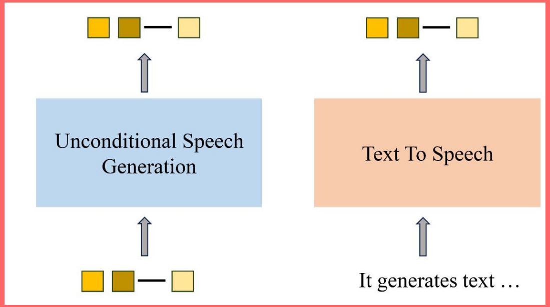
| Tokenizer | #Q | SS - Audio SI-SDRi↑ Rec | SS - Audio SI-SDRi↑ Sep | SS - Music SI-SDRi↑ Rec | SS - Music SI-SDRi↑ Sep | SS - Music SAR↑ | SS - Music SIR↑ | SEC ACC↑ | MGC ACC↑ |
|---|---|---|---|---|---|---|---|---|---|
| Continuous | – | – | 15.07 | – | 13.29 | 9.56 | 11.99 | 92.91 | 87.00 |
| Enc-SMA-24 | 2 | 0.76 | 7.03±0.49 | 3.36 | 1.49±2.04 | -2.80±1.68 | **5.96±1.52** | 34.83±0.47 | **70.33±1.70** |
| | 8 | 3.87 | **9.53±0.33** | 7.99 | **1.98±0.36** | **-1.95±0.33** | 5.26±0.22 | **37.00±0.73** | 54.67±3.86 |
| | 32 | **5.76** | -1.73±0.09 | **11.10** | -11.72±0.35 | -15.00±0.02 | -0.42±0.01 | 35.43±1.45 | 39.67±1.25 |
| DAC-SMA-24 | 2 | 0.12 | 3.84±0.48 | 2.37 | 1.01±0.17 | -3.59±0.09 | **5.92±0.28** | 31.03±1.84 | 50.00±0.82 |
| | 8 | 3.33 | 5.62±0.21 | 6.66 | -11.77±0.1 | -10.62±2.35 | -5.52±3.68 | 28.60±0.79 | 47.67±3.09 |
| | 32 | 4.73 | -4.92±0.32 | 8.54 | -11.32±0.12 | -12.70±0.17 | -2.05±0.41 | 36.67±0.92 | 50.00±0.82 |
| SQ-SMA-16 | 4 | 3.62 | 6.54±0.22 | 5.53 | -3.62±0.87 | -5.84±0.86 | 1.42±0.32 | 31.37±1.37 | 42.67±0.47 |
| WT-SMA-24 | 1 | -24.05 | -16.72±0.08 | -2.66 | -4.52±0.04 | -8.32±0.07 | 2.65±0.11 | 34.50±0.82 | 48.00±1.41 |
| Mixture | – | – | -16.5 | – | -7.71 | 50.01 | -inf | – | – |

# Main Takeaways

| Tokenizer Type | Reconstruction | Downstream | Low-Resource Robustness | Model Scalability | Convergence Speed | efficiency |
|---|---|---|---|---|---|---|
| Acoustic | High | Low | ✗ | Requires large models | Slow | High |
| Semantic | Low | High | ✓ | Can work with smaller models | Fast | Low |
| Hybrid | Moderate | Moderate | ✗ | Depends on data | Moderate | Moderate |




# Acoustic Language Models Evaluation




# Acoustic Language Models Evaluation

[Zero Resource]

[SALMon]

