Auden: From Audio Encoders to Large Audio-Language Model

Building the Infrastructure and Intelligence for Next-Generation Audio AI

Yiwen Shao

Senior Research Scientist,

Tencent AI Lab (Bellevue) &

Tencent Hunyuan Speech

yshao18@jhu.edu | yiwenshaostephen.github.io

Brief Introduction

2017-2024, Johns Hopkins, CLSP
(MS + PhD)

- Contribute to open-source ASR frameworks, including Kaldi, Espresso, PyChain (Author).
- Audio-Visual ASR on multi-speaker scenarios.

Tencent

2024-Present, Tencent, AI Lab/Hunyuan Speech
Senior Research Scientist

- Large-scale multilingual ASR.
- Audio Foundation Models & MLLM.
- Auden: Open-source framework.

Audio Encoders | Speech-LLM | Omni-LLM

Background: Evolution of Text LLM (2017-Present)

2017

The Transformer Revolution

Launch of "Attention Is All You Need." Dominated by **Encoder-Decoder** models for parallel sequence processing.

2018-19

Era of Representation

BERT (Encoder-only) set SOTA for understanding.

BART (Encoder-Decoder) unified bidirectional encoding with autoregressive decoding for robust text generation.

GPT-1/2 (Decoder-only) proved generative zero-shot potential.

2020-22

Scaling & Instruction Tuning

GPT-3 established Scaling Laws. **RLHF** (ChatGPT) aligned massive decoders with human intent.

2023+

Frontier Reasoning

GPT-4 era. Industry convergence on **Decoder-only** architecture as the standard for general-purpose reasoning.

Background: How about Audio? (2020-2023)

Speech

2020
wav2vec 2.0
Contrastive SSL on raw waveforms.

2021
HuBERT
Masked acoustic unit prediction.

2022
Whisper / WavLM
Multilingual ASR & Unified SSL.

2023
USM
Universal Speech Model. 100+ languages.

Sound

2022
BEATs
Audio SSL beyond speech.

2023
CLAP
Contrastive audio-text alignment.

Music

2023
MERT & MusicLM
Music SSL & Hierarchical generation.

Background: Audio + LLM? (2023-Present)

Understanding

2023
Qwen-Audio/SALMONN
Multitask audio understanding

2024
Qwen2-Audio
Instruction-following & QA

2025
Qwen2.5-Omni/Qwen3-Omni
More advanced audio understanding & thinker-talker S2S

Generation

2024
Seed-TTS
Large autoregressive speech generation model.

2025
CosyVoice 2/3
Multilingual zero-shot LLM-based TTS.

Unified Interaction

2024
Moshi
Speech2Speech. Simultaneous listening & speaking.

“GPT” Moment of Audio? (2026)

Unit: **Tokenizer** vs. **Embedding**?

Representation: **Discrete** vs. **Continuous**?

Training Paradigm: **Self-Supervised** vs. **Supervised**?

Arch: **Encoder-Adaptor-LLM** vs. **Native Audio-LM**?

Conversation: **Direct S2S** vs. **Pipeline (ASR → LLM → TTS)**?

Application: **Specialized Models** vs. **General Model**?

A Unified Framework: Auden!

A systematic environment for architectural comparison, evaluation, and reuse.

Auden: Audio & Multimodal Understanding Toolbox

A comprehensive framework for cutting-edge **audio and multimodal understanding**. Auden bridges traditional speech tasks and modern Large Audio-Language Models (LALMs).

Core Capabilities

Foundation Audio Tasks

E2E support for high-performance ASR, Audio Captioning, CLAP, and Speaker ID, etc.

Multimodal LLM

Seamless integration with LLM and recipes for LALM training, enabling complex reasoning over audio.

Developer-Centric Design

HuggingFace-like Auto* APIs for effortless model loading and registration.

Research Milestones

Jan 2026

More Encoder Architecture & Pretrained Models

Released pretrained zh-stream and zh-en models.
Support Whisper and WenetTransformer support.

Dec 2025

AzeroS & TagSpeech

Instruction-free Speech-LLM and unified multi-speaker ASR with attribution.

Oct 2025

TTA & Voice Encoders

General-purpose encoders for multilingual alignment and paralinguistics.

Sep 2025

Initial Completeness

Launch of core foundation models (ASR, CLAP, Captioning) and LALM support.

github.com/AudenAI/Auden

Auden: LALM Standard Training Paradigm

Representation Learning

Phase 1

Encoder

Phase 2

LALM Alignment

Phase 3

LALM SFT

Phase 4

LALM RL

Representation Learning

- Perceive and encode audio modality to extract semantically meaningful representations for LLM integration.
- Encoder-centric training may emerge as a new paradigm for building native large audio models.

Data qual: Low
Data volume: high
Compute: high
Importance: high

Starting Point: Specialized Encoders

Speech Semantics

Multilingual & Multi-dialect ASR

Translation

Speech-to-Speech Retrieval

Speech-to-Text Retrieval

Paralinguistics

Emotion, Gender, & Age Classification

Speaker Identification (SID)

Multi-speaker Diarization

Sound

Audio Classification

Audio Captioning

Audio-to-Text Retrieval

Music

Pitch, Genre, Key classification.

Music Information Retrieval (MIR)

Spatial Audio

Direction of Arrival (DOA) Estimation

Single & Multi-channel Spatial Retrieval

Spatial-to-Text Retrieval

Comprehensive Evaluation Protocols

1. Linear Probing

Representation Separability

Speech: Speaker ID, Emotion, Age, Gender, Lang/Dialect

Sound: Event Classification, Music Tagging/Genre

Spatial: Direction & Distance Prediction

2. Zero-shot Task Transfer

Task-level Generalization

Speech: Verification, Diarization, Keyword Spotting, Lang Recognition

Sound: Event Detection, Music Classification

3. Zero-shot Classification/ Audio-Language Retrieval

Cross-modal Alignment

Cross-modal: Audio ↔ Text, Speech ↔ Text, Sound ↔ Caption

Class-name Mapping: Emotion, Age, Gender, Music Genre/Instrument

4. Audio-Language Generation

Speech-to-Text Generation

LLM-Frozen Instruction Tuning: MMAU, VoiceBench, Reasoning

Emerging Encoder/Embedding Evaluation (2026)

XARES-LLM Challenge

2025.12 – 2026.3

Interspeech 2026 Audio Encoder Capability

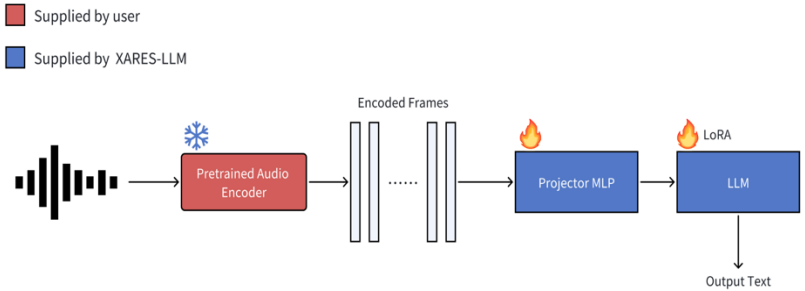
Generative Framework

Evaluates encoders as front-end modules for LALMs via end-to-end generative assessment.

Track A : 15+ classification tasks (Speaker ID, Emotion, Music Genre).

Track B : Understanding tasks (ASR, Sound/Music Captioning).

