Audio Encoder Capability Challenge 2025

UNIVERSITY OF SURREY

DataoceanAI

Workshop Agenda

Start Time	End Time	Duration	Topics	Presenters
10:15 AM	10:25 AM	0:10	Welcome and Introduction	Wenwu Wang
10:25 AM	10:35 AM	0:10	Baseline Sharing	Yadong Niu
10:35 AM	10:40 AM	0:05	Award winners with certificates	Wenwu Wang
10:40 AM	10:50 AM	0:10	Share thoughts on neural audio codec	Wenwu Wang
10:50 AM	10:55 AM	0.05	Q&A by Wenwu Wang	Wenwu Wang
10:55 AM	11:00 AM	0:05	Share thoughts on data related topics	Helen Wang
11:00 AM	11:10 AM	0:10	Presentation 1 by Team Audiocodec	Team Audiocodec
11:10 AM	11:15 AM	0:05	Q&A by Team Audiocodec	
11:15 AM	11:25 AM	0:10	Presentation 2 by Team SAMoVA	Team SAMoVA
11:25 AM	11:30 AM	0:05	Q&A by Team SAMoVA	

Challenge Introduction

Motivation

"Strongly inspired by the HEAR benchmark, this challenge introduces several key enhancements: a diverse task set, a focus on real-world applications, a combination of parameterized and parameter-free evaluation, and a new open-sourced, efficient evaluation system."

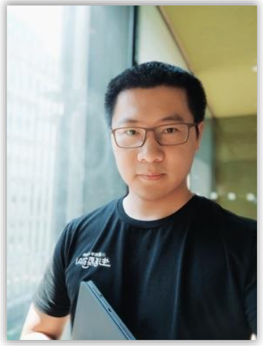
UNIVERSITY OF SURREY

DataoceanAI

Organization committee

Wenwu Wang

UNIVERSITY OF SURREY



Junbo Zhang

Yadong Niu

Helen Wang

DataoceanAI

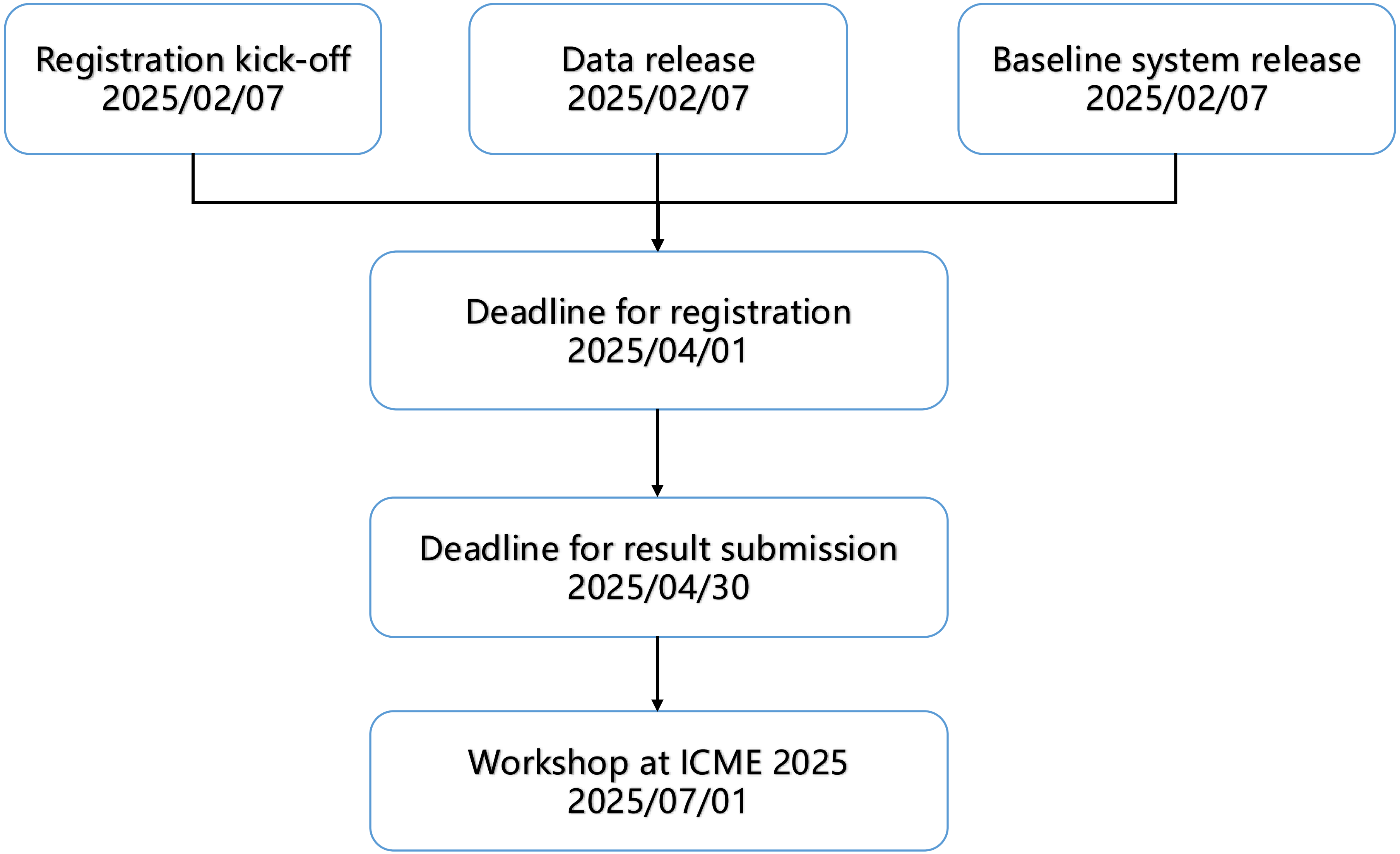
Guanbo Wang

DataoceanAI

Chris Wu

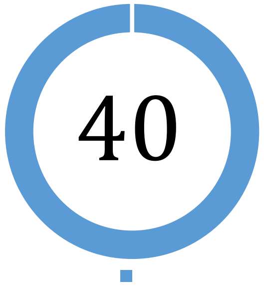
DataoceanAI

Timeline

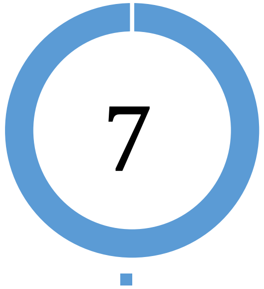

Registration kick-off
2025/02/07
Data release
2025/02/07
Baseline system release
2025/02/07
Deadline for registration
2025/04/01
Deadline for result submission
2025/04/30
Workshop at ICME 2025
2025/07/01

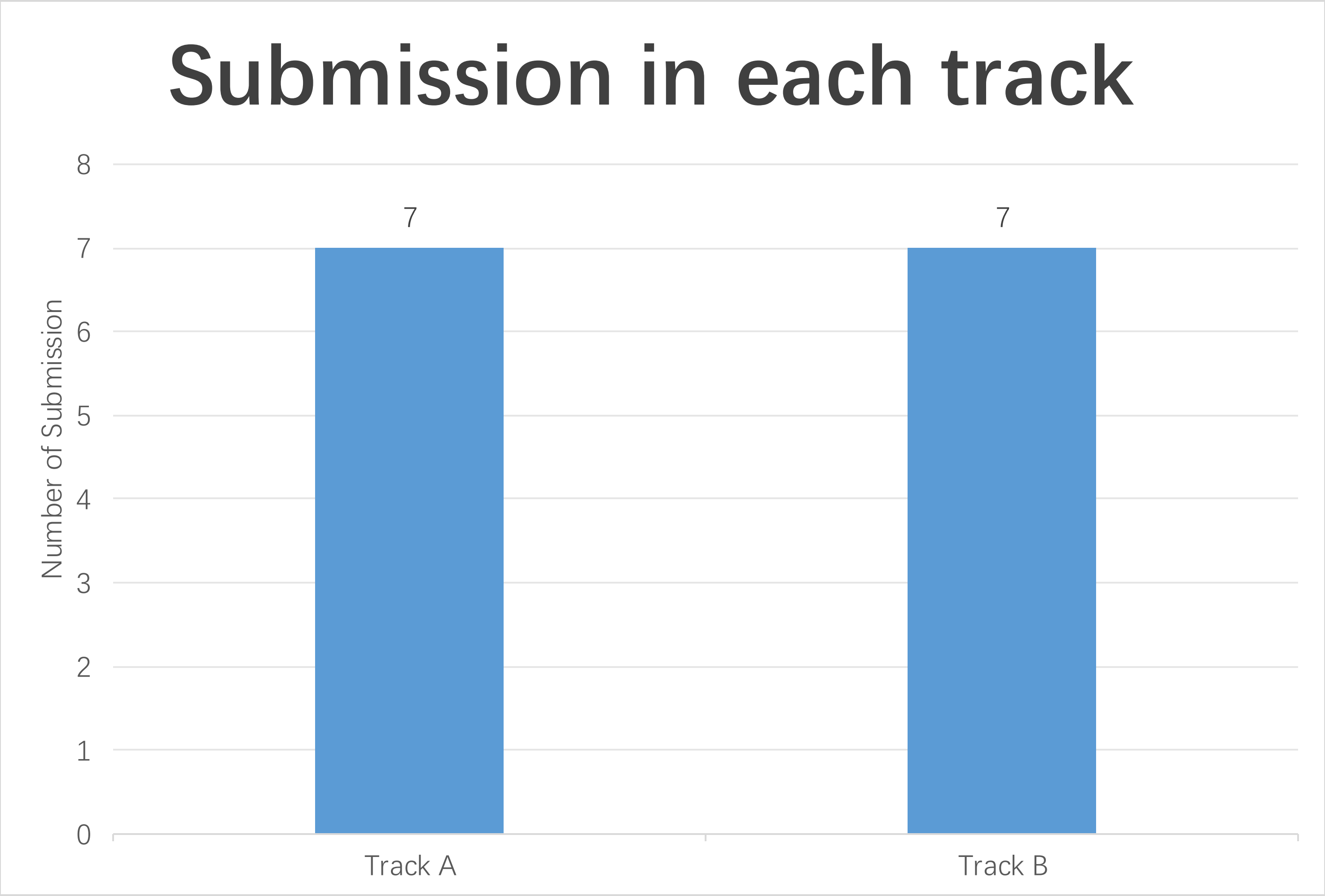
Statistics

Registration



Submission





Leaderboard

Track A: MLP Results		
Affiliation	**Team**	**Weighted Averaged Score**
ByteDance	audiocodec	0.865
ByteDance	GAEBT	0.860
ByteDance	AudioX	0.836
Carnegie Mellon University	CMU	0.827
Alibaba	Aluminumbox	0.807
NTT	Probin	0.709
IRIT	SAMoVA	0.516

Leaderboard

Track B: KNN Results		
Affiliation	**Team**	**Weighted Averaged Score**
ByteDance	audiocodec	0.792
ByteDance	GAEBT	0.782
ByteDance	AudioX	0.778
NTT	Probin	0.710
Carnegie Mellon University	CMU	0.707
Alibaba	Aluminumbox	0.641
IRIT	SAMoVA	0.480

Best Result in Each Track

Track	Affiliation	Team	Weighted Averaged Score
Track A: MLP Results	ByteDance	audiocodec	0.865
Track B: KNN Results	ByteDance	audiocodec	0.792

Baseline Sharing

Yadong Niu

Motivation

- Audio models exhibit a generalization gap:
  - ✓speech-trained models not work well for sounds
  - ✓Sound pretraining does not work for speech

- Self-supervised learning (SSL) audio representations show promise but lack exploration in scaled model and dataset sizes for cross-domain generalization.