# Speech Language Modeling

| Tokenizer | #Q | Spoken Content | | | | Acoustic Consistency | | | Sem.-Ac. Align. |
|---|---|---|---|---|---|---|---|---|---|
| | | sBLIMP↑ | sWUGGY↑ | sSC↑ | tSC↑ | Gender↑ | Sent.↑ | Spk↑ | Sentiment↑ |
| HuBERT 25Hz | 1 | **60.89** | **70.51** | **53.23** | **71.46** | 69.50 | 62.50 | 69.00 | **53.00** |
| Enc-SMA-24 | 8 | 51.14 | 51.29 | 50.18 | 48.20 | 70.50 | 56.50 | 65.00 | 50.00 |
| DAC-SMA-16 | 8 | 51.51 | 50.73 | 48.95 | 51.52 | 81.00 | 60.00 | 77.00 | 50.00 |
| ST-S-16 | 8 | 51.08 | 56.89 | 48.42 | 55.74 | 66.50 | 58.00 | 65.00 | 49.50 |
| ST-S-16* | 8 | 52.75 | 63.46 | 47.56 | 60.60 | 67.00 | 59.50 | 65.50 | 50.00 |
| Mimi-S-24 | 8 | 52.25 | 62.21 | 51.52 | 54.30 | 77.50 | 71.50 | 78.00 | 52.00 |
| Mimi-S-24* | 8 | 60.17 | 67.57 | 51.68 | 68.51 | 76.50 | 77.00 | 76.00 | 52.00 |
| DWavL-S-16 | 6 | 53.96 | 69.10 | 51.41 | 62.42 | **92.00** | 70.00 | **86.50** | 49.00 |
| SQ-SMA-16 | 4 | 51.58 | 51.41 | 51.79 | 55.10 | 83.00 | 64.00 | 84.50 | 50.50 |
| WT-SMA-24 | 1 | 51.22 | 54.60 | 52.00 | 52.75 | 81.50 | **78.50** | 69.00 | 50.50 |

# Main Takeaways

- Semantic and acoustic performance in SLMs varies significantly across tokenizer types.

# Main Takeaways

- Semantic and acoustic performance in SLMs varies significantly across tokenizer types.

- Semantically distilled tokenizers, particularly those with semantic stream overweighting, showed promising results close to HuBERT.

# Main Takeaways

- Semantic and acoustic performance in SLMs varies significantly across tokenizer types.

- Semantically distilled tokenizers, particularly those with semantic stream overweighting, showed promising results close to HuBERT.

- Overall, our findings suggest that, for now, there is no single tokenizer that excels across all spoken and acoustic tasks.

# Text-to-Speech (VALL-E).

| Tokenizer | #Q | UTMOS↑ | dWER↓ | SpkSim↑ |
|---|---|---|---|---|
| Enc-SMA-24 | 8 | 2.31 | 4.77 | **0.91** |
| Enc-S-24 | 8 | **3.77** | 5.74 | **0.91** |
| DAC-SMA-24 | 8 | 2.47 | 11.71 | 0.88 |
| ST-S-16 | 8 | 2.91 | 5.35 | **0.91** |
| Mimi-S-24 | 8 | 2.60 | 7.93 | **0.91** |
| DWavL-S-16 | 6 | 3.42 | **4.32** | 0.90 |
| WT-SMA-24 | 1 | 2.85 | 4.67 | 0.88 |

# Main Takeaways

- Overall, achieving strong TTS performance with discrete tokenizers remains challenging, especially under constrained training conditions

# Main Takeaways

- Overall, achieving strong TTS performance with discrete tokenizers remains challenging, especially under constrained training conditions

- Training with semantic tokenizers leads to more robust and effective TTS performance compared to acoustic or semantically distilled tokenizers.

# Main Takeaways

- Overall, achieving strong TTS performance with discrete tokenizers remains challenging, especially under constrained training conditions

- Training with semantic tokenizers leads to more robust and effective TTS performance compared to acoustic or semantically distilled tokenizers.

- When scaling data and model, acoustic tokenizers, such as EnCodec, can be competitive with or even outperform semantic ones.