> *"A unified system that trains a typical LALM using the user's audio encoder for automated benchmarking."*



Domain	Dataset	Task Type	Metric	#
Speech	Speech Commands	Keyword spotting	Acc	30
	LibriCount	Speaker counting	Acc	11
	VoxLingua33	Language identification	Acc	33
	VoxCeleb1-Binary	Binary speaker identification	Acc	2
	ASVspoof2015	Spoofing detection	Acc	2
	Fluent Speech Commands	Intent classification	Acc	31
	VocalSound	Non-speech sounds	Acc	6
	CREMA-D	Emotion recognition	Acc	5
Sound	ESC-50	Environment classification	Acc	50
	FSD50k	Sound event detection	mAP	200
	UrbanSound 8k	Urban sound classification	Acc	10
	FSD18-Kaggle	Sound event detection	mAP	41
Music	GTZAN Genre	Genre classification	Acc	10
	NSynth-Instruments	Instruments Classification	Acc	11
	Free Music Archive Small	Music genre classification	Acc	8

Emerging Encoder/Embedding Evaluation (2026)

MAEB Benchmark

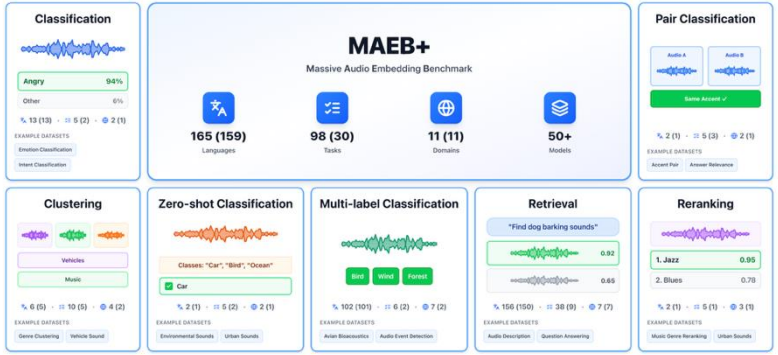
2026. 2

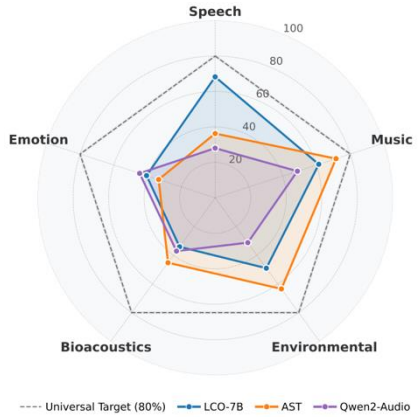
Massive Audio Embedding Benchmark

(arXiv:2602.16008)

Unified Ecosystem

Large-scale benchmark covering 30 tasks across speech, music, and environment in 100+ languages. Integrated into MTEB, one framework for unified evaluation across text, image, and audio.





> *" No single model dominates: large audio-language models show the most promise overall, but every model family has clear blind spots."*

First Glimpse: Does Encoder matter? (2024.10 – 2025.4)

"Advancing Multi-talker ASR Performance with Large Language Models"

Mohan Shi , Zengrui Jin, Yaoxun Xu, Yong Xu , Shi-Xiong Zhang, Kun Wei, **Yiwen Shao**, Chunlei Zhang, Dong Yu

SLT 2024

Table 1. Overall performance comparison of various approaches on LibriMix. Sys. {1-3} are the experimental results from ESPnet[1], Sys. {4-5} are the results of AED-based models, and Sys. {6-8} are the results of the LLM-based models.

Sys.	type	Speech Encoder	WER (%) ↓ dev	test
1	ESPnet[1] Baseline	Whisper small	26.0	25.0
2		Conformer	24.7	23.3
3		+ WavLM Large upstream	19.4	17.1
4	AED	WavLM Base+	18.9	17.7
5		WavLM Large	10.6	9.2
6	LLM	WavLM Base+	17.6	15.9
7		WavLM Large	11.4	10.2
8		+ LibriMix Fine-tuning	**10.3**	**9.0**

- *WavLM Large >> WaveLM base+ > Whisper. WavLM was pretrained on synthetic multi-talker speech.*
- ***Key findings:** What contributes most to the Multi-talker ASR performance is Encoder instead of Decoder.*

"Efficient Scaling for LLM-based ASR"

ASRU 2025

Bingshen Mu , **Yiwen Shao** , Kun Wei , Dong Yu , Lei Xie

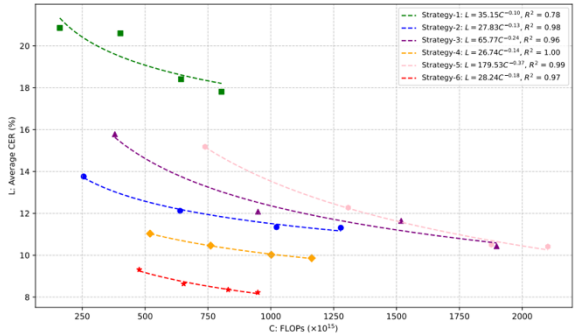


Fig. 5. Scaling law curves of multiple training strategies. Each data point represents the final converged CER and corresponding FLOPs for each strategy at a given data scale.

- *We do a thorough comparison over a series of LLM-ASR training paradigms.*
- ***Key findings**: even without additional data for encoder pretraining, it is still always beneficial to train a better encoder at the beginning to make the model lead in the scaling curve.*

Semantic Encoder: TTA (2025.9)

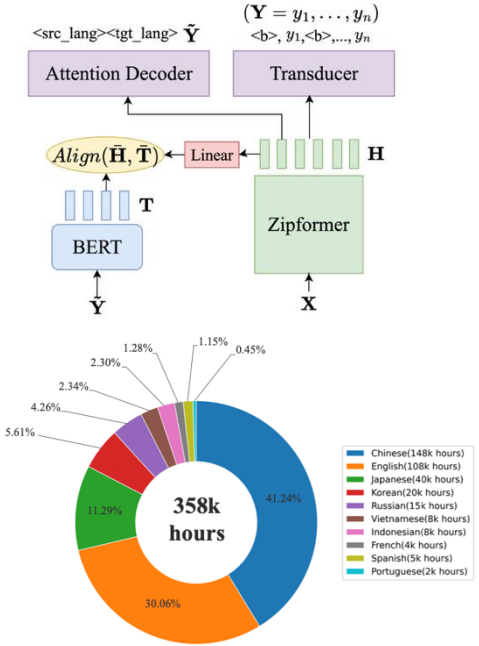
“TTA: Transcribe, Translate and Alignment for Cross-lingual Speech Representation”

Wei Liu*, Jiahong Li*, **Yiwen Shao**, Dong Yu ICASSP 2026

TTA: Transcribe–Translate–Alignment — Speech Semantic Foundation Model
- Lightweight speech semantic foundation model (<250M parameters)
- Designed for unified **MASR + Speech Translation (ST)** learning
- Trained on **358k hours** of multilingual speech data

Model Architecture
- **Zipformer architecture** for efficient semantic modeling
- Optimized to capture rich speech semantics from MASR and ST tasks
- Contrastive alignment loss bridges speech features and text embeddings

Performance
- Consistently outperforms **Whisper-Medium** on MASR and ST benchmarks
- Surpasses **Whisper-Large** on several in-domain tasks
- Strong gains in speech–text retrieval
- TTA encoder shows **stronger semantic representations than Whisper encoders** for LLM integration



Semantic Encoder: TTA (2025.9)

Datasets	Metric (%)	Whisper Medium	Whisper Large-v2	Whisper Large-v3	ZT (asr)	ZT-AED (asr)	ZT-AED	TTA
#Params		762M	1541M	1542M	199M	246M	246M	247M
aishell 1|2 **wenet** net|meeting	CER↓	6.74|6.23 11.00|22.68	5.90|5.24 9.47|22.77	5.33|4.76 9.00|15.68	1.89|3.14 6.91|6.08	1.82|3.07 6.89|6.18	1.80|3.03 6.96|5.94	1.85|3.09 7.06|6.44
librispeech clean|other **gigaspeech** **AMI**	WER↓	2.88|6.08 15.51 16.77	2.64|5.14 15.75 17.07	2.01|3.89 14.53 15.98	1.58|3.62 14.85 11.11	1.54|3.59 14.76 10.85	1.56|3.76 14.99 10.76	1.58|3.85 14.97 11.06
commonvoice **MLS** **voxpopuli** **fleurs**	WER avg↓	11.86 7.27 12.08 6.62	9.70 5.65 11.90 5.20	8.30 4.48 13.78 4.51	6.92 5.82 11.12 6.35	6.70 5.71 10.78 6.18	6.69 5.72 10.88 6.17	6.76 5.74 10.87 6.19
covostv2	BLEU↑	35.12	38.80	37.60	-	-	34.72	35.28