- Need for a unified encoder to bridge speech, sound, and music domains efficiently.

Proposed Method: Dasheng

- **Design:** Masked Autoencoder (MAE) + Unprecedented Scale

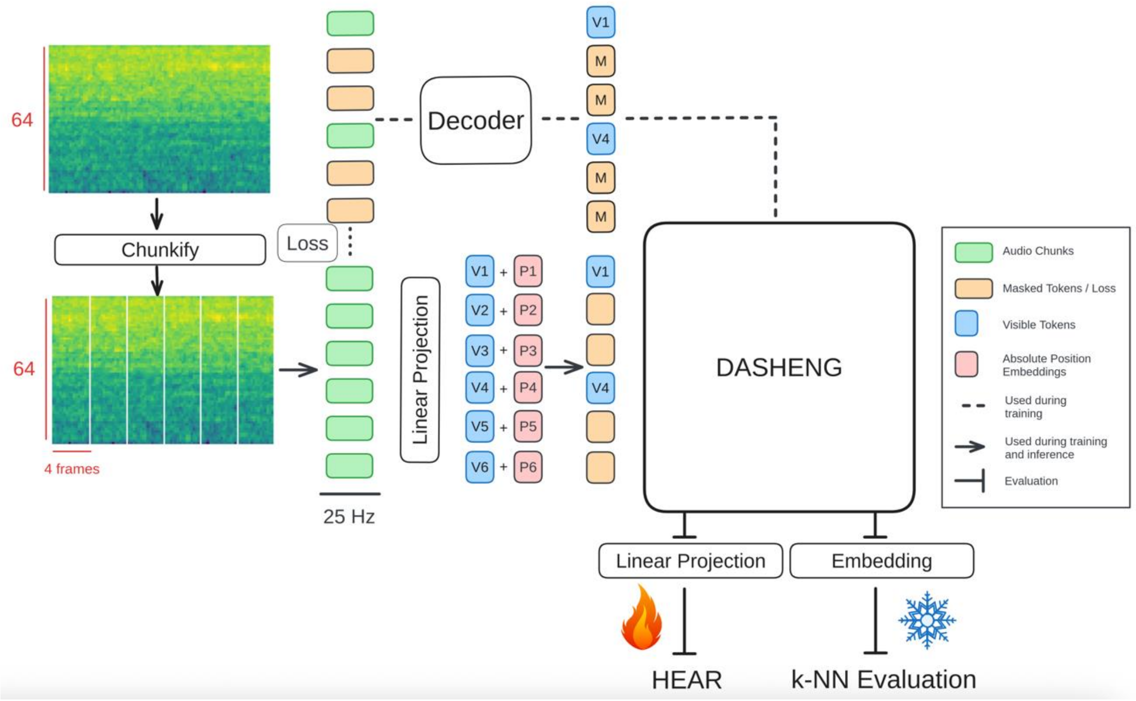


Fig. 1. The Dasheng Training Framework

Features

✓training-data: 272k hours

✓parameters: max to 1.2B

✓frame-level representation

Dominant Results

- Evaluation on HEAR Benchmark (Dasheng: Blue line in Fig.2)
- Embeddings inherently capture rich information across domains

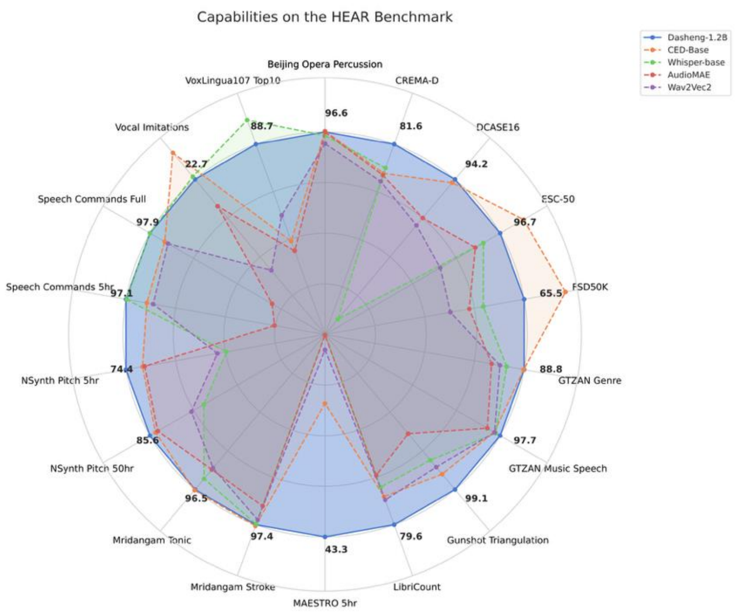


Fig. 2. Evaluation on the HEAR benchmark

Table 1. Evaluation using a k-NN classifier

Domain	Task	AudioMAE	Dasheng Base	Dasheng 0.6B	Dasheng 1.2B
Env	ESC50	53.1	61.9	66.6	**68.6**
	FSDK18	43.4	70.3	**72.1**	**72.1**
	US8k	58.2	73.9	75.9	**77.7**
Music	NS_{Inst}	67.2	70.0	70.9	**71.2**
Speech	SPC1	56.9	93.6	93.4	**95.9**
	SPC2	5.9	86.0	87.3	**90.9**
	VoxCeleb1	2.9	34.2	37.8	**39.4**
	RAVDESS	28.7	58.1	61.8	**61.9**
	FSC	7.6	52.3	57.6	**62.4**

Dominant Results

- Evaluation: X-ARES Benchmark
  - ✓Domain: speech, sound, music
  - ✓Tracks: Parametric & Non-parametric

Domain	Dataset	Task Type	Metric	n-classes	Track B	Hidden
Speech	ASV2015	Spoofing detection	EER	2	✓	✗
	CREMA-D	Emotion recognition	Acc	5	✓	✗
	Fluent Speech Commands	Intent classification	Acc	248	✓	✗
	LibriCount	Speaker counting	Acc	11	✓	✗
	LibriSpeech	Gender classification	Acc	2	✓	✗
	LibriSpeech	Speech Recognition	iWER	-	✗	✗
	Speech Commands	Keyword spotting	Acc	30	✓	✗
	VocalSound	Non-speech sounds	Acc	6	✓	✗
	VoxCeleb1	Speaker identification	Acc	1251	✓	✗
	VoxLingua107	Language identification	Acc	33	✓	✗

Table 2. Average result on X-ARES Benchmark

Domain	Dataset	Task Type	Metric	n-classes	Track B	Hidden
Sound	Clotho	Sound retrieval	Recall@1	-	✗	✗
	DESED	Sound event detection	Segment-F1	10	✓	✗
	ESC-50	Environment classification	Acc	50	✓	✗
	Finger snap sound	Sound event detection	Acc	2	✓	✓
	FSD18-Kaggle	Sound event detection	mAP	41	✗	✗
	FSD50k	Sound event detection	mAP	200	✗	✗
	Inside/outside car	Sound event detection	Acc	2	✓	✓
	Key scratching car	Sound event detection	Acc	2	✓	✓
	LiveEnv sounds	Sound event detection	mAP	18	✗	✓
	Subway broadcast	Sound event detection	Acc	2	✓	✓
	UrbanSound 8k	Urban sound classification	Acc	10	✓	✗
Music	Free Music Archive Small	Music genre classification	Acc	8	✓	✗
	GTZAN Genre	Genre classification	Acc	10	✓	✗
	MAESTRO	Note classification	Acc	88	✓	✗
	NSynth-Instruments	Instruments Classification	Acc	11	✓	✗
	NSynth-Pitch	Pitches Classification	Acc	128	✓	✗

Dominant Results

- Evaluation: X-ARES Benchmark

Table 3. Average result on X-ARES Benchmark

Accessibility	Method	Dasheng	Wav2vec2	Whisper	Data2vec
Public Dataset	**MLP**	0.699	0.490	0.632	0.598
	KNN	0.504	0.262	0.299	0.388
Total Dataset	**MLP**	0.801	0.664	0.740	0.694
	kNN	0.683	0.469	0.475	0.455

Conclusion

- Audio encoders are vital for unified audio understanding, closing domain gaps through scaled SSL.

- Collaborative focus on speech, sound, and music enables generalizable models with real-world impact.

- **Future: unified audio understanding ＋ generation ?**

Thanks

Award Winners with Certificates

Wenwu Wang

1st Place of Track A


Certificate of honor

THIS CERTIFICATE IS PRESENTED TO

Linping Xu

(audiocodec)

The 1st place

of Track A:Linear Fine-Tuning on Task-Specific Data

ICME 2025 Audio Encoder Capability Challenge

In conjuntion with ICME 2025

June 30th - July 4th 2025, Nantes, France

Junbo Zhang yadong Niu Xiaoman Wang

UNIVERSITY OF SURREY

DataoceanAI

Team: audiocodec   Linping Xu

2nd Place of Track A

Team: GAEBT　　Jiawei Jiang

2nd Place of Track A

Team: AudioX　　Qingbo Huang

3rd Place of Track A

Team: CMU

Shikhar, Hyejin Shim, Samuele Cornell, Kwanghee Choi,
Satoru Fukayama, Jeeweon Jung, Soham Deshmukh,
Shinji Watanabe

3rd Place of Track A

Certificate of honor

THIS CERTIFICATE IS PRESENTED TO

Xiang Lyu

(Aluminumbox)

The 3rd place

of Track A:Linear Fine-Tuning on Task-Specific Data

ICME 2025 Audio Encoder Capability Challenge

In conjuntion with ICME 2025

June 30th - July 4th 2025, Nantes, France

Junbo Zhang yadong Niu Xiaoman Wang

UNIVERSITY OF SURREY

DataoceanAI

Team: Aluminumbox Xiang Lyu

3rd Place of Track A

Certificate of honor

THIS CERTIFICATE IS PRESENTED TO

Daisuke Niizumi

(Probin M2D)

The 3rd place

of Track A:Linear Fine-Tuning on Task-Specific Data

ICME 2025 Audio Encoder Capability Challenge

In conjuntion with ICME 2025

June 30th - July 4th 2025, Nantes, France

Xiaoman Wang

UNIVERSITY OF SURREY

DataoceanAI

Team: Probin M2D Daisuke Niizumi

3rd Place of Track A

Certificate of honor

THIS CERTIFICATE IS PRESENTED TO

Ludovic Tuncay

(SAMoVA)

The 3rd place

of Track A:Linear Fine-Tuning on Task-Specific Data

ICME 2025 Audio Encoder Capability Challenge

In conjuntion with ICME 2025

June 30th - July 4th 2025, Nantes, France

Xiaoman Wang

UNIVERSITY OF SURREY

DataoceanAI

Team: SAMoVA Ludovic Tuncay

1st Place of Track B

Certificate of honor

THIS CERTIFICATE IS PRESENTED TO

Linping Xu

(audiocodec)

The 1st place

of Track B:Unparameterized Evaluation

ICME 2025 Audio Encoder Capability Challenge

In conjuntion with ICME 2025

June 30th - July 4th 2025, Nantes, France

Junbo Zhang Yadong Niu Xiaoman Wang

UNIVERSITY OF SURREY

DataoceanAI

Team: audiocodec Linping Xu

2nd Place of Track B


Certificate of honor

THIS CERTIFICATE IS PRESENTED TO

Jiawei Jiang

(GAEBT)