# Audio Generation

| Tokenizer | #Q | Text Cond. Generation | | | Uncond. Generation | | | Reconstruction | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | FAD↓ | KLD↓ | CLAP↑ | FAD↓ | KLD↓ | CLAP↑ | FAD↓ | KLD↓ | CLAP↑ |
| Enc-SMA-24 | 8 | 3.771 | 1.555 | .279 | 5.996 | **1.897** | .202 | 3.806 | 0.456 | .281 |
| Enc-M-32 | 4 | 10.110 | 1.788 | .295 | 13.400 | 2.840 | .175 | 12.611 | 1.387 | .251 |
| Enc-A-16 | 4 | **1.955** | 1.576 | **.300** | **3.548** | 2.064 | .205 | **1.816** | 0.419 | .273 |
| DAC-SMA-44 | 9 | 6.929 | 1.959 | .267 | 6.732 | 2.041 | **.212** | 2.206 | **0.242** | **.299** |
| DAC-SMA-24 | 9 | 7.708 | 1.966 | .253 | 8.196 | 2.183 | .199 | 4.124 | 0.446 | .281 |
| SQ-SMA-16 | 4 | 7.733 | 3.078 | .151 | 5.977 | 2.301 | .175 | 3.460 | 0.460 | .268 |
| WT-SMA-24 | 1 | 2.594 | **1.463** | .291 | 4.441 | 2.224 | .193 | 5.018 | 0.892 | .253 |

# Main Takeaways

- Our findings highlight the critical role of domain-specific training for audio tokenizers.

# Main Takeaways

- Our findings highlight the critical role of domain-specific training for audio tokenizers.

- Our results also show that the best reconstruction performance does not correlate with the best modeling performance.

# Main Takeaways

- Our findings highlight the critical role of domain-specific training for audio tokenizers.
- Our results also show that the best reconstruction performance does not correlate with the best modeling performance.
- We also emphasize the need for more robust evaluation metrics.

# Music Generation

| Tokenizer | #Q | Text Cond. Generation | | | Uncond. Generation | | | Reconstruction | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | FAD↓ | KLD↓ | CLAP↑ | FAD↓ | KLD↓ | CLAP↑ | FAD↓ | KLD↓ | CLAP↑ |
| *MusicCaps* | | | | | | | | | | |
| Enc-SMA-24 | 8 | 11.173 | 2.246 | .108 | 4.632 | 0.904 | .275 | 2.209 | 0.259 | **.358** |
| Enc-M-32 | 4 | **4.264** | **2.006** | **.150** | **2.715** | 0.890 | **.282** | 1.995 | 0.356 | .339 |
| DAC-SMA-44 | 9 | 8.398 | 2.214 | .119 | 3.724 | **0.784** | **.282** | **0.927** | **0.182** | .340 |
| DAC-SMA-24 | 9 | 9.403 | 2.127 | .093 | 4.001 | 0.820 | .277 | 1.335 | 0.209 | **.358** |
| SQ-SMA-16 | 4 | 14.211 | 2.810 | .064 | 5.163 | 0.979 | .270 | 2.078 | 0.258 | .338 |
| WT-SMA-24 | 1 | 17.050 | 2.792 | .056 | 5.550 | 1.105 | .251 | 1.984 | 0.414 | .336 |
| *FMA* | | | | | | | | | | |
| Enc-SMA-24 | 8 | 15.380 | 2.161 | .059 | 14.478 | 1.827 | .065 | 1.013 | 0.287 | .141 |
| Enc-M-32 | 4 | 8.871 | **1.299** | **.078** | 8.357 | **1.006** | **.079** | 0.784 | 0.344 | .153 |
| DAC-SMA-44 | 9 | **8.115** | 1.543 | .062 | 6.398 | 1.100 | .075 | **0.494** | **0.196** | **.158** |
| DAC-SMA-24 | 9 | 8.789 | 1.746 | .039 | 7.002 | 1.405 | .043 | 0.708 | 0.222 | .125 |
| SQ-SMA-16 | 4 | 9.426 | 2.412 | .048 | **4.690** | 1.592 | .070 | 0.956 | 0.327 | .133 |
| WT-SMA-24 | 1 | 16.511 | 1.881 | .030 | 6.890 | 1.414 | .047 | 0.631 | 0.368 | .129 |