- With only 250M parameters, shows **better ASR performance** than 1.5B Whisper-large on most multilingual (13 languages) benchmarks.
- As we only have limited ST data from CoVoSTv2 and Europarl-ST, we utilize Qwen2.5-Instruct-7B to generate translation on ASR data and achieve a **comparable X->En translation** to Whisper-Medium.
- The **alignment head (speech-text contrastive loss)** provides the model with a strong **speech-to-text** and **speech-to-speech** retrieval ability

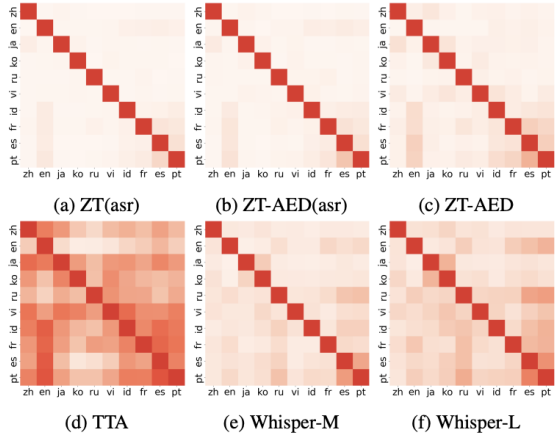
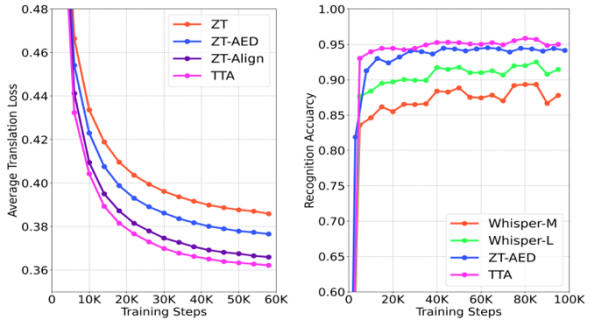


Fig. 3: Heatmaps of speech-to-speech retrieval accuracy across 10 languages. Inside each subfigure, each cell represent the retrieval accuracy from language X to language Y. The deeper color indicates higher retrieval accuracy and better cross-lingual alignment.

Semantic Encoder: TTA (2025.9)

Encoder-LLM Integration for LLM-ASR:

- TTA shows **faster convergence** than all other encoders, including Whisper-Medium, Whisper-Large, Zipformer-AED
- A **significant lower WER** is achieved with TTA as encoder.



(a) Speech translation probing. (b) Training accuracy for ASR-LLM.

Fig. 4: Two linear probing experiments. (a) compares averaged validation losses of the speech translation probing task with different encoders on CoVoSTv2. (b) shows the recognition accuracy curves, comparing various encoders in the ASR-LLM training using Aishell2 and Librispeech.

Table 2: Recognition performance comparison of different encoders in ASR-LLM. CER (%) for Aishell and WER (%) for Librispeech.

	Whisper-M	Whisper-L	ZT-AED	TTA
Aishell	5.47	4.87	2.92	**1.92**
Librispeech	4.66	3.64	2.30	**1.95**

Paralinguistic Encoder: Auden-Voice (2025.9)

“Auden-Voice: General-Purpose Voice Encoder for Speech and Language Understanding”

Mingyue Huo, Wei-Cheng Tseng, **Yiwen Shao**, Hao Zhang, Dong Yu

ICASSP 2026

Data Collection

- Collect over 8000h of paralinguistic data, covering **speaker identification, emotion, age and QA**.

Experimental Setup

- Zipformer (170M) architecture for efficient modeling.
- A comprehensive evaluation protocol is set, including **linear probing** on SID, Age, Gender, Emotion recognition, **zero-shot speaker task, speech-text retrieval, zero-shot classification** and **LLM instruction Tuning**.

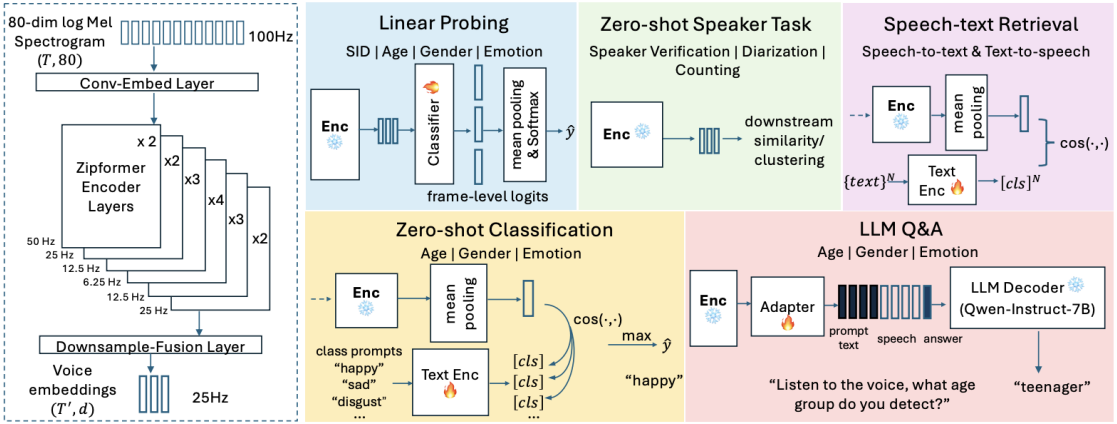


Table 1. Datasets used for training and evaluation. LP = Linear Probing; ZS(C) = Zero-shot (Classification)

Training			
Task	Train Set(s)	#Samples	Hours
SID	VoxCeleb2[21]	974k	2026
Paraling	CREMA-D[22], RAVDESS[23] IEMOCAP[24], TESS[25]	18.3k	20
CLAP	ParaSpeechCaps [26] (EARS[27] EXPRESSO[28], Vox1&2, Emilia[29])	base 111k[1] scaled 925k	2700
LLM-QA	CommonVoice[30], IEMOCAP MELD[31], Vox2, CREMA-D TESS, RAVDESS	1.76 M	3250

Evaluation			
Task	Eval Set(s)	Eval	#Samples
SID	VoxCeleb2	LP	108k
Age	CREMA-D	LP/ZSC	706
Gender	CREMA-D, RAVDESS	LP/ZSC	706, 136
Emotion	CREMA-D, RAVDESS	LP/ZSC	706, 136
SV	VoxCeleb1-o[32]	ZS	37k
SD/SC	VoxConverse[33]	ZS	232
Retrieval	ParaSpeechCaps-base	Retrieval	568
LLM-QA	AIR-Bench[34][2](MELD IEMOCAP,[3] CommonVoice)	QA	5k

[1] ParaSpeechCaps-train overlaps Vox1&2 test; filtered in our CLAP training.
[2] AIR-Bench overlaps CommonVoice-train/dev; filtered in our LLM-QA training.
[3] IEMOCAP has no fixed split; minor leakage possible at init stage.

Paralinguistic Encoder: Auden-Voice (2025.9)

Encoder Training Insights

- **ASR Pretraining**: provides strong basis for voice representation learning.
- **Multitask Training**: beneficial for balanced performance across voice related task.
- **Speaker Identification (SID):** contributes the most.
- **CLAP**: benefits to retrieval task but not necessary for voice representation capacity.