The 2nd place

of Track B:Unparameterized Evaluation

ICME 2025 Audio Encoder Capability Challenge

In conjuntion with ICME 2025

June 30th - July 4th 2025, Nantes, France

Junbo Zhang yadong Niu Xiaoman Wang

Team: GAEBT Jiawei Jiang

2nd Place of Track B

Certificate of honor

THIS CERTIFICATE IS PRESENTED TO

Qingbo Huang

(AudioX)

The 2nd place

of Track B:Unparameterized Evaluation

ICME 2025 Audio Encoder Capability Challenge

In conjuntion with ICME 2025

June 30th - July 4th 2025, Nantes, France

UNIVERSITY OF SURREY

DataoceanAI

Team: AudioX　　Qingbo Huang

3rd Place of Track B

Team: CMU

Shikhar, Hyejin Shim, Samuele Cornell, Kwanghee Choi, Satoru Fukayama, Jeeweon Jung, Soham Deshmukh, Shinji Watanabe

3rd Place of Track B

Team: Aluminumbox Xiang Lyu

3rd Place of Track B

Certificate of honor

THIS CERTIFICATE IS PRESENTED TO

Daisuke Niizumi

(Probin M2D)

The 3rd place

of Track B:Unparameterized Evaluation

ICME 2025 Audio Encoder Capability Challenge

In conjuntion with ICME 2025

June 30th - July 4th 2025, Nantes, France

Xiaoman Wang

UNIVERSITY OF SURREY

DataoceanAI

Team: Probin M2D Daisuke Niizumi

3rd Place of Track B

Team: SAMoVA Ludovic Tuncay

Reminder

If you did not join ICME in person， we will send the

e-certificate by email after the workshop.

CVSSP
Centre for Vision, Speech and Signal Processing

UNIVERSITY OF SURREY

Neural Audio Codec

Wenwu Wang

Centre for Vision, Speech and Signal Processing (CVSSP)

University of Surrey

United Kingdom

UKRI Research England

gain GAMES AND INNOVATION NEXUS

BRITISH COUNCIL

EPSRC Engineering and Physical Sciences Research Council

Outline

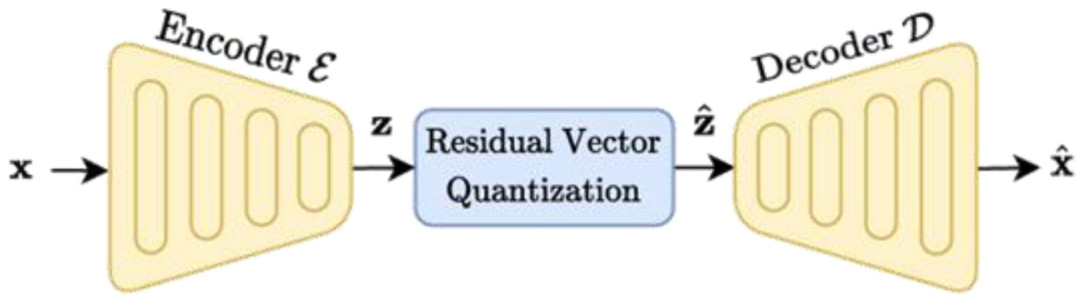
- Background: Audio Codec & Neural Audio Codec
- Motivation: Limitations with Current Neural Codecs
- Proposed Method: **SemantiCodec**
- Experimental Results
- Conclusions and Future Works

Audio Codec

- A device or software that encodes or decodes digital audio data for transmission, storage, or playback.
- **Co**mpressor-**dec**ompressor → “Codec”
- MP3, FLAC, AAC, Vorbis, etc.

Neural Audio Codec

- SoundStream (Zeghidour et al. 2021)
- Encodec (Défossez et al. 2021)
- Descript (Kumar et al. 2023)
- HiFi-Codec (Yang et al. 2023)
- SpeechTokenizer (Zhang et al, 2023)
- Codec Superb Benchmark (Wu et al. 2024)
  - https://github.com/voidful/Codec-SUPERB



Limitations of Current Neural Codecs

- **High token rate (long token sequence)**
  - e.g., 6kbps Descript audio codec has 600 tokens per second
  - Make auto-regressive modeling challenging and computational expensive
- **Poor reconstruction quality at low bit rate (e.g., 0.6 kbps).**
  - Most previous studies work on bit rate > 2kbps
  - Can we go further under 1.0 kbps?
- **Falling short in capturing semantic information**
  - *For example, latent encodings given by 6kbps achieved an average accuracy of only 33% on the HEAR benchmark, while AudioMAE latent encodings achieved an accuracy of 61% (without finetuning).*

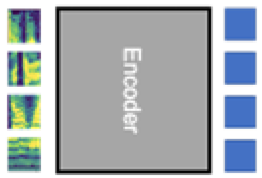
Motivations

- Shorter sequence: Lower token rates at 25, 50, or 100 tokens per second.
- Better reconstruction at lower bit rate: 0.3~1.4 kbps
- Improved semantic in the codec tokens (which potentially can lead to better language modelling)

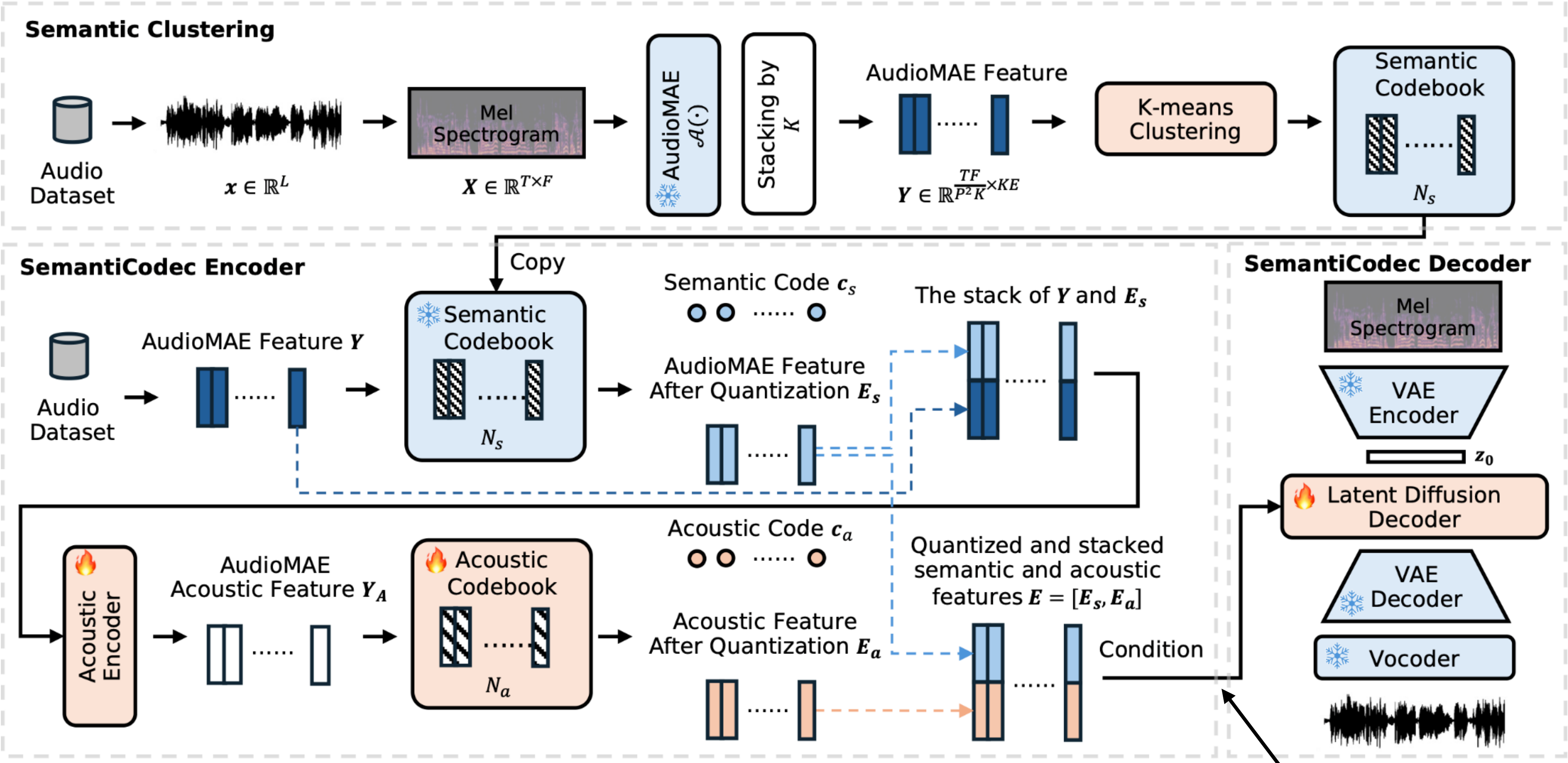
SemantiCodec

Large scale k-means is challenging
AudioSet + Million Song Dataset + GigaSpeech
https://github.com/haoheliu/kmeans_pytorch

- Ultra-low bit rate (0.31 kbps ~1.40 kbps, token rate 25, 50, or 100 per second) & Strong semantics in the token & Variable vocabulary sizes



AudioMAE Encoder



CrossAtten

https://haoheliu.github.io/SemantiCodec/

Experimental Evaluations

- **Metrics:**
  - MEL: Mel spectrogram distance
  - STFT: Short-time Fourier Transform Distance
  - ViSQOL: Virtual Speech Quality Objective Listener Score
  - WER: Word Error Rate
  - Accuracy: Audio classification task accuracy
  - MUSHRA Scores
- **Datasets (we only use audio):**
  - GigaSpeech (10K hours), Million Song Dataset (510K music tracks), MedleyDB (122 tracks), MUSDB18 (10 hours), AudioSet (2M), VGGSound (190K), WavCaps (7K hours)

Experimental Evaluations

- **Baselines:**
  - Encodec (EC): 23M parameters
  - Descript Codec (DAC): 74M
  - HiFi-Codec (HC): 63M
- **Evaluations:**
  - Reconstruction performance evaluation: LibriTTS clean test set (300 speech utterances), AudioSet general sound (500 audio signals), MUSDB18 (50 songs covering vocals, drums, bass, etc.); 1050 audio clips in total
  - Semantic information evaluation: HEAR benchmark (NSPitch, ESC-50, LibriCount, CREMA-D, Vocal Imitations (VoImit), Speech Commands(SC))
  - Subjective evaluations: MUSHRA test: 10 raters, 10% of evaluation data (25 music tracks, 30 speech recordings, and 50 general sound samples)