# Main Takeaways

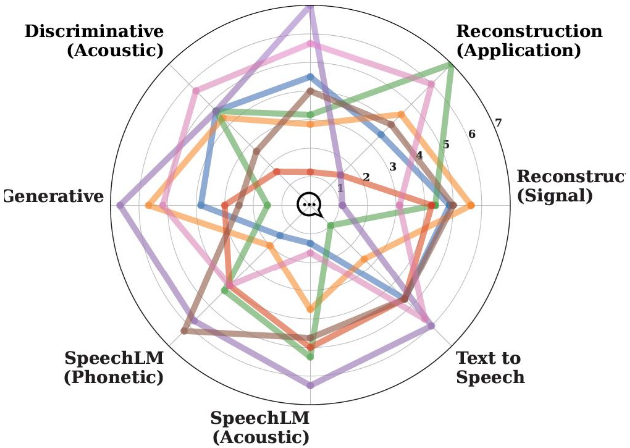
- Same as audio, domain-specific training is crucial for music tokenizers.

# Main Takeaways

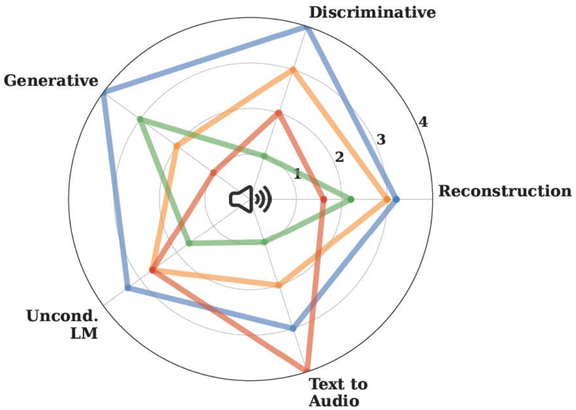
- Same as audio, domain-specific training is crucial for music tokenizers.

- Tokenizers with higher sample rates and multi-codebook, associated with higher bitrates, tend to perform better.

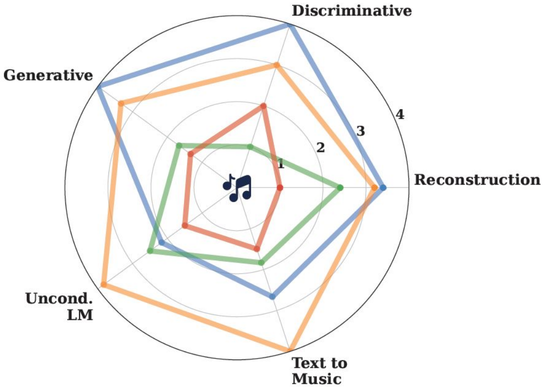
# General Trend




(a) Speech Ranking

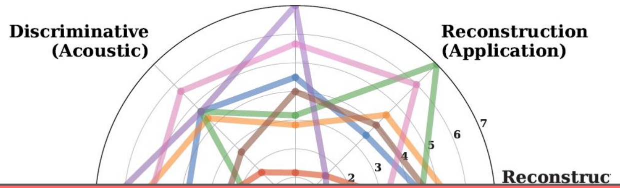



(b) Audio Ranking

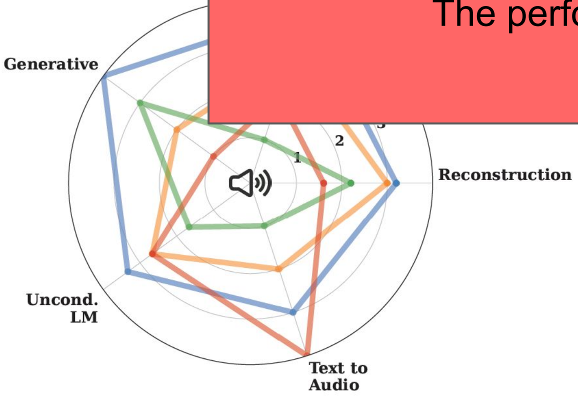



(c) Music Ranking

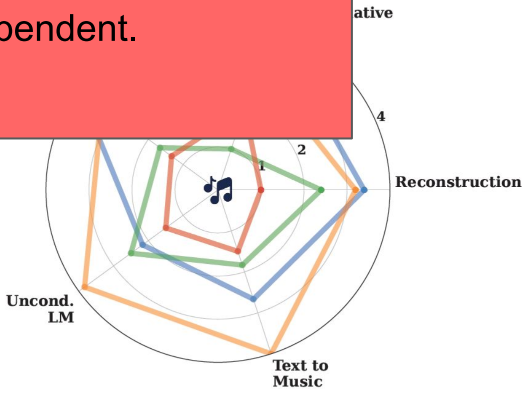
# General Trend




No tokenizer consistently outperforms others on all axes.
The performance is strongly task- and domain-dependent.