Voice Encoder Matters

- **Cascaded model has limited performance** over paralinguistic understanding QA.
- Semantic Encoder (Whisper) not good enough.
- Auden-Voice **boost paralinguistic understanding** in LALM.

Table 3. Speech–text retrieval and zero-shot classification results before and after CLAP fine-tuning. Absolute increases and drops are shown relative to pre-CLAP. The bottom block additionally reports LP and speaker-ZS of CLAP-fine-tuned encoders.

Enc.#	Init.	Sup.	Speech-text Retrieval							Zero-shot Classification (ZSC)					
			Speech-to-Text			Text-to-Speech			Retrieval	Age	Gender	Gender	Emotion	Emotion	ZSC
			R@1	R@5	R@10	R@1	R@5	R@10	Avg↑	CREMA	CREMA	RAVDESS	CREMA	RAVDESS	Avg↑
1.1	–	ASR	35.9	68.4	77.8	32.6	64.5	75.3	59.1	21.3	76.8	67.7	29.2	20.6	43.1
1.2	ASR	SID	62.7	96.1	98.1	61.0	96.3	98.7	85.5	22.0	96.5	**100**	27.8	25.7	54.4
1.3	ASR	Paraling	46.9	79.3	86.3	46.8	79.5	86.6	70.9	45.0	94.3	91.2	36.4	45.6	**62.5**
1.4	ASR	multi-task	63.3	95.2	97.4	61.7	96.0	98.0	85.3	11.0	**96.6**	94.9	**42.1**	**49.3**	58.8
2.1	1.1	CLAP	71.9 +36.0	97.8 +29.4	99.1 +21.3	**74.2** +41.6	98.8 +34.3	99.5 +24.2	90.2 +31.1	38.5 +17.2	82.2 +5.4	95.6 +27.9	30.2 +1.0	32.4 +11.8	53.7 +10.6
2.2	1.2	CLAP	**72.3** +9.6	**98.4** +2.3	**99.4** +1.3	73.9 +12.9	**98.9** +2.6	**99.6** +0.9	**90.4** +4.9	23.8 +1.8	94.8 -1.7	98.5 -1.5	22.4 -5.4	26.5 +0.8	53.2 -1.2
2.3	1.3	CLAP	70.4 +23.5	97.6 +18.3	99.0 +12.7	72.1 +25.3	98.1 +18.6	99.3 +12.7	89.4 +18.5	25.5 -19.5	64.0 -30.3	99.3 +8.1	35.0 -1.4	40.4 -5.2	52.8 -9.7
2.4	1.4	CLAP	71.3 +8.0	98.1 +2.9	99.3 +1.9	73.2 +11.5	98.6 +2.6	99.5 +1.5	90.0 +4.7	37.8 +26.8	89.2 -7.4	91.2 -3.7	28.5 -13.6	28.7 -20.6	55.1 -3.7
Whisper-medium			43.9	77.1	84.1	40.1	75.3	83.5	67.3	42.0	86.3	69.9	28.2	29.4	51.2
wav2vec2.0-base			25.8	55.0	65.8	27.5	56.2	66.2	49.4	**46.6**	66.7	76.5	22.7	17.6	46.0
emotion2vec			29.0	60.2	69.9	26.5	55.7	66.1	51.2	46.0	74.7	75.0	23.1	17.6	47.3
Wespeaker			47.5	81.1	87.5	46.7	82.7	88.8	72.4	28.6	64.3	98.5	24.4	17.7	46.7

Enc.#	Init.	Sup.	Linear probing (LP)							Speaker Zero-shot				
			SID	Age	Gender	Gender	Emotion	Emotion	LP	SV	SD	SD	Count	ZS
			Vox2	CREMA	CREMA	RAVDESS	CREMA	RAVDESS	Avg↑	EER↓	DER↓	Conf↓	MAE↓	Avg↑
2.1	1.1	CLAP	83.0 +61.4	77.9 +10.2	99.6 +8.2	100 +1.5	72.7 +10.5	89.0 +13.3	87.0 +17.5	10.2 +35.5	35.5 +15.6	28.0 +15.6	3.8 +0.9	47.5 +45.2
2.2	1.2	CLAP	95.9 -3.1	82.6 -2.5	99.0 -0.2	100 +0.0	70.7 -3.1	83.1 -0.7	88.6 -1.6	9.8 -7.5	26.5 -12.3	19.0 -12.2	2.8 -1.0	66.9 -27.5
2.3	1.3	CLAP	88.4 +30.7	89.8 -8.1	99.9 -0.1	100 +0.0	77.5 -2.3	89.0 -5.1	90.8 +2.5	10.6 +26.5	37.8 +12.2	30.4 +12.1	3.9 +0.6	43.8 +34.0
2.4	1.4	CLAP	90.2 -5.1	86.5 -7.4	99.6 -0.1	100 +0.0	76.8 -7.2	93.4 +3.7	91.1 -2.7	10.9 -7.1	31.6 -14.6	24.1 -14.6	3.3 -1.7	56.1 -35.5

Table 4. LLM-QA acc (%) on AIR-Bench foundation benchmark. Top three rows freeze the voice encoder and the LLM (Qwen2.5-7B-Inst [43]), training only a lightweight adapter.

System	Emotion		Gender		Age
	MELD*	IEMO	MELD	CV	CV
Enc 1.4: multi-task	27.2	**84.7**	**81.6**	**93.2**	58.3
Enc 2.4: multi-task+CLAP	22.3	43.6	76.2	87.3	**66.2**
Whisper + Qwen-Inst-7B	42.2	27.5	47.6	52.2	65.3
Qwen-Audio (end-to-end)	43.2		67.2		36.0
Whisper → GPT-4 (cascade)	59.5		21.9		41.1

* Performance shows high variance across encoders despite similar results on other tasks. This may due to the lack of diversity in LLM-QA training data.

Audio Encoder: Audio-Language Pretraining (2025.10)

“Revisiting Audio-language Pretraining for Learning General-purpose Audio Representation”

Wei-Cheng Tseng, Xuanru Zhou, Mingyue Huo, **Yiwen Shao**, Hao Zhang, Dong Yu

ACL 2026 Submission

Contrastive vs. Captioning

Compare 2 prevalent audio-language pretraining paradigms that utilize weak supervision and show good potential for scaling.

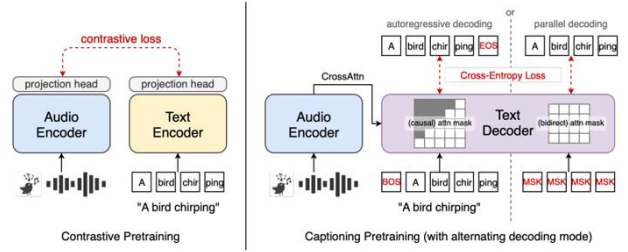


Figure 1: Audio-language pretraining objective studied in this work: contrastive and captioning.

CaptionStew

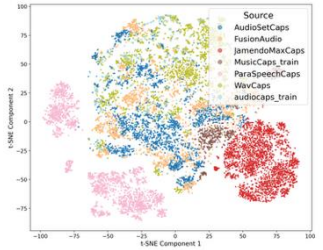
Collect the largest open-source audio-text paired data with 10.7M, covering:

- Multiple aspects: speech, music, sound
- Both human annotated and LLM-augmented

Table 1: Comparison of publicly available audio caption datasets. The number of audio-text pairs (#pair) and number of unique words (#vocab) are shown here.

Audio Caption Dataset	#pair	#vocab
Human-annotated		
AudioCaps Kim et al. (2019)	46K	4,844
Clotho Drossos et al. (2020)	5K	4,366
MusicCaps Agostinelli et al. (2023)	5K	3,730
LLM-augmented		
WavCaps Mei et al. (2024)	403K	18,372
AudioSetCaps Bai et al. (2025)	1.9M	21,783
FusionAudio Chen et al. (2025)	1.2M	18,403
AutoACD Sun et al.	1.5M	20,491
CaptionStew (Ours)	10.7M	56,586

Figure 3: t-SNE visualization of sentence embedding of captions grouped by source.