Samples Comparison

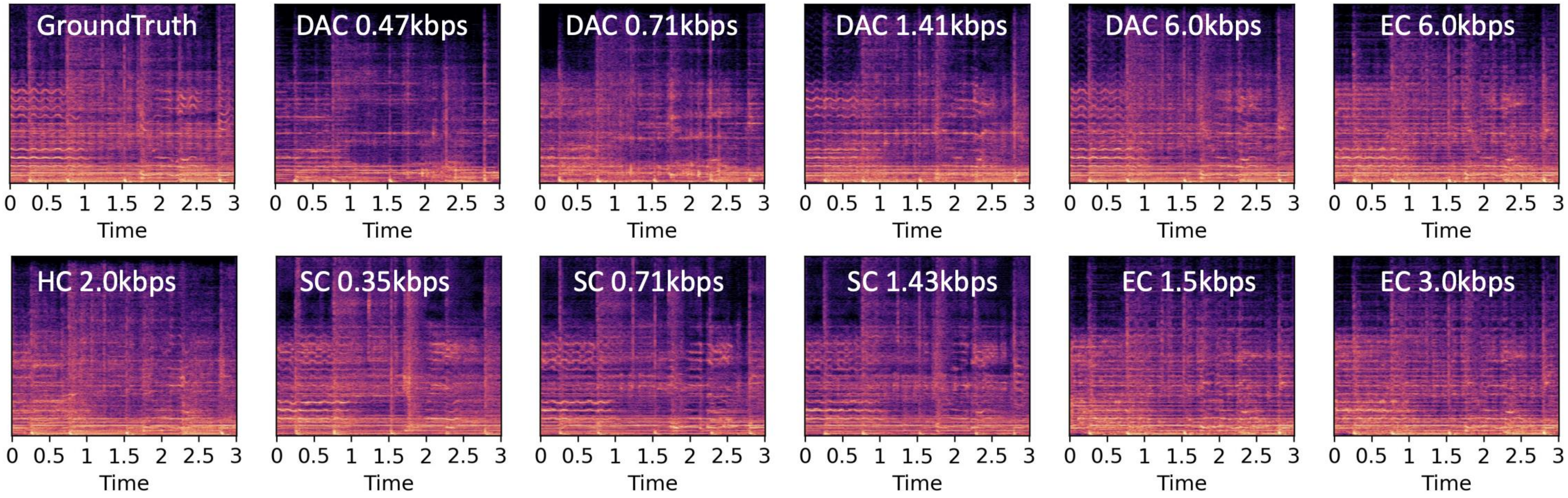
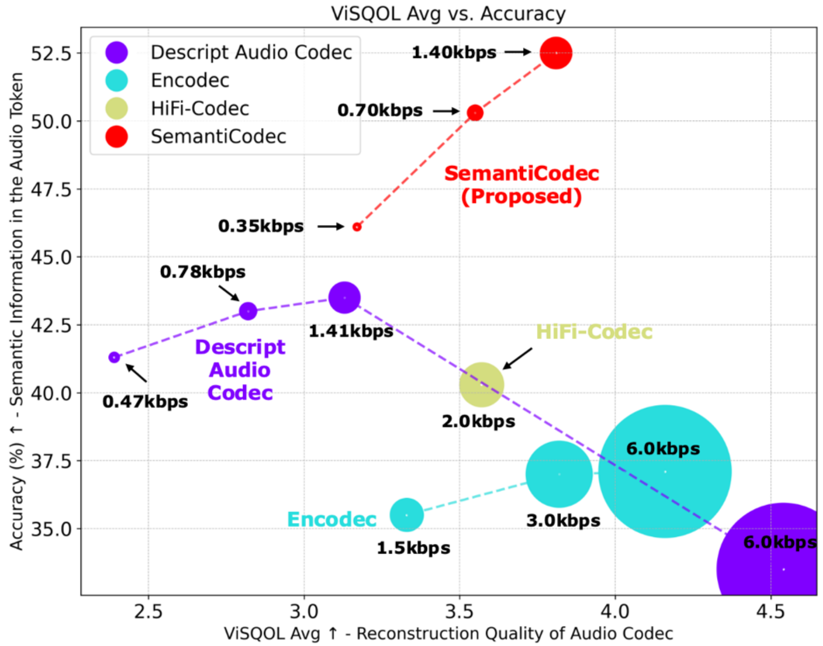
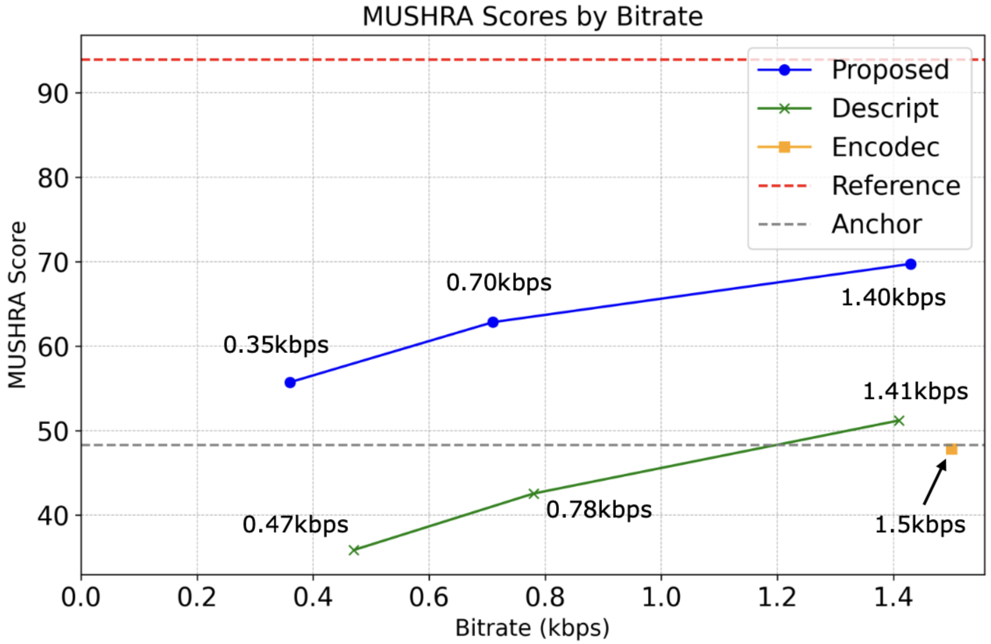


Fig. 3. The log-STFT spectrogram of the ground truth audio and the reconstruction audio with different audio codecs. DAC, EC, HC, and SC are the descript codec, Encodec, HiFi-Codec, and SemantiCodec, respectively.

Visual Comparison

- Better reconstruction with a lower bit rate
- Better semantic in the audio token (potentially better Audio LLM?)





Sound demos: https://haoheliu.github.io/SemantiCodec/

Demos

	Original	HiFi-Codec (2.0 kbps)	Encodec (1.5 kbps)	DAC (1.41 kbps)	SemantiCodec (1.43 kbps)	DAC (0.47 kbps)	SemantiCodec (0.35 kbps)
Music (MUSDB18)							
General Audio (AudioSet)							
Speech (Libri)							

More sound demos: https://haoheliu.github.io/SemantiCodec/

H. Liu, X. Xu, Y. Yuan, M. Wu, W. Wang, M.D. Plumbley, "SemantiCodec: An Ultra Low Bitrate Semantic Audio Codec for General Sound," IEEE Journal on Selected *Topics in Signal Processing*, vol. 18, no. 8, pp. 1448--1461, 2024.

Conclusion & Future Work

- We have presented SemantiCodec, which produces audio tokens of **more semantic**, **lower token rate**, with thus enabling ultra-low bit rate audio codec, and works for any length audio.
- Future works:
  - How does the semantic of audio token related with LM performance?
  - Can shortened audio token sequence alleviate LM robustness issue?
  - Specialized SemantiCodec

Acknowledgements

Thanks to Haohe for providing slides.

The project is funded in part by EPSRC, British Council and GAIN program (Research England).

Take Aways

- Arxiv Paper: https://arxiv.org/abs/2405.00233
- Project Page: https://haoheliu.github.io/SemantiCodec/
- Open-source Code: https://github.com/haoheliu/SemantiCodec-inference
- Demos: https://haoheliu.github.io/SemantiCodec/

Dataocean AI Company Overview

Helen Wang

DataoceanAI

High-Quality ASR Dataset for Incarbin and outside

DataoceanAI

Overview

Volume: 15,000 Sentences (10,000 for training / 5,000 for evaluation)

Participants: 430 speakers

Recording Location: Inside and outside vehicles

Scenarios: Outdoor and underground parking lots

DataoceanAI

High-Quality ASR Dataset for Incarbin and outside

Data Collection Details (In-Car)

- 5 fixed speaker positions
- Microphone fixed at point A
- 7,500 sentences total (1,500 per point)

Data Collection Details (Outside-Car)

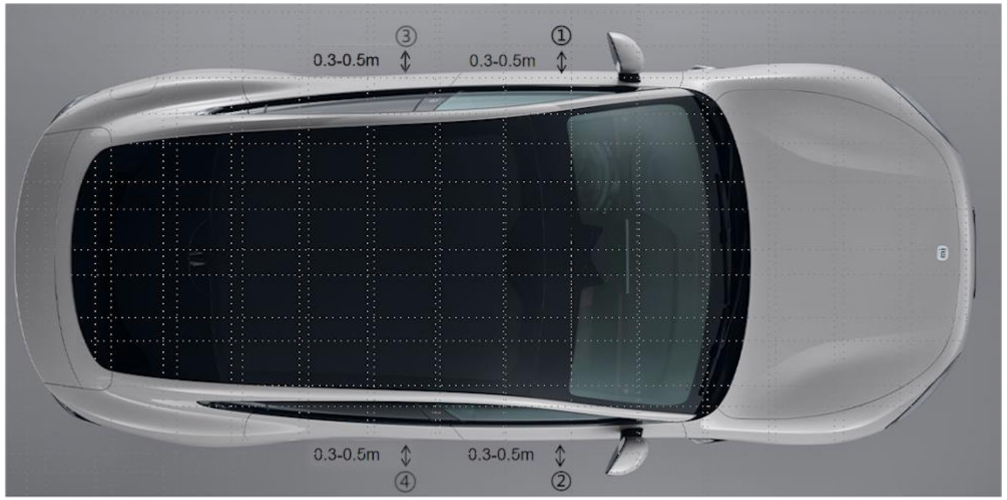
- 4 recording positions around the vehicle
- Microphone fixed at point A

Recording Conditions

- Windows fully closed
- Window slightly open near the sound source

Recording Conditions

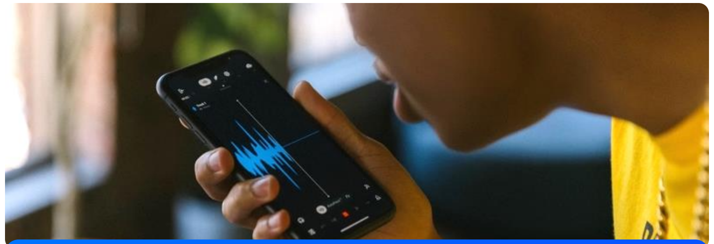
- Windows fully open
- Windows fully closed
- Windows slightly open



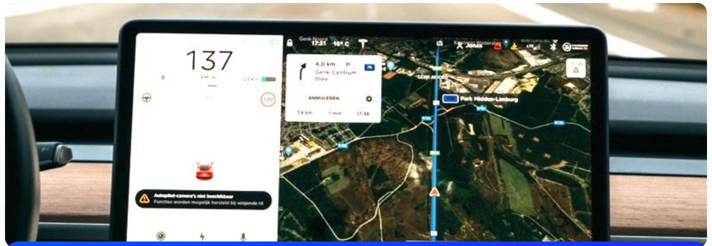
www.