(b) Audio Ranking




(c) Music Ranking

# Study 3:
# Ablation Studies

ESPnet-C
odec

# Summary of Models Used in the Ablation Study Across 16kHz and 44.1kHz Setups.

| Model | Base | Quantization | Distillation | Data Domains | | |
|---|---|---|---|---|---|---|
| | | | | Speech | Audio | Music |
| RVQ-S | DAC | RVQ | - | ✓ | | |
| RVQ-S+ | | | ✓ | ✓ | | |
| RVQ-A | | | - | | ✓ | |
| RVQ-M | | | - | | | ✓ |
| RVQ-3 | | | - | ✓ | ✓ | ✓ |
| SVQ-S | DAC | SVQ | - | ✓ | | |
| SVQ-S+ | | | ✓ | ✓ | | |
| SVQ-A | | | - | | ✓ | |
| SVQ-M | | | - | | | ✓ |
| SVQ-3 | | | - | ✓ | ✓ | ✓ |
| FSQ-S | DAC | FSQ | - | ✓ | | |
| FSQ-A | | | - | | ✓ | |
| FSQ-M | | | - | | | ✓ |
| FSQ-3 | | | - | ✓ | ✓ | ✓ |
| K-means-S | Unit-HifiGAN | K-means | - | ✓ | | |

# Ablation Experiments on Reconstruction Performance (speech).

| | SDR↑ | | SI-SNR↑ | | PESQ↑ | | UTMOS↑ | | DNSMOS P835↑ | | WER↓ | | Spk Sim↑ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model\SR(kHz) | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 |
| Ground truth | - | - | - | - | - | - | **4.09** | **4.09** | 3.18 | **3.18** | 2.83 | 2.83 | - | - |
| RVQ-S | **4.08** | 8.24 | 1.15 | **6.38** | **2.59** | 3.24 | **3.35** | 3.62 | 3.16 | **3.16** | **2.04** | **2.63** | 0.67 | **0.90** |
| RVQ-S+ | 1.63 | **8.40** | -1.77 | 5.67 | 2.22 | **3.30** | 3.12 | **3.65** | 3.12 | 3.14 | 2.12 | 3.17 | **0.69** | 0.89 |
| RVQ-A | 0.59 | 6.92 | -3.36 | 4.27 | 2.02 | 2.86 | 1.81 | 3.15 | 2.58 | 3.06 | 2.47 | 2.92 | 0.51 | 0.73 |
| RVQ-M | 2.80 | 6.74 | -0.68 | 4.61 | 2.00 | 2.41 | 1.64 | 2.33 | 2.64 | 2.68 | 3.20 | 2.85 | 0.44 | 0.56 |
| RVQ-3 | 2.46 | 7.50 | -1.12 | 4.63 | 2.43 | 3.06 | 2.71 | 3.33 | 2.96 | 3.08 | 2.22 | 2.66 | 0.61 | 0.87 |
| SVQ-S | -4.90 | 0.92 | -13.59 | -2.04 | 1.43 | 1.69 | 2.19 | 2.61 | 3.09 | 3.10 | 13.16 | 8.28 | 0.35 | 0.53 |
| SVQ-S+ | -4.45 | -0.64 | -11.74 | -3.64 | 1.42 | 1.63 | 2.14 | 2.40 | 3.00 | 3.07 | 13.71 | 7.89 | 0.36 | 0.52 |
| SVQ-A | -11.80 | -5.04 | -33.46 | -9.56 | 1.19 | 1.18 | 1.25 | 1.26 | 1.86 | 1.97 | 31.40 | 34.68 | 0.19 | 0.14 |
| SVQ-M | -5.19 | -4.70 | -11.17 | -7.89 | 1.20 | 1.14 | 1.29 | 1.24 | 2.19 | 1.56 | 14.88 | 33.87 | 0.15 | 0.11 |
| SVQ-3 | -5.90 | -3.09 | -15.13 | -7.12 | 1.29 | 1.79 | 1.44 | 2.57 | 3.10 | 3.01 | 22.24 | 6.71 | 0.26 | 0.49 |
| FSQ-S | 3.89 | 4.58 | **1.75** | 2.47 | 2.08 | 2.10 | 3.29 | 3.06 | **3.21** | 3.12 | 3.57 | 4.02 | 0.48 | 0.66 |
| FSQ-A | 1.14 | 1.00 | -1.89 | -1.58 | 1.94 | 1.74 | 2.84 | 2.41 | 3.15 | 2.97 | 5.12 | 7.59 | 0.43 | 0.39 |
| FSQ-M | -1.08 | -2.09 | -4.39 | -4.61 | 1.39 | 1.22 | 1.57 | 1.26 | 2.66 | 2.04 | 24.46 | 20.26 | 0.17 | 0.16 |
| FSQ-3 | 2.41 | 1.29 | 0.01 | -1.28 | 1.97 | 1.79 | 3.06 | 2.57 | 3.20 | 3.01 | 4.35 | 6.71 | 0.44 | 0.49 |
| K-means-S | -18.21 | - | -42.98 | - | 1.05 | - | 2.28 | - | 2.46 | - | 6.78 | - | 0.13 | - |