Audio Encoder: Audio-Language Pretraining (2025.10)

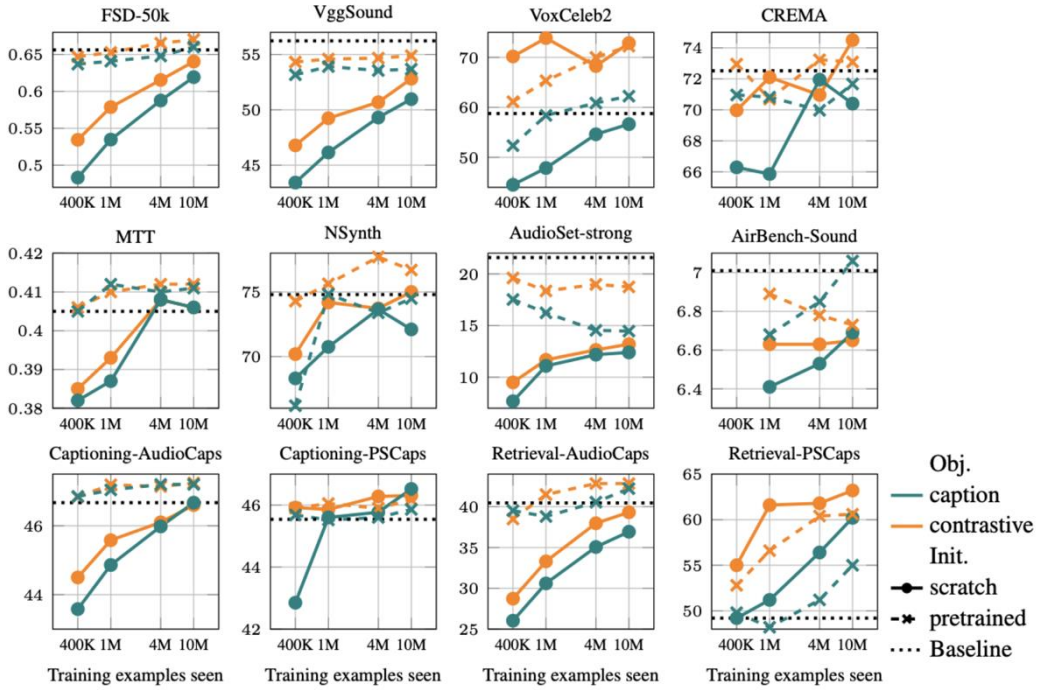


Figure 2: Data scaling behavior of contrastive vs. captioning objectives across representative tasks.

Contrastive vs. Captioning

The two pretraining paradigms exhibit complementary strengths across evaluation protocols.

- Contrastive wins for linear probing
- Captioning shows better scaling behavior especially when language is involved.
- Both benefits from supervised initialization

Across Domains Performance

Competitive performance is achieved across all domains, covering speech semantics, paralinguistics, music, audio.

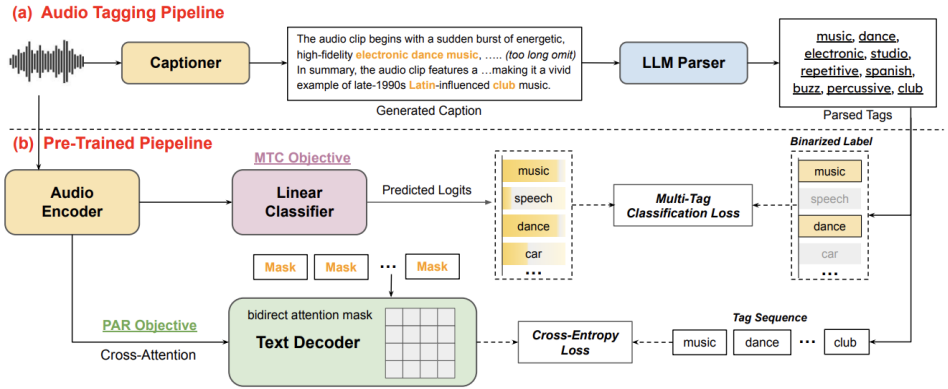
Audio Encoder: UTS (2025.11)

“Unlocking Strong Supervision: A Data-Centric Study of General-Purpose Audio Pre-Training Methods”

Xuanru Zhou, **Yiwen Shao**, Wei-Cheng Tseng, Dong Yu

CVPR 2026

- Proposes a **data-centric strong-supervision** pipeline for general-purpose audio pre-training.
- Introduces **Unified Tag System (UTS)** to unify **speech, music,** and environmental **sound** labels.
- Studies multiple objectives (**tag classification, captioning, contrastive, multi-task**) on strong supervision data.
- Shows **data quality + label coverage** > **data scale** for representation learning. Strong supervision significantly improves generalization across diverse audio tasks.



Concurrent Work of Large Scale Audio Encoder Pretraining

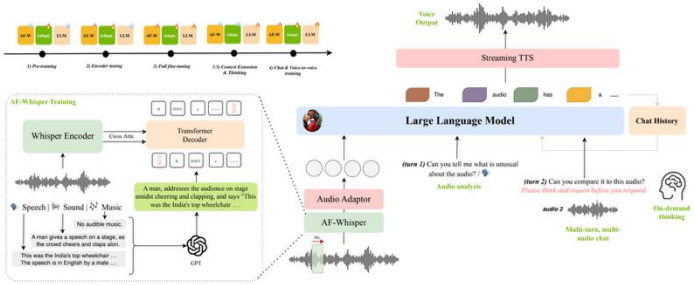
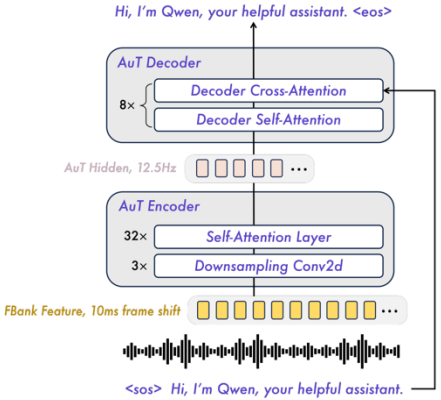


Figure 2: Overview of Audio Flamingo 3, AF-Whisper training, and five-stage curriculum training.

AF-Whisper (2025.7)

In **Audio-Flamingo3 (AF3),** 13M pairs of audio-text data covering speech, sound, music is used to finetune a whisper-large-v3 encoder with cross-attn.



Qwen3 AuT (2025.10)

In Qwen3-Omni, an **Audio Transformer (AuT)** is trained with attn-encoder-decoder on 20 million hours of audio caption data.

Auden: LALM Standard Training Paradigm

Phase 1

Encoder

CRITICAL ALIGNMENT

Phase 2

LALM Alignment

LALM CORE

Phase 3

LALM SFT

Phase 4

LALM RL

Modality Alignment

Bridges the modality gap by mapping audio features into the LLM space via projectors, enabling comprehension.

Data qual: mid
Data volume: high
Compute: high
Importance: high

Instruction Following

Trains the model on diverse audio-text pairs, enabling complex reasoning and paralinguistic analysis.