Off-The-Shelf Datasets

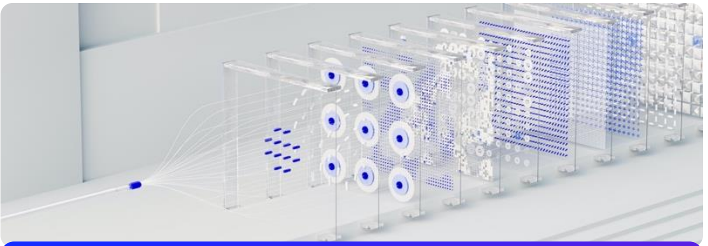
Over 1600 OTS datasets ready to go!



1179 SPEECH RECOGNITION SPEECH SYNTHESIS



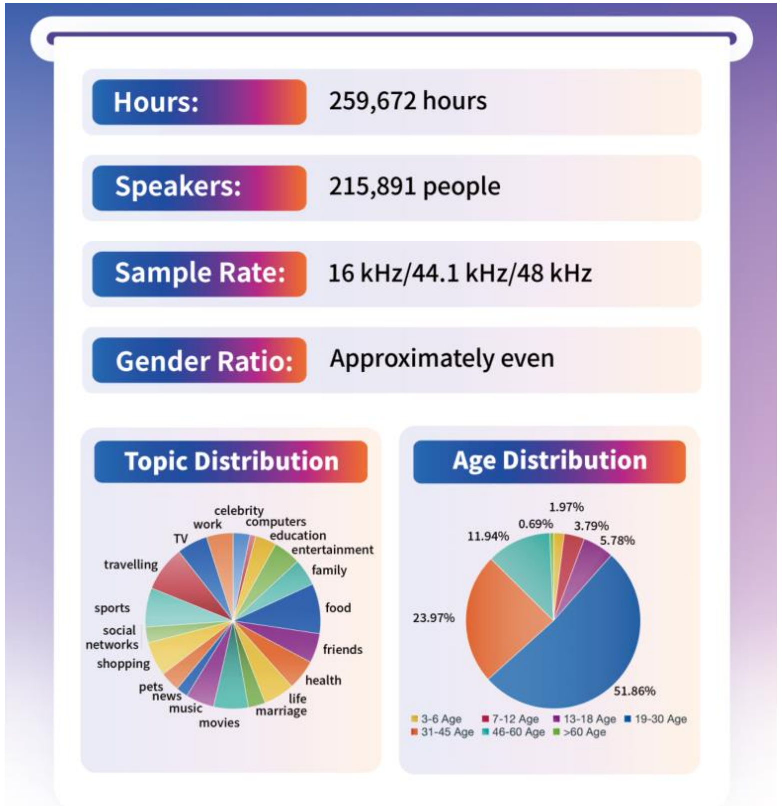
138 COMPUTER VISION

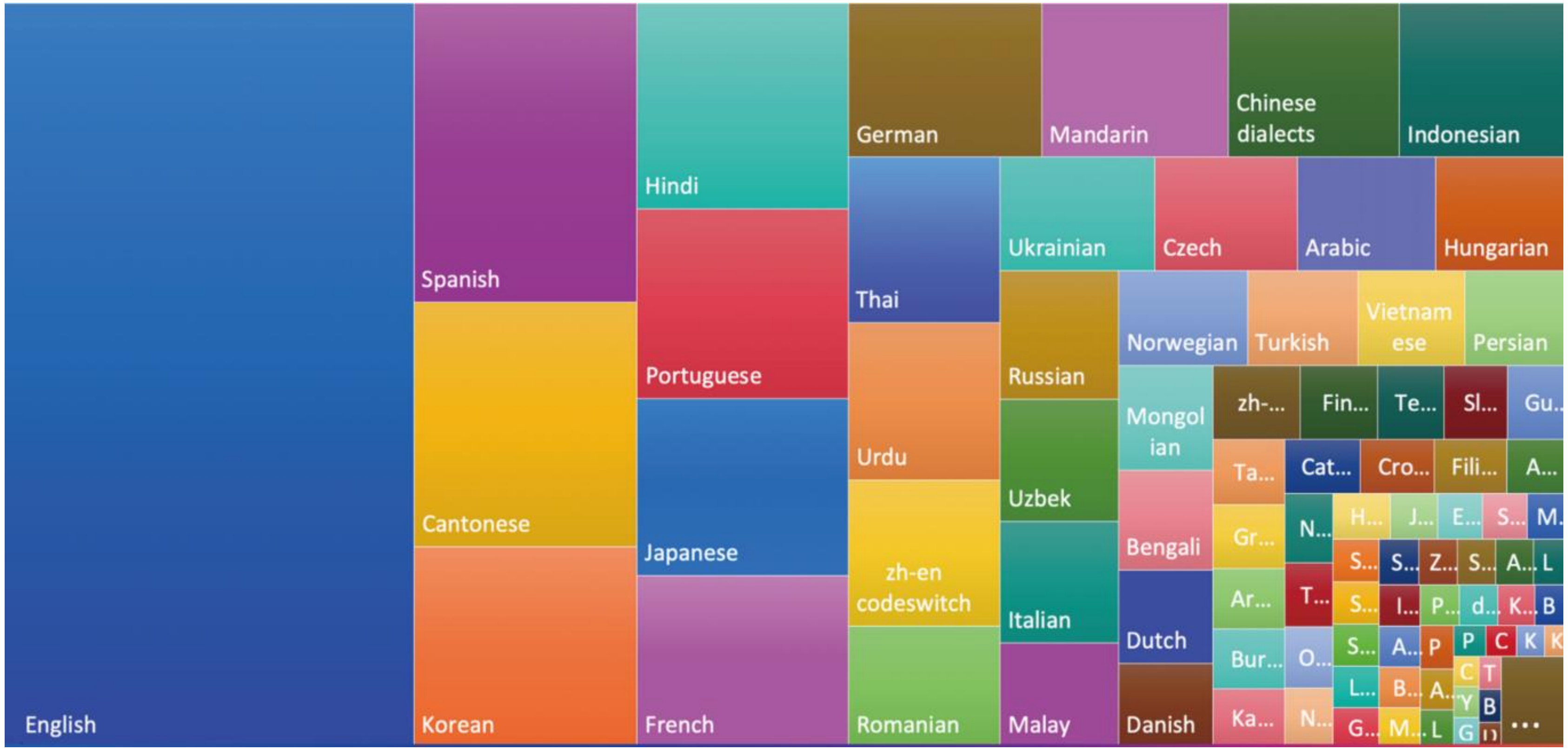


360 NATURAL LANGUAGE PROCESSING

Application Scenarios		
Personal assistant, voice input, smart home, intelligent customer service, robot, voice navigation, intelligent broadcast, voice translation, mobile social networking, virtual human, smart finance, etc.	Intelligent driving, mobile socialization, virtual humans, smart finance, smart transportation, smart city, OCR recognition, etc.	CoT, coding, machine translation, intelligent Q&A, information extraction, sentiment analysis, etc.

High Quality 100+ Languages Speech Dataset





Thanks！

Website: Dataoceanai.com

Presentation 1 by Team audiocodec

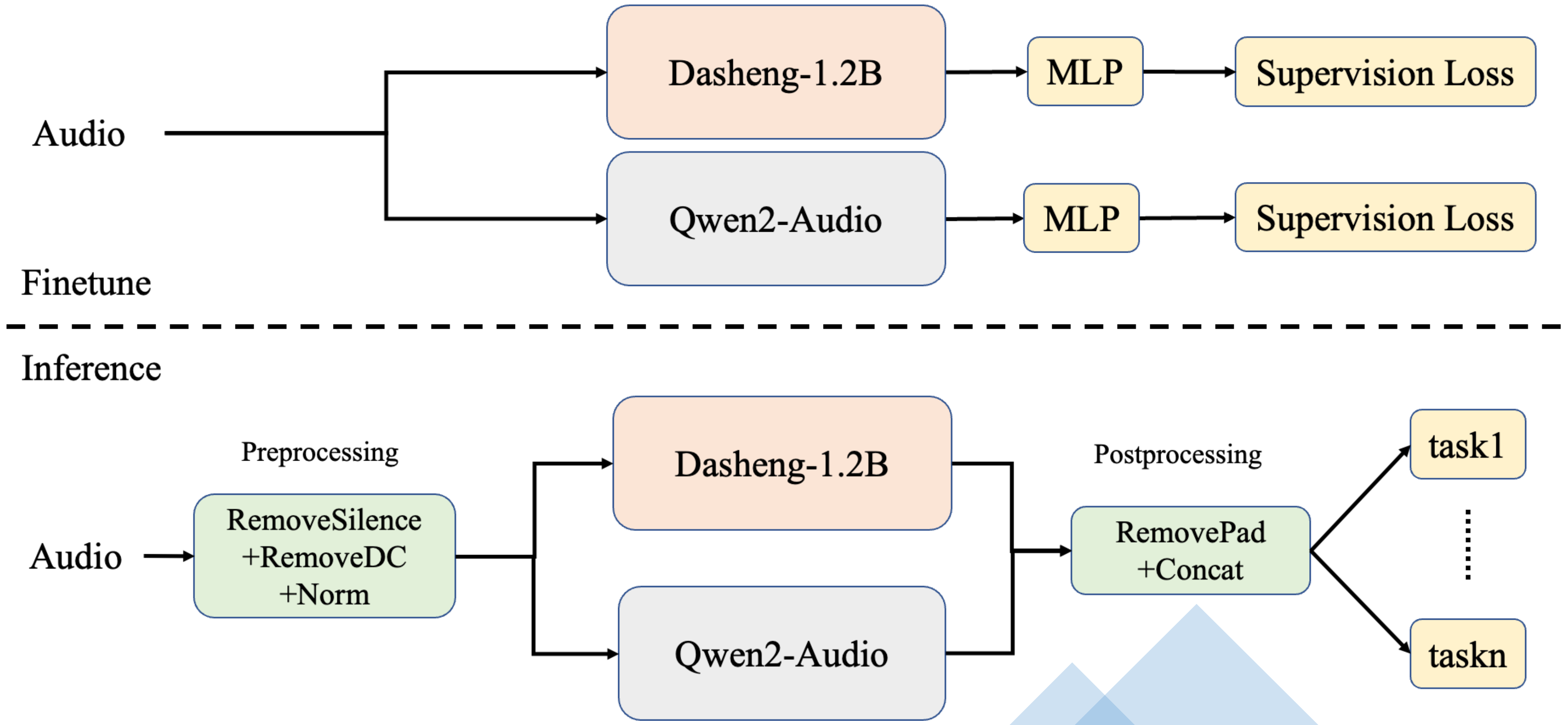
DQFAudio Encoder: A Solution for Audio Downstream Tasks Based on Model Ensemble and FineTuning

ByteDance MMLab
Team: audiocodec
Linping Xu
xulinping.678@bytedance.com

ByteDance 字节跳动

- **Tasks Analysis**
- **Pre & Postprocessing**
- **Model Ensemble**
- **Model Fine-tuning**