# Ablation Experiments on Reconstruction performance (audio and music).

| | Audio | | | | | | | | | | Music | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SDR↑ | | CI-SDR↑ | | SI-SNR↑ | | VISQOL↑ | | SingMOS↑ | | SDR↑ | | CI-SDR↑ | | SI-SNR↑ | | VISQOL↑ | | SingMOS↑ | |
| Model\SR(kHz) | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 | 16 | 44.1 |
| RVQ-S | -4.05 | 2.85 | -4.02 | 2.51 | -10.07 | -0.02 | 4.18 | 3.81 | 2.59 | **2.66** | 0.45 | 6.80 | 0.44 | 6.75 | -2.29 | 4.83 | **4.21** | 4.17 | 2.67 | 2.60 |
| RVQ-S+ | -8.90 | 2.52 | -8.83 | 2.19 | -17.57 | -1.14 | 4.13 | 3.79 | 2.57 | 2.63 | -6.73 | 6.60 | -6.70 | 6.55 | -11.10 | 4.46 | 4.07 | **4.24** | 2.64 | 2.70 |
| RVQ-A | -0.59 | 3.65 | -0.61 | 3.21 | -5.56 | 0.48 | **4.22** | 3.78 | 2.63 | 2.65 | 5.78 | 8.24 | 5.70 | 8.17 | 2.87 | 5.97 | **4.21** | 4.12 | 2.71 | 2.72 |
| RVQ-M | **0.65** | 3.76 | **0.60** | 3.35 | **-4.07** | **1.06** | 4.13 | 3.62 | 2.63 | 2.63 | **6.34** | 8.33 | **6.25** | 8.26 | **3.82** | **6.48** | 4.13 | 3.99 | 2.70 | 2.68 |
| RVQ-3 | 0.14 | **3.99** | 0.11 | **3.50** | -4.47 | 0.70 | 4.17 | **3.83** | 2.61 | 2.65 | 5.75 | **8.54** | 5.68 | **8.46** | 3.32 | 6.34 | 4.12 | 4.07 | 2.68 | 2.72 |
| SVQ-S | -14.80 | -8.43 | -14.72 | -8.07 | -31.86 | -14.73 | 3.83 | 3.28 | 2.55 | 2.59 | -14.25 | -4.78 | -14.22 | -4.76 | -27.16 | -7.77 | 3.68 | 3.64 | 2.59 | 2.70 |
| SVQ-S+ | -14.74 | -9.39 | -14.66 | -9.00 | -30.69 | -16.37 | 3.84 | 3.23 | 2.54 | 2.59 | -15.10 | -5.24 | -15.06 | -5.23 | -27.69 | -8.80 | 3.69 | 3.60 | 2.58 | 2.68 |
| SVQ-A | -13.08 | -6.80 | -13.00 | -6.46 | -30.70 | -12.80 | 3.86 | 3.32 | 2.56 | 2.59 | -6.54 | -0.49 | -6.52 | -0.40 | -14.76 | -3.21 | 3.68 | 3.52 | 2.65 | 2.69 |
| SVQ-M | -6.53 | -5.50 | -6.47 | -5.19 | -13.43 | -10.62 | 3.94 | 3.33 | 2.59 | 2.58 | -0.35 | 0.52 | -0.35 | 0.52 | -2.94 | -1.80 | 3.87 | 3.33 | 2.70 | 2.63 |
| SVQ-3 | -9.28 | -7.47 | -9.21 | -7.12 | -19.50 | -14.32 | 3.88 | 3.34 | 2.52 | 2.62 | -2.75 | -1.35 | -2.73 | -1.35 | -6.75 | -4.63 | 3.72 | 3.54 | 2.63 | **2.73** |