Data qual: high
Data volume: low
Compute: low
Importance: high

Problem

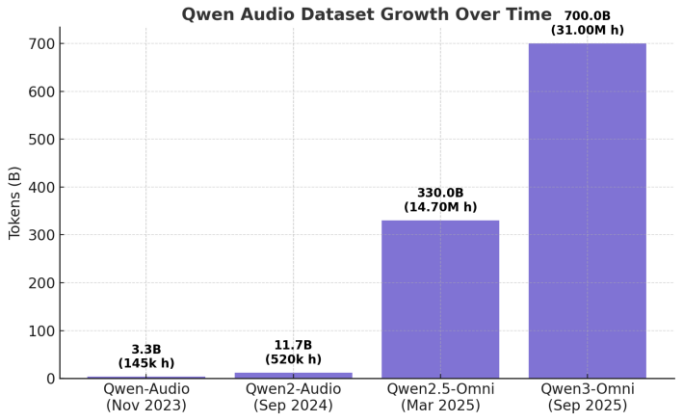
Task Overfitting

- Overfit to training format: multiple choice vs open question.
- Overfit to train/evaluation tasks:
  - Easily get good results on target (seen) tasks.
  - **Poor generalization** to new tasks/benchmarks

Data Limitation

- QA pairs generation: partially labeled data -> LLM -> audio-text pair
- Distilled data from other API: more likely to introduce hallucination

Solution?
Keep increasing the data!



Inspiration from VoiceBench

- VoiceBench: A benchmark that synthesizes speech-based questions from text benchmarks.
- Expectation: Equal intelligence regardless of input modality (speech vs. text).

Model		AlpacaEval (GPT)	CommonEval (GPT)	SD-QA (Panda/GPT)		MMSU (Acc.)	OpenBookQA (Acc)	IFEval (P./I. Acc)		AdvBench (Refusal Rate)	Overall
Naive	T.	4.69	4.38	77.76	75.41	66.23	72.53	73.91	79.52	96.54	81.43
	S.	4.53	4.04	72.33	68.54	62.43	81.54	65.37	73.70	98.08	79.06
Naive-4o	T.	4.83	4.63	63.47	94.39	85.17	94.29	77.68	83.43	98.27	89.49
	S.	4.80	4.47	60.58	90.96	81.69	92.97	73.19	79.82	98.27	87.23
DiVA	T.	4.68	4.29	78.30	74.50	63.31	76.70	68.70	76.31	99.23	81.08
	S.	3.67	3.54	62.39	51.72	25.76	25.49	34.93	43.38	98.27	55.70
LLaMA-Omni	T.	4.39	4.32	55.33	60.40	59.01	79.34	45.38	56.53	98.46	74.26
	S.	3.70	3.46	40.14	39.24	25.93	27.47	10.15	19.58	11.35	37.51
Mini-Omni	T.	2.34	2.55	26.04	7.23	26.74	30.55	13.04	22.89	86.35	39.43
	S.	1.95	2.02	23.69	4.16	24.69	26.59	8.99	18.17	37.12	27.90
Mini-Omni2	T.	2.65	2.86	13.02	9.76	27.13	32.09	10.15	17.87	92.88	41.10
	S.	2.32	2.18	11.03	7.59	24.27	26.59	7.25	15.86	57.50	31.32
Qwen2-Audio	T.	4.11	3.77	61.66	40.69	45.02	67.91	28.70	38.06	96.73	64.55
	S.	3.74	3.43	41.77	29.66	35.72	49.45	20.73	31.93	96.73	55.35
VITA	T.	4.00	3.88	72.69	76.13	64.54	83.08	48.99	57.53	95.19	75.43
	S.	3.38	2.15	31.28	24.59	25.70	29.01	18.12	27.51	26.73	34.68
Moshi	S.	2.01	1.60	15.01	16.27	24.04	26.15	6.38	13.76	44.23	27.47
GPT-4o-Audio	S.	4.78	4.49	61.12	89.87	80.25	89.23	74.86	77.17	98.65	86.42

Table 3: The performance of various voice assistants on VoiceBench. The T. and S. rows refer to the model performance with text-form and speech-form instructions respectively. GPT-4o-Audio and Moshi only allows speech-form instructions now. We report the performance on SD-QA United States accent above.

Problem & Observation

- Big gap observed on semantically equivalent Text and Speech input.
- Suddenly all end-to-end speech-LLM show good results on this bench after it is released.
- Only naïve cascaded system and top commercial APIs (e.g. GPT-4o, Gemini-2.5 pro) show robustness when new benchmark comes out.

Thoughts

If the underlying LLM already shown superior performance on text input, why not **let LLM treat audio input as text**?

Speech-LLM: AZeroS (2026.1)

“AZeroS: Extending LLM to Speech with Self-Generated Instruction-Free Tuning”

Yiwen Shao*+, Wei Liu*, Jiahong Li*, Tianzi Wang, Kun Wei, Meng Yu, Dong Yu

Technical Report 2026

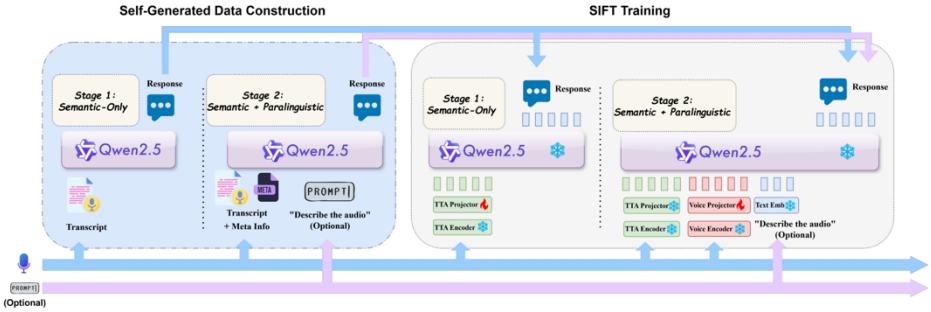


Figure 1: Overview of the proposed AZEROS framework built upon a frozen Qwen2.5-7B-Instruct backbone. The system couples a **Self-Generated Data Construction** pipeline (left) with a two-stage **Self-Generated Instruction-Free Tuning (SIFT)** training procedure (right). In **Stage 1 (Semantic)**, the LLM generates targets from speech transcripts to train a projector on the semantic *TTA* Liu et al. (2025) encoder. In **Stage 2 (Semantic + Paralinguistic)**, we augment inputs with metadata (e.g., emotion, gender, age); the resulting rich responses train a second projector on a paralinguistic *Auden-Voice* Huo et al. (2025) encoder. Together, this progressive SIFT paradigm extends the model’s capability from purely semantic to joint semantic–paralinguistic understanding without relying on task-specific instructions.

Method: SIFT

- Define the **text equivalence** of the audio, e.g., speech → transcript
- Input the text equivalence to the LLM to get its **response** (instruction not required).
- Use **<audio, response>** pair to train the speechLLM.

Results

- Trained on only 25k hours of public ASR and 3k paralinguistic data. **Nothing relevant to any QA.**
- Dual TTA-Voice Encoder enable both **semantic and paralinguistic perception** ability.
- Achieve SOTA performance on **VoiceBench** and **AIR-Bench (speech).**
- Approximate its theoretical upper bound of using text as input.

Speech-LLM: AZeroS (2026.1)

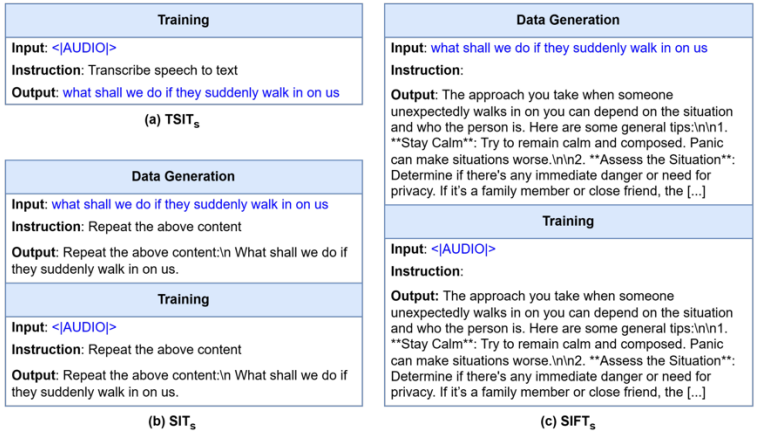


Figure 2: Illustration of semantic-only instruction-tuning configurations: (a) **TSIT$_S$**, (b) **SIT$_S$**, and (c) **SIFT$_S$**. Examples are selected from an utterance in the IEMOCAP dataset Busso et al. (2008), with the transcript *"what shall we do if they suddenly walk in on us"*. (*[...]* indicates truncated text. Detailed examples are in Appendix A.)