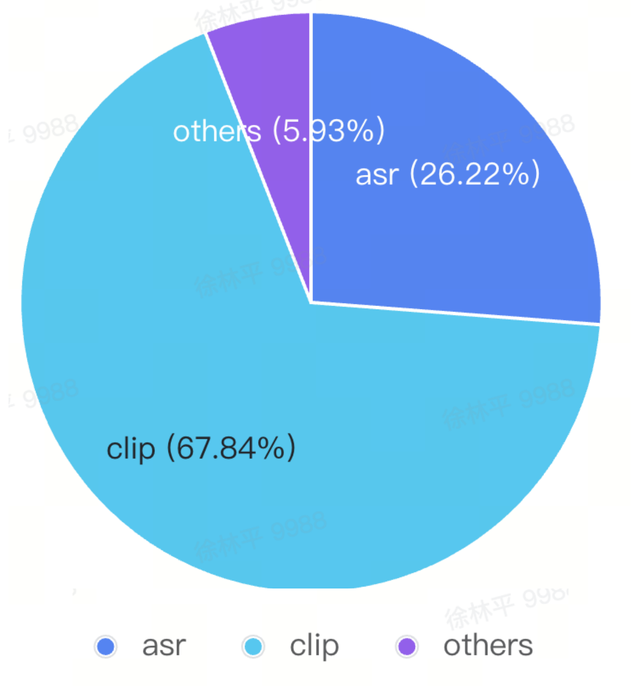
Simple System Description



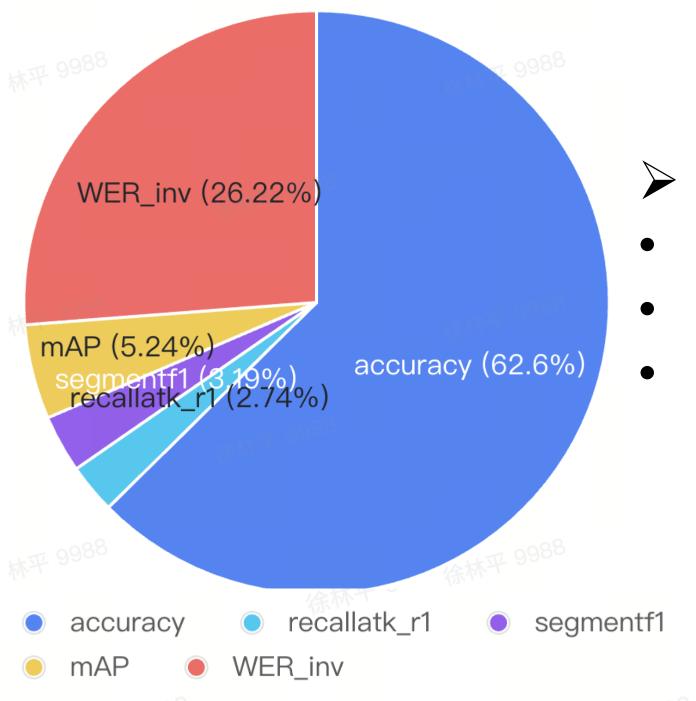
■ Tasks Analysis

➢ 23 Tasks Evaluation Weight
- Clip-Level Tasks: 67+% (by task type)
- Classification Tasks: 62+% (by type & metrics)
- Speech Recognition: 26+%

Tasks type eval weight



Metrics type eval weight



➢ Model Selection & Fine-tuning
- Integrate **audio and speech** model
- Fine-tuning focus **clip classification** tasks

➢ Evaluation Efficiency optimization
- 14 quick-representative tasks for development
- Process split: Encoder Inference ‖ Task Evaluation
- Benefits: **Faster results • Higher GPU efficiency**

Note: *Development-time Xares evaluation stats, not synced with latest commit.*

■ Pre & Postprocessing

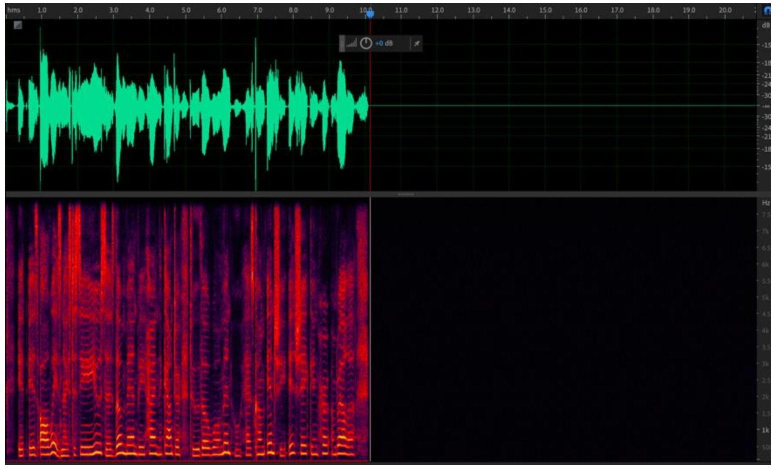
➢ Evaluation Audio Analysis

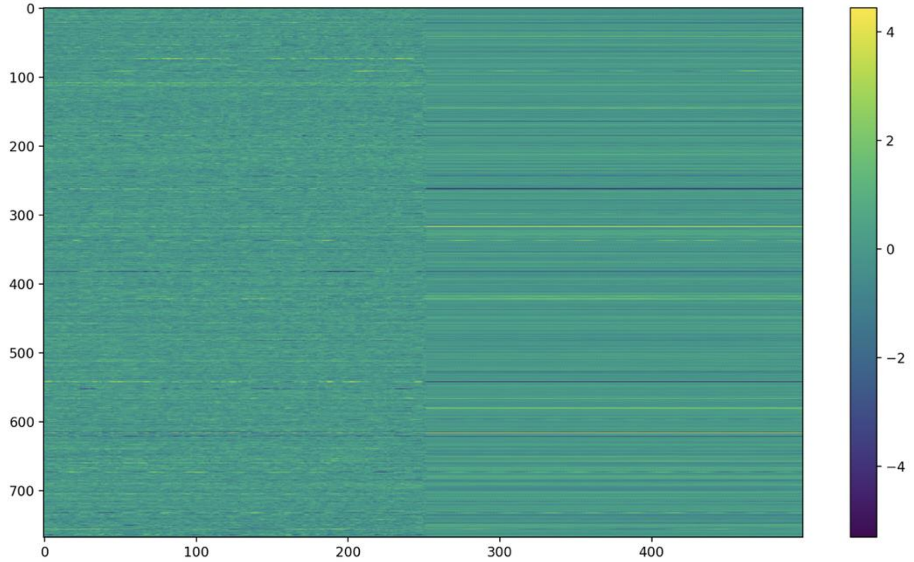
- **Long silent segments** (≥0.5s)
- **Fixed-length input** requirement (e.g., Dasheng 10s, Whisper 30s)
- **Silence padding** for sub-length sequences when inference

➢ Silent segments influence

- **Dilutes effective information** in audio features
- Degrades encoder's performance especially Clip-Level Tasks

Silent Padding Impact on Embeddings





ByteDance 字节跳动

■ Pre & Postprocessing

- Preprocessing:
  - Detect & filter silent segments (10ms units)
  - Applied DC removal + amplitude normalization

- Postprocessing:
  - Remove invalid features from silent padding

- Impact: 14 tasks' weighted scores up! 📈
  - Remove Silence
    - MLP: 0.705 → 0.712
    - KNN: 0.524 → 0.582
  - Remove DC and Norm
    - KNN: 0.582 → 0.598 🌟

	Preprocessing & Postprocessing			14 tasks' weighted scores	
BatchSize	RemSil	RemDC	Norm	MLP	KNN
16	\	\	\	0.704	0.504
1	\	\	\	0.705	0.524
1	\	on	\	0.705	0.523
1	on	\	\	**0.711**	0.582
1	on	on	\	**0.712**	**0.596**
1	on	on	on	0.710	**0.598**
16	on	on	on	0.707	0.546

Note:
Best inference performance achieved at batch size=1.
When batch size = 16, some invalid features are retained, due to the tensor [B, C, T] requirement.

■ Model Ensemble

➢ Task Divergence
Audio vs. speech tasks exhibit different characteristics, with speech recognition carrying significant evaluation weightage

➢ Parameter Flexibility
The challenge imposes no constraints on model parameter count, enabling architecture scaling.

➢ Model Domain Preferences
SSL Audio Encoders can be broadly categorized by application scenarios into Speech and Audio.

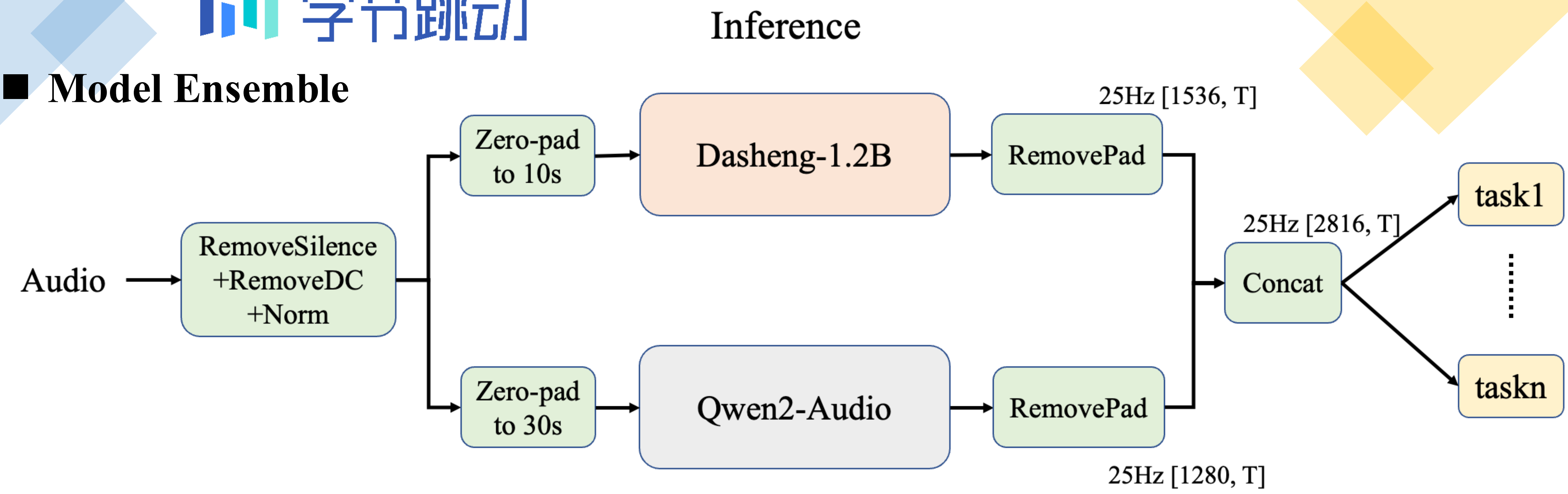
SSL for Speech & Audio	
Speech	Audio
Wav2vec2	Dasheng
Whisper	Beats
Qwen2-Audio encoder	…

1+1 > 2 ？

Ensembling Speech and Audio models may yield better performance than single models.