| FSQ-S | -7.32 | -4.26 | -7.26 | -4.04 | -14.22 | -8.12 | 4.05 | 3.32 | 2.60 | 2.63 | -5.04 | -0.37 | -5.02 | -0.37 | -8.42 | -2.84 | 3.80 | 3.65 | 2.69 | 2.71 |
| FSQ-A | -2.79 | -2.00 | -2.76 | -1.00 | -7.05 | -5.37 | 4.09 | 3.53 | 2.63 | 2.65 | 0.90 | 2.62 | 0.80 | 2.60 | -1.14 | 0.49 | 4.00 | 3.92 | 2.70 | **2.73** |
| FSQ-M | -3.37 | -3.25 | -3.33 | -3.05 | -8.23 | -6.85 | 4.02 | 3.41 | 2.62 | 2.60 | 1.54 | 2.70 | 1.52 | 2.77 | -0.67 | 0.66 | 3.99 | 3.54 | 2.71 | 2.66 |
| FSQ-3 | -3.22 | -2.36 | -3.19 | -2.24 | -7.86 | -6.01 | 4.06 | 3.53 | 2.61 | 2.62 | 0.89 | 2.75 | 0.88 | 2.73 | -1.38 | 0.19 | 3.96 | 3.74 | 2.69 | 2.68 |
| K-means-S | -21.01 | - | -20.89 | - | -47.37 | - | 3.14 | - | **2.78** | - | -19.53 | - | -19.49 | - | -46.03 | - | 2.82 | - | **2.87** | - |

# Main Takeaways

- **Data Domains.** Our experiments confirm that reconstruction quality consistently peaks when models are evaluated on domains matching their training data.

# Main Takeaways

- **Data Domains.** Our experiments confirm that reconstruction quality consistently peaks when models are evaluated on domains matching their training data.

- **Sampling Rate.** we recommend that future research on discrete audio representation should consider sampling rate as a critical design parameter, with careful optimization based on the selected quantization approach and target application domain.

# Main Takeaways

- **Distillation Effect.** Distillation from pretrained speech representations can enhance model performance on certain metrics for signal reconstruction. But, s a potential trade-off between achieving high performance in specialized tasks and maintaining broader generalization capabilities.

# Main Takeaways

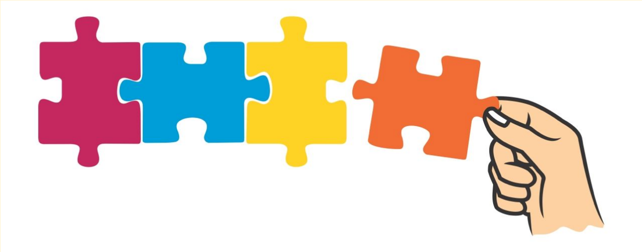
- **Distillation Effect.** Distillation from pretrained speech representations can enhance model performance on certain metrics for signal reconstruction. But, s a potential trade-off between achieving high performance in specialized tasks and maintaining broader generalization capabilities.