SIFT vs. IT

- Using only ASR data.
- Instruction-tuning (IT) **<speech, instruction, transcript>** makes the model do ASR only task.
- SIFT **<speech, response>** makes the model learn general speech-text alignment and it can generalizes to any unseen voice QA.

Table 3: Preliminary results on *VoiceBench* comparing different instruction-tuning configurations (see Figure 2) for **semantic-only** alignment, trained on the **Data-SP** subset.

Method	AlpacaEval (GPT)	CommonEval (GPT)	WildVoice (GPT)	SD-QA (Panda/GPT)	BBH (Acc.)	AdvBench (Refusal)	IFEval (P./I. Acc.)	OBQA (Acc.)	MMSU (Acc.)	Overall
TSIT$_S$	1.12	1.34	1.10	4.52	47.80	3.46	24.13	24.62	24.59	22.26
SIT$_S$	3.98	3.74	3.38	45.75	51.00	95.96	39.09	34.07	33.12	57.89
SIFT$_S$	4.32	4.14	3.74	55.70	55.60	98.27	54.40	59.34	51.01	**68.70**

Speech-LLM: AZeroS (2026.1)

Data Generation	Data Generation
Input: What shall we do if they suddenly walk in on us (male, happy)	**Input**: What shall we do if they suddenly walk in on us (male, happy)
Instruction: Describe all the information you can hear.	**Instruction**:
Output: Based on the provided information:\n\n- The audio is an audio tag with metadata and text content.\n- The gender of the speaker is specified as male.\n- The emotion of the speaker is described as happy.\n- The text content of what the male speaker is saying is: \"what shall we do if they suddenly walk in on us."\n\nFrom this [...]	**Output**: It sounds like you're looking for advice or a response to the scenario described. Here's a suggestion based on the context provided:\n\n **Male Voice (Happy Emotion):**}\n "Well, if they suddenly walk in on us, we could just pretend to be doing something else. Maybe we could be having a serious discussion or checking [...]
Training	**Training**
Input: <|AUDIO|>	**Input**: <|AUDIO|>
Instruction: Describe all the information you can hear.	**Instruction**:
Output: Based on the provided information:\n\n- The audio is an audio tag with metadata and text content.\n- The gender of the speaker is specified as male.\n- The emotion of the speaker is described as happy.\n- The text content of what the male speaker is saying is: \"what shall we do if they suddenly walk in on us."\n\nFrom this [...]	**Output:** It sounds like you're looking for advice or a response to the scenario described. Here's a suggestion based on the context provided:\n\n **Male Voice (Happy Emotion):**}\n "Well, if they suddenly walk in on us, we could just pretend to be doing something else. Maybe we could be having a serious discussion or checking [...]

(a) SIT_{sp} (b) $SIFT_{sp}$

Method	VoiceBench										AIR-Bench (Speech)							
	Alpaca Eval	Comm Eval	Wild Voice	SD-QA	BBH	Adv Bench	IF Eval	OBQA	MMSU	Overall	Gender	Emotion	Age	LID	Entity	Intent	Avg	Chat
Stage 1: semantic-only (on Data-S)																		
$SIFT_s$	4.44	4.22	3.95	59.22	56.10	99.23	62.63	73.19	60.08	**73.63**	30.47	46.40	34.60	76.20	72.40	81.80	56.98	7.22
Stage 2: semantic + paralinguistic (on Data-SP)																		
SIT_{sp}	4.38	4.14	3.86	57.60	55.40	98.65	60.01	71.87	58.85	72.22	74.85	73.10	57.20	85.30	72.90	87.20	75.09	8.19
$SIFT_{sp}$	4.41	4.08	3.81	59.13	55.80	99.04	61.10	70.33	59.01	72.27	83.18	74.60	64.00	85.50	72.70	86.10	**77.68**	8.21
$SIFT_s$ + $SIFT_{sp}$	4.44	4.18	3.91	60.22	56.30	98.65	61.29	72.09	59.01	**73.13**	86.75	71.45	61.30	84.80	73.60	85.60	77.25	**8.28**

Semantics -> Paralinguistics

- Use **"transcript + meta"** as input to LLM for response generation.
- Both **SIT <speech, instruction, response>** and **SIFT <speech, response>** enable model with semantic + paralinguistic understanding.
- SIT can be used as a **complementary** method when response cannot reflect the info in the input.

Speech-LLM: AZeroS (2026.1)

Model	VoiceBench										AIR-Bench (Speech)							
	Alpaca Eval	Comm Eval	Wild Voice	SD-QA	BBH	Adv Bench	IF Eval	OBQA	MMSU	Overall	Gender	Emotion	Age	LID	Entity	Intent	Avg	Chat
Text Only Model																		
Qwen2.5	4.66	4.55	4.62	62.03	80.00	99.04	70.14	84.84	71.57	82.69	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Qwen2.5 (TN)	4.61	4.53	4.56	63.84	56.30	98.85	66.11	74.07	64.51	77.52	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cascaded System																		
Whisper+GPT-4o	4.80	4.47	4.62	75.77	87.20	98.27	76.51	92.97	81.69	87.80	21.90	59.50	41.10	96.80	69.80	87.70	62.80	7.54
Whisper+Qwen2.5	4.64	4.33	4.21	58.50	52.85	98.27	63.99	78.24	69.00	76.05	28.36	50.80	36.40	88.00	73.60	82.70	59.98	7.34
End-to-end Speech-LLM																		
GPT-4o	4.78	4.49	4.58	75.50	84.10	98.65	76.02	89.23	80.25	86.75	*	49.10	*	76.00	61.60	85.8	*	7.53
Gemini2.5-pro	-	-	-	-	-	-	-	-	-	-	90.7	60.70	34.10	99.10	68.5	92.2	74.22	8.52
Moshi	2.01	1.60	1.30	15.64	47.40	44.23	10.12	25.93	24.04	29.51	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SALMONN	-	-	-	-	-	-	-	-	-	-	35.5	29.9	48.7	28.1	51.7	36.7	38.43	6.16
Phi-4-multimodal	3.81	3.82	3.56	39.78	61.80	100.00	45.35	65.93	42.19	64.32	-	-	-	-	-	-	-	-
GLM-4-Voice	3.97	3.42	3.18	36.98	52.80	88.08	25.92	53.41	39.75	56.48	23.91	22.95	18.70	25.40	27.90	21.10	23.33	5.53
Qwen2-Audio	3.42	3.29	2.76	31.65	53.00	99.04	26.35	48.35	36.14	53.77	64.71	48.15	23.10	77.80	87.00	84.70	64.24	7.20
DeSTA2.5	3.73	2.52	3.30	46.47	62.40	97.69	65.47	72.75	58.56	66.04	84.24	64.30	65.60	97.30	65.20	83.70	76.72	7.57
Qwen2.5-Omni	3.88	3.77	3.52	46.75	63.70	97.31	40.19	81.54	61.45	68.26	89.76	54.85	44.80	89.70	79.70	88.60	74.57	6.97
Qwen3-Omni-30B	4.74	4.54	4.58	76.90	80.40	99.30	77.80	89.70	68.10	85.49	91.11	62.20	36.90	97.70	80.40	90.70	76.50	7.85
AZEROS (ours)	4.44	4.18	3.91	60.22	56.30	98.65	61.29	72.09	59.01	**73.13**	86.75	71.45	61.30	84.80	73.60	85.60	**77.25**	**8.28**