ByteDance 字节跳动

■ Model Ensemble



➢ Model Integration Trials:

- Dasheng
- Dasheng + Whisper
- Dasheng + Qwen2-Audio Encoder

Key Note: Embedding alignment required in temporal dimension for ensemble models

Model	MLP	KNN
Dasheng1.2B	0.731	0.560
Dasheng1.2B+Whisper	0.743	0.586
Dasheng1.2B+Qwen2Audio	**0.761**	**0.682**

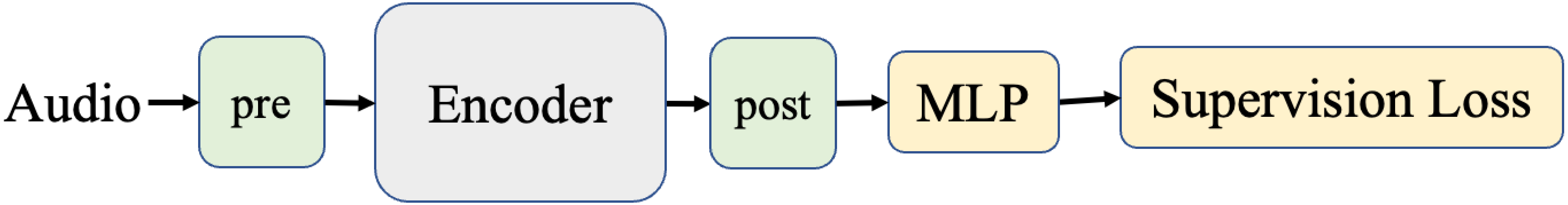
■ Model Fine-tuning & Ensemble

➢ Audioset for Fine-tuning

- Multi-task fitness: 527 classes
- Audio variety: speech, music environmental sounds
- Data scale: 5100+h

➢ Fine-tuning Strategy

- Approach: Fine-tune Qwen2-Audio Encoder and Dasheng separately, then fused embeddings.
- Model: Encoder + MLP



Fine-tuning Details

Model	Optimizer	learning rate	Masking Rate	Layer Freezing	others
Dasheng 1.2B	AdamW8bit	1.00E-05	0	ALL Layers	4*V100 BatchSize=12 EpochLength=500
Qwen2-Audio Encoder				Last 5 Layers	

In model fine-tuning, pre-/post-processing are applied

ByteDance 字节跳动

Model Fine-tuning & Ensemble

- Fine-tuning Strategy & Key Result
- Dasheng Base Model Results
  - **Fine-tuning significantly improved KNN scores**
  - Minimal impact on MLP performance

- Same Method to Larger Models
  - Applied same fine-tuning pipeline to: Dasheng 1.2B and Qwen2-Audio Encoder
  - Observed **consistent KNN improvements across models**

- Best solution
  - **Dasheng** 1.2 Finetuned + **Q**wen2Audio Finetuned Audio Encoder

Model	MLP	KNN
Dasheng base	0.704	0.504
Dasheng base*	0.703	0.622
Dasheng1.2B	0.731	0.560
Dasheng1.2B*	0.731	0.647
Dasheng1.2B+Qwen2Audio	**0.761**	0.682
Dahseng1.2B*+Qwen2Audio	0.756	0.722
Dasheng1.2B*+Qwen2Audio*	0.759	**0.726**

* Indicates the model has undergone fine-tuning.

ByteDance 字节跳动

■ Challenge Results

DQFAudio Encoder secured **first place** in the Weighted Averaged Score.

Results of the ICME 2025 Audio Encoder Capability Challenge

Track 1 MLP Results | Track 2 KNN Results

Track 1 MLP Results

Click on column headers to sort the table

Affiliation	Team	Report	Weighted Averaged Score ↓	asvspoof2015	clotho	cremad	desed	esc50	finger_snap	fluentspeechcommands
ByteDance	audiocodec	download	**0.865**	0.995	0.055	0.858	0.596	0.968	0.885	0.992
ByteDance	GAEBT	download	**0.860**	0.997	0.054	0.868	0.637	0.965	0.885	0.988
ByteDance	AudioX	download	**0.836**	0.986	0.058	0.862	0.602	0.964	0.884	0.991
Carnegie Mellon University	CMU	download	**0.827**	0.983	0.033	0.810	0.568	0.905	0.873	0.954
Alibaba	Aluminumbox	download	**0.807**	0.980	0.027	0.772	0.556	0.871	0.873	0.958
NTT	Probin	download	**0.709**	0.924	0.045	0.715	0.738	0.978	0.875	0.683
IRIT	SAMoVA	download	**0.516**	0.884	0.013	0.426	0.305	0.341	0.853	0.027

THANKS.

ByteDance 字节跳动

Presentation 2 by Team SAMoVA

Presented by **Ludovic TUNCAY** for **IEEE ICME 2025**
01.07.2025

Audio-JEPA

Joint-Embedding Predictive Architecture for Audio Representation Learning

Ludovic TUNCAY [1]
ludovic.tuncay@irit.fr

Étienne LABBÉ [1]
etienne.labbe@irit.fr

Emmanouil BENETOS [2]
emmanouil.benetos@qmul.ac.uk

Thomas PELLEGRINI [1]
thomas.pellegrini@irit.fr

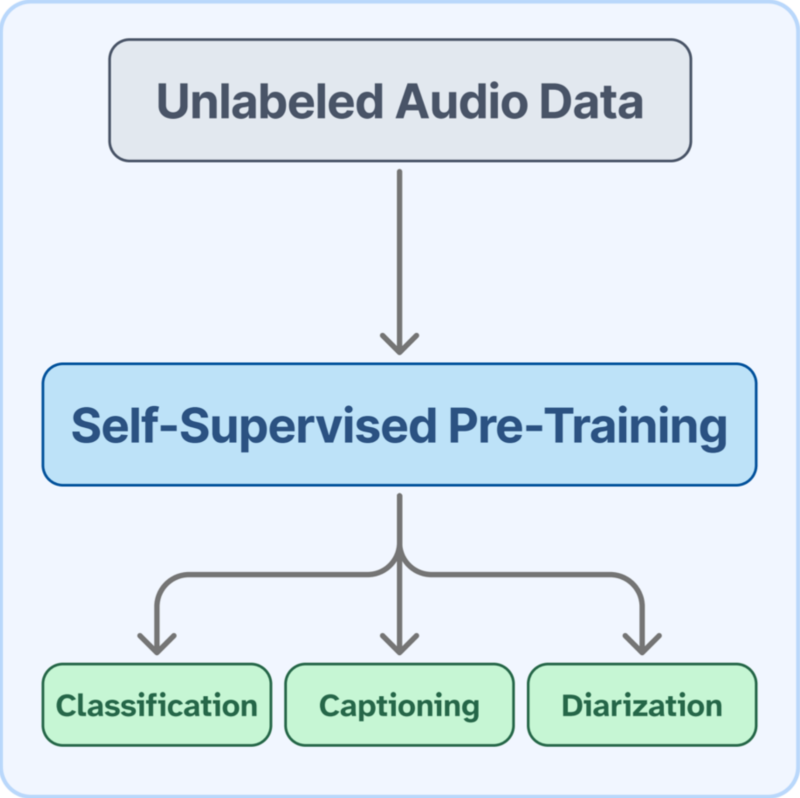
[1] IRIT, Université de Toulouse, CNRS, Toulouse INP | Toulouse, France
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London | UK

Presentation Overview

1. Introduction to SSL for Audio
2. The JEPA Paradigm
3. Audio-JEPA
4. Experimental Setup and Results
5. Limitations, Insights and Future Work

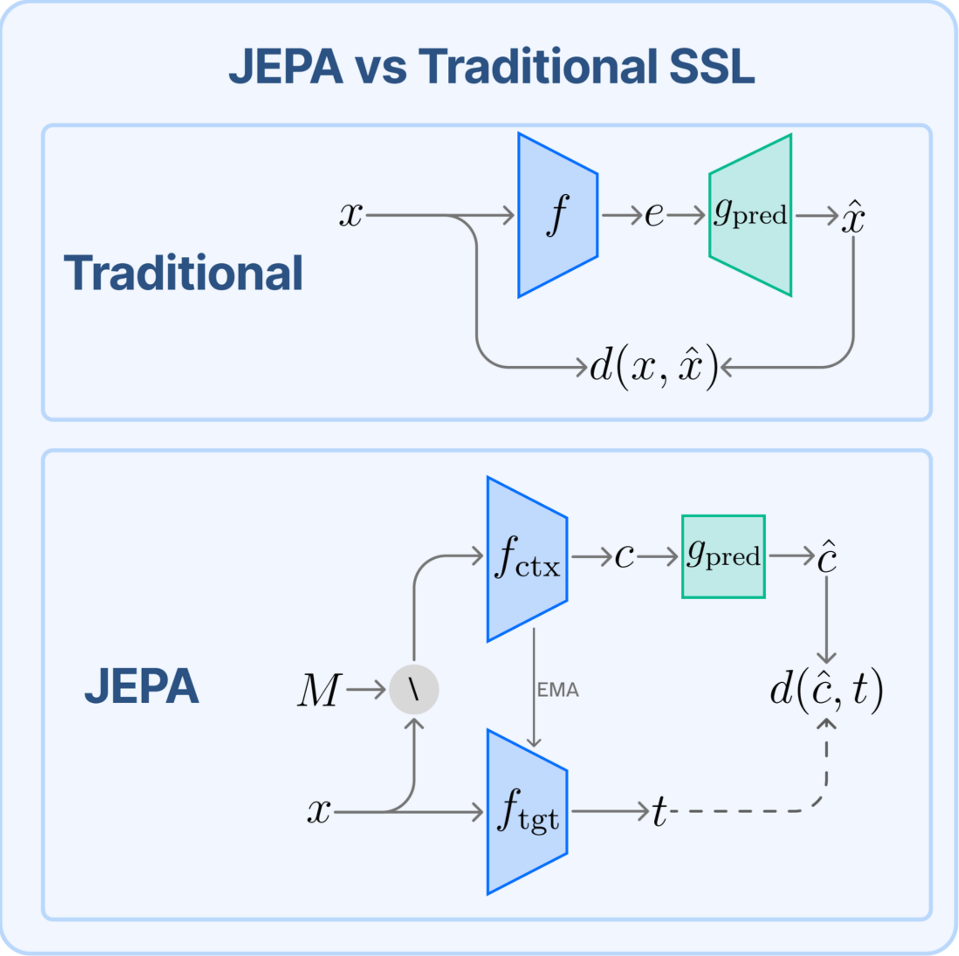
Introduction to Self-Supervised Learning for Audio

- Self-supervised learning leverages unlabeled data to learn useful representations
- Essential for audio, where annotated data is scarce and expensive to create
- Mordern approaches pre-train on large-scale datasets, then adapt to downstream tasks
- Reduces dependence on labeled data while maintaining strong performance