- **Quantization Methods.** Our experiments demonstrate that different quantization methods significantly impact codec performance.  The RVQ modeling consistently outperforms other quantization approaches across most reconstruction metrics.

# Listen to Some Examples

| Task | Dataset | ~Hours | Data Link |
|---|---|---|---|
| *Reconstruction* | | | |
| Speech | LibriSpeech (Korvas et al., 2014) | 6 | Link |
| Music | MUSDB (Rafii et al., 2017) | 10 | Link |
| Audio | Audioset (Gemmeke et al., 2017) | 56 | Link |
| *Downstream* | | | |
| ASR (En) | LibriSpeech (Korvas et al., 2014) | 1,000 | Link |
| ASR (Welsh) | CommonVoice 17.0 (Ardila et al., 2020) | 8 | Link |
| ASR (Basque) | CommonVoice 17.0 (Ardila et al., 2020) | 116 | Link |
| Speaker ID / Verification | VoxCeleb1 (Nagrani et al., 2017) | 350 | Link |
| Emotion Recognition | IEMOCAP (Busso et al., 2008) | 7 | Link |
| Keyword Spotting | Speech Commands (Warden, 2018) | 18 | Link |
| Intent Classification | SLURP (Bastianelli et al., 2020) | 10 | Link |
| Speech Enhancement | VoiceBank (Valentini-Botinhao et al., 2016) | 10 | Link |
| Speech Separation | Libri2Mix (Cosentino et al., 2020) | 400 | Link |
| Music Genre Classification | GTZAN (Tzanetakis & Cook, 2002) | 8 | Link |
| Music Source Separation | MUSDB (Rafii et al., 2017) | 10 | Link |
| Sound Event Classification | ESC-50 (Piczak, 2015) | 2 | Link |
| Audio Separation | FUSS (Wisdom et al., 2021) | 23 | Link |
| *Acoustic LM* | | | |
| Speech Language Modeling | LibriHeavy (Kang et al., 2024) | 56,000 | Link |
| Text-to-Speech | LibriTTS (Zen et al. (2019)) | 960 | Link |
| Audio Generation | Data Mix (see Sec. 3.3.3 for details) | 4050 | - |
| Music Generation | FMA (Defferrard et al., 2017) | 702 | Link |
| *Ablation* | | | |
| Speech | Data Mix (see Sec. 4 for details) | 1,000 | - |
| Music | Data Mix (see Sec. 4 for details) | 1,000 | - |
| Audio | Data Mix (see Sec. 4 for details) | 1,000 | - |

# Conclusion and Future Directions



# Scaling Limitations and Generalizability

# Scaling Limitations and Generalizability

## Correlation between Reconstruction and Downstream Performance

Scaling Limitations and Generalizability

Correlation between Reconstruction and Downstream Performance

Fair and Consistent Evaluation

Scaling Limitations and Generalizability

Correlation between Reconstruction and Downstream Performance

Fair and Consistent Evaluation

Benchmark vs. Reported Performance Gap

Scaling Limitations and Generalizability

Correlation between Reconstruction and Downstream Performance

Fair and Consistent Evaluation

Benchmark vs. Reported Performance Gap

Semantic Distillation Beyond Speech

Scaling Limitations and Generalizability

Correlation between Reconstruction and Downstream Performance

Fair and Consistent Evaluation

Benchmark vs. Reported Performance Gap

Semantic Distillation Beyond Speech

Discrete vs. Continuous Representations

- Scaling Limitations and Generalizability
- Correlation between Reconstruction and Downstream Performance
- Fair and Consistent Evaluation
- Benchmark vs. Reported Performance Gap
- Semantic Distillation Beyond Speech
- Discrete vs. Continuous Representations
- Toward Unified Tokenizers

Scaling Limitations and Generalizability

Correlation between Reconstruction and Downstream Performance

Fair and Consistent Evaluation

Benchmark vs. Reported Performance Gap

Semantic Distillation Beyond Speech

Discrete vs. Continuous Representations

Toward Unified Tokenizers

Trustworthiness


THANK YOU