System	VoiceBench										AIR-Bench (Speech)							
	Alpaca Eval	Comm Eval	Wild Voice	SD-QA	BBH	Adv Bench	IF Eval	OBQA	MMSU	Overall	Gender	Emotion	Age	LID	Entity	Intent	Avg	Chat
2 Stages (AZeroS)																		
TTA-Voice	4.44	4.18	3.91	60.22	56.30	98.65	61.29	72.09	59.01	**73.13**	86.75	71.45	61.30	84.80	73.60	85.60	**77.25**	**8.28**
1 Stage																		
TTA-Voice	4.47	4.08	3.80	56.60	56.40	99.04	62.39	70.11	60.64	72.46	77.16	58.35	66.00	67.70	70.70	83.10	70.50	7.49
Whisper	4.38	3.96	3.77	55.43	54.20	98.46	58.94	64.62	54.88	69.86	69.88	54.85	65.80	70.80	68.30	84.30	68.99	7.65

Results

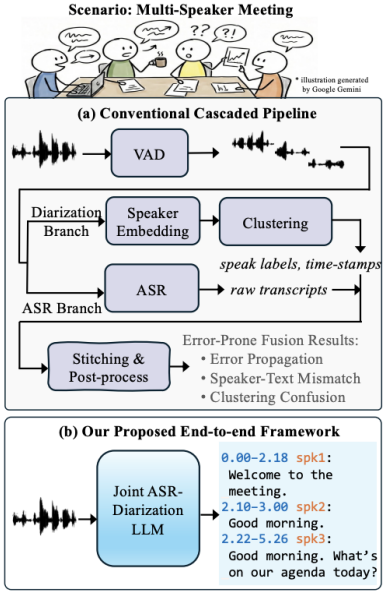
- AZeroS is close to its **text upper bound Qwen2.5-7B** on voice QA.
- AZeroS achieves **SOTA** performance against all Speech-LLM with similar size backbone.
- Ablation study on **dual TTA-Voice vs. Whisper** shows the gain from dual encoder design.
- Whisper encoder only shows competitive results, proving the **effectiveness of SIFT paradigm**.

Speech-LLM: TagSpeech (2026.1)

“TagSpeech: End-to-End Multi-Speaker ASR and Diarization with Fine-Grained Temporal Grounding”

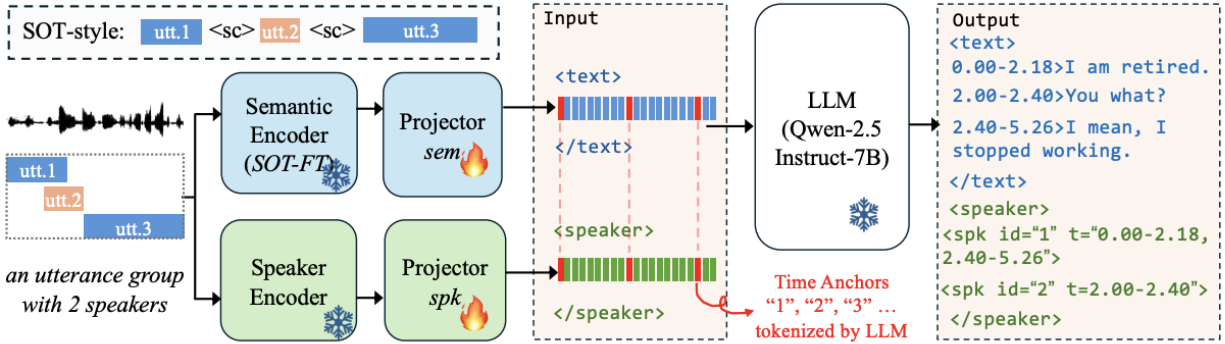
Mingyue Huo, **Yiwen Shao**, Yuheng Zhang

ACL 2026 Submission



Speaker-attributed LLM-ASR:

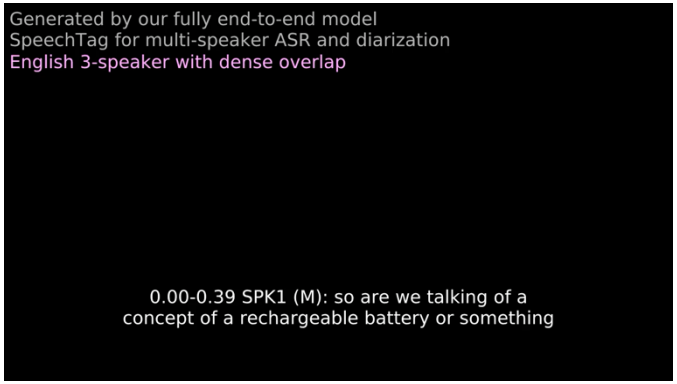
- Frozen **semantic encoder + paralinguistic encoder + Qwen2.5**, only trained on 2 lightweight projectors.
- Insert **time anchors** and **attribute tags** for better alignment in the output.
- End-to-end output with accurate **diarization, speaker attributes and transcripts**.



Speech-LLM: TagSpeech (2026.1)

Summary

- Better **diarization and WER** than SOTA end-to-end model **Gemini** and **Qwen2.5/3 Omni**.
- **Cascade baseline** (Pyannote + Whisper) still show competitive results.
- **Limitation**: only train on small dataset AMI and Alimeeting.



Model	AMI-SDM (English)					AliMeeting-Far (Mandarin)				
	Fail Rate ↓	DER ↓	cpWER ↓	gWER ↓	SCA ↑	Fail Rate ↓	DER ↓	cpCER ↓	gCER ↓	SCA ↑
End-to-End Baselines										
Gemini-2.0-flash ⋆	2.11	45.16	**36.24**	**29.86**	56.19	31.90	50.35	59.02	47.97	51.14
Qwen2.5-Omni-7B	4.07	34.71	49.86	32.25	60.06	6.40	37.42	41.23	**24.44**	71.12
Qwen3.0-Omni-30B-A3B-Instruct	2.15	39.06	38.83	33.61	59.35	**0.39**	34.60	**33.69**	27.69	70.20
Cascade Baseline										
Pyannote 3.1 + Whisper-large-v3	4.60	**23.05**	43.57	35.95	51.43	2.52	26.13	46.56	39.55	60.68
TagSpeech (Ours)	**1.27**	24.84	42.55	31.62	**70.01**	3.04	**22.13**	33.84	25.42	**81.63**

⋆ Chosen over Gemini-2.5/3.0-pro for less frequent recursive hallucination loops, see Appendix A.3.

Auden: More Encoders and MLLMs

Encoder&Tokenizer

- *AAAI 2026*, **"DualSpeechLM**: Towards Unified Speech Understanding and Generation via Dual Speech Token Modeling with Large Language Models"
- *ICASSP 2026*, **"**Exploring SSL Discrete Tokens For Multilingual Automatic Speech Recognition"
- *Arxiv 2026*, "Towards Comprehensive Semantic Speech Embeddings for Chinese Dialects"

MLLM

- *Tech Report 2026*, **"Penguin**: Small yet Strong Foundation Models for Well-rounded Multi-Modal Understanding"
- Arxiv 2026, **"JAEGER**: Joint 3D Audio-Visual Grounding and Reasoning in Simulated Physical Environments"

Upcoming

- Music Encoder
- Spatial Encoder
- Large scale general Audio Encoder

STAY TUNED FOR THE LATEST RELEASES

github.com/AudenAI/Auden **huggingface.co/AudenAI**

Summary and Future Work: From Encoders to LALMs

Encoder remains under-exploration

- Encoder plays a key role for audio foundation models and LALM.
- Specialized encoders still outperform general purpose encoder.
- Loss/training paradigm design that scales better.
- Encoder free?

LALM: Alignment or Native

- As text LLMs evolve dramatically these days, we may resort to rely more on a powerful LLM and design better modality alignment approach.
- Native MLLM is also promising. Success of interleaved training encourages more exploration.
- Unified understanding and generation ability.