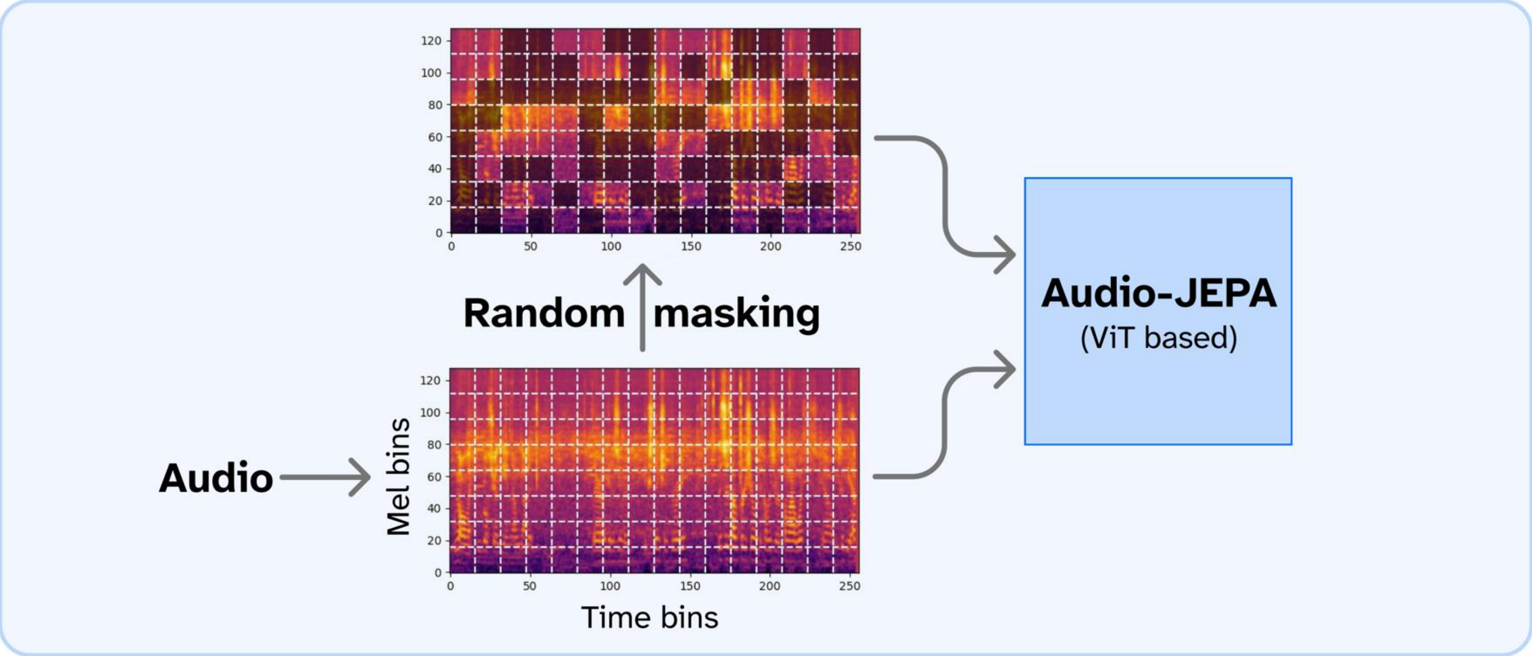
The JEPA Paradigm: Masked Latent Prediction

- Joint-Embedding Predictive Architectures (JEPA) **predict high-level latent features of masked regions**
- Rather than reconstructing raw inputs, JEPA encourages semantic feature learning by **reconstructing latent features**
- Proven effective for images (I-JEPA), video (V-JEPA, V-JEPA 2) and more modalities; **our work adapts it for audio**
- Focuses on **capturing meaningful semantic** structure because it works in the feature space



Audio-JEPA: Adapting JEPA to the Audio Domain

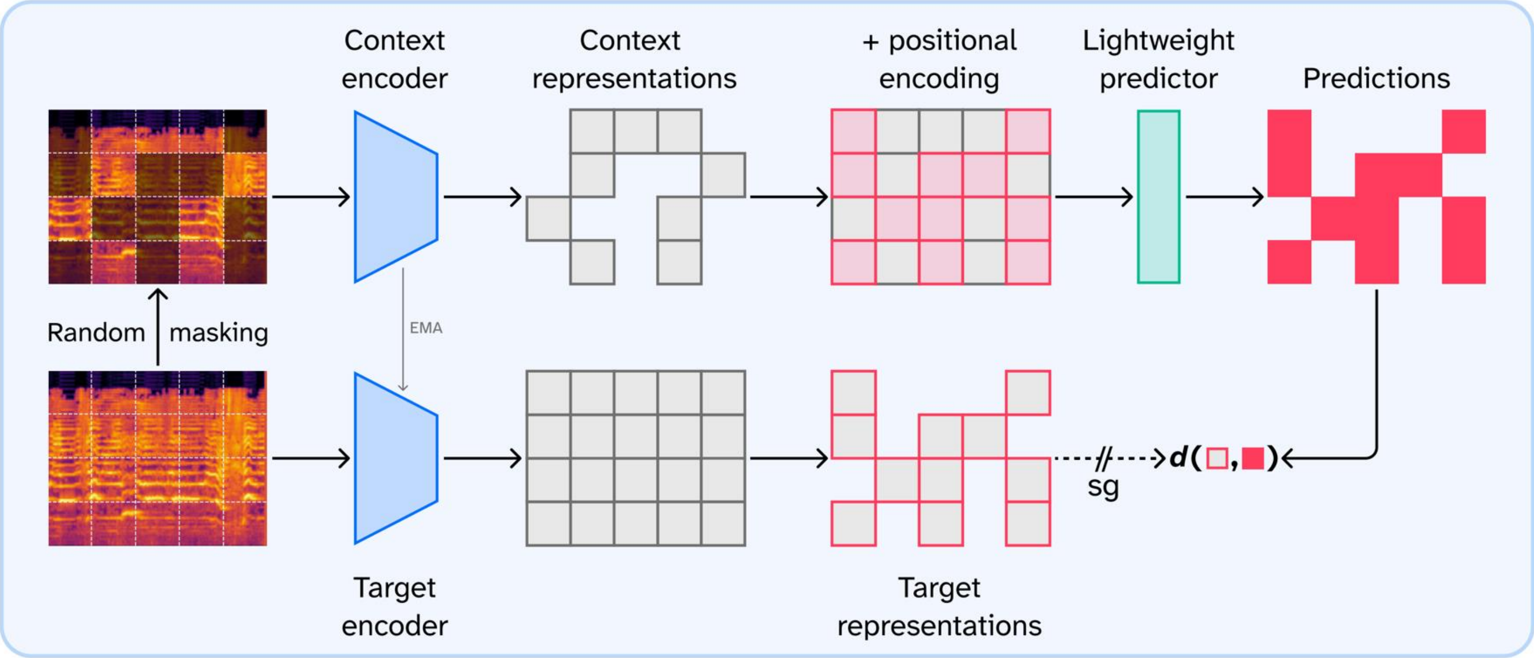
- Inputs: Convert audio waveforms to Mel-spectrograms (128 bands, 256 time bins)
- Randomly mask 40–60% of spectrogram patches for each sample
- Treats spectrograms as images. Feeds them into a Vision Transformer (ViT) backbone
- Each 16×16 patch spans and equivalent of 625ms of audio



Audio-JEPA: Overview of the Architecture

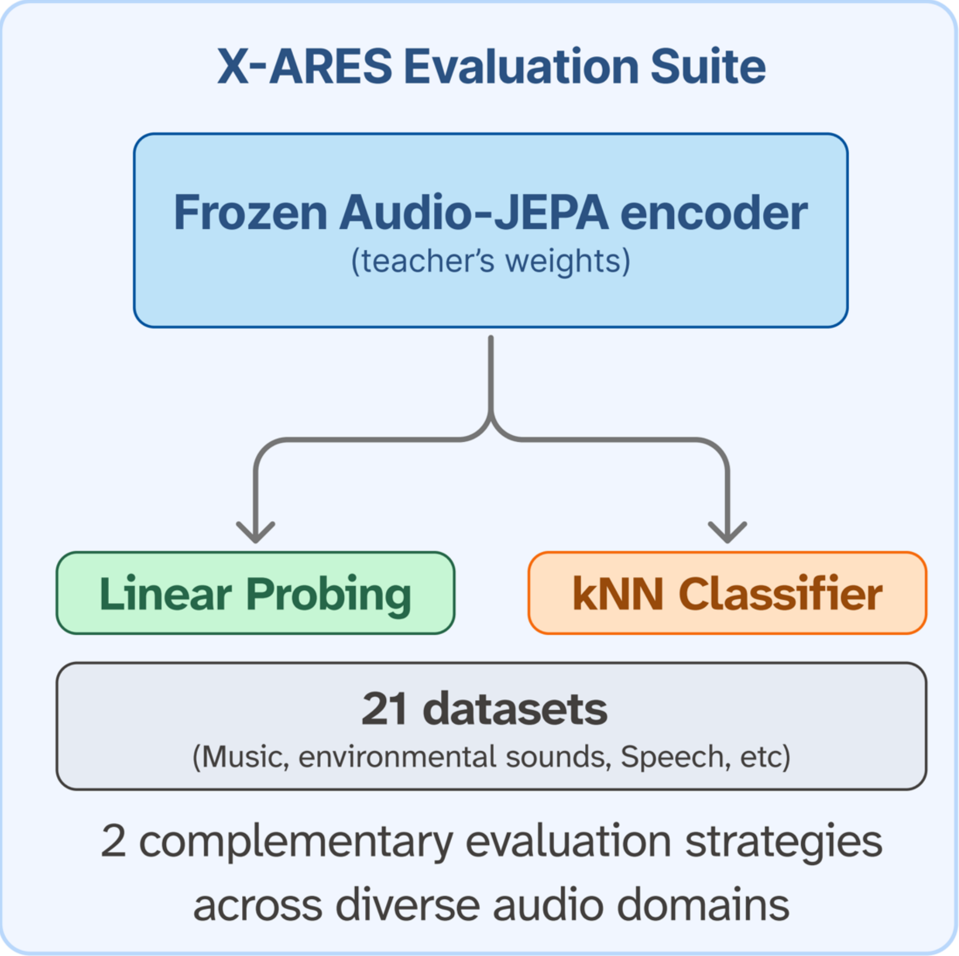
1. **Context Encoder** (ViT based)
   Encodes visible patches
2. **Target Encoder** (EMA of context encoder):
   Encodes the whole audio to create targets
3. **Lightweight Predictor** (ViT based)
   Predict encodings of masked patches

- Training minimizes L2 distance in latent space



Experimental Setup

- **Pre-training data:** ~1.9M AudioSet clips (10s, 32KHz audio, 5338 hours total)
- **Input Processing:** 256 time bins and 128 band Mel-spectrograms divided into 16x16 patches or "chunks" (each spectrogram is 8x16 patches)
- **Training details:** 100k steps (~13 epochs), 256 batch size, 4 NVIDIA V100 GPUs
- **Resource efficiency:** 14 hours total training time (vs. days for wav2vec2)



Key Findings and Results

- Strong kNN results on **music and environmental sound** recognition tasks (ESC-50, FMA-small, GTZAN)
- Competitive with or superior to wav2vec2/ data2vec on these tasks using **only 1/5 of the training data**
- Performance **lags behind on speech tasks** (speaker verification, keyword spotting), probably due to the poor temporal resolution
- **Random masking** outperforms block masking for audio representation learning

Performance on key datasets (kNN)

Dataset	Audio-JEPA	Wav2Vec2	Data2Vec
ESC-50 (environmental)	**14.0%**	8.1%	4.0%
FMA-Small (music)	**44.9%**	25.1%	10.6%
GTZAN (music)	**45.2%**	30.3%	10.8%
Speech Commands	4.4%	20.8%	**85.2%**

Training Resources Comparison

5,338 hours
Audio-JEPA
(100k steps)

VS

~27,000 hours
Wav2Vec2/Data2vec
(400k steps)

Limitations, Insights, and Future Works

Current limitations

- Weakness on fine-grained speech discrimination
- Large patch size (625ms in time dimension) limits temporal precision
- Linear probe performance lags behind kNN results due to the non-linear nature of the embedding space created by JEPA[1]
- Underperforms on tasks requiring precise localization

Future directions

1. Attention pooling for head evaluation
2. Modernize transformer backbones
3. Systematic hyperparameter tuning
4. Targeted training for speech specific tasks

Despite being a straightforward adaptation of JEPA, Audio-JEPA demonstrates strong potential for audio representation learning while using significantly fewer resources than previous methods.

All code and pretrained models are open-sourced

Audio-JEPA

[1] A. Bardes et al., Revisiting Feature Prediction for Learning Visual Representations from Video.

Q&A

Audio-JEPA

Take a group photo

Thanks， see you next year!