Audio Representation Learning and Downstream Task Modeling: A Survey

Kaihong Wang[1], Liang He[1,2†],
[1]Department of Electronic Engineering, and Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[2]School of Computer Science and Technology, School of Intelligence Science and Technology, Xinjiang University, Urumqi 830017, China
{wkh25, heliang}@mails.tsinghua.edu.cn

Abstract

Audio representation learning has revolutionized acoustic intelligence, enabling audio foundation models to generalize and perform well across various downstream tasks. To better characterize the paradigm shift in general representation extraction and modeling, this survey presents a unified decoupled framework separating audio processing into audio-to-feature representation learning and feature-to-target downstream modeling. We systematically review modeling domains, representation forms, architectures, and pre-training paradigms, while categorizing five integration mechanisms and downstream modeling paradigms. Furthermore, we summarize representative systems and discuss key trade-offs, providing practical guidelines for designing next-generation audio foundation models.

1 Introduction

Raw audio signals entangle complex, dynamic, and diverse acoustic factors. While historical research bottlenecked representational capacity by relying on rigid hand-crafted features like MFCCs (Davis and Mermelstein, 1980; Reynolds and Rose, 1995) and Log-Mel spectrograms (Cotton and Ellis, 2011; Rakotomamonjy and Gasso, 2015) or heuristic mixtures (Scheirer and Slaney, 1997; Lu et al., 2002), data-driven deep learning (Hinton et al., 2012; Abdel-Hamid et al., 2014) and self-supervised learning (SSL) (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2023) have redefined this domain. By learning universal representations directly from unlabeled data, pre-trained models now support diverse downstream tasks, ranging from speech language modeling (Borsos et al., 2023a; Chen et al., 2024a) and low-latency dialogue (Défossez et al., 2024; Roy et al., 2026) to unified generation (Chen et al., 2025b; Tian et al., 2026d) and diffusion-based understanding (Zhou et al., 2026a,b).

Current audio representation extraction paradigm has shifted from task-specific networks to large-scale pre-trained models capturing general representations. These robust representations diversify mapping mechanisms and inspire unified architectures, such as discrete representations with autoregressive prediction for audio language modeling (Borsos et al., 2023a; Chen et al., 2024a) and low-latency voice interaction (Défossez et al., 2024; Roy et al., 2026), unified understanding and generation via joint prediction (Chen et al., 2025b; Tian et al., 2026d), and diffusion language models for audio understanding (Zhou et al., 2026a,b). However, existing task-centric surveys (Xie et al., 2025b; Yeow et al., 2025; Guo et al., 2026b; Zhu et al., 2025; Retkowski et al., 2025) or those focusing solely on autoregressive speech-language models (Yang et al., 2025; Peng et al., 2025a) neglect representation injection and downstream modeling, failing to capture this paradigm shift. To address this, we propose a decoupled perspective of representation and modeling that first delineates representation space construction and evaluation before discussing downstream adaptation strategies across tasks.

Our main contributions are threefold: (a) we establish a decoupled taxonomy that untangles audio foundation model research into "audio-to-feature" representation extraction (Section 2) and "feature-to-target" downstream modeling (Section 3); (b) we systematically categorize five representation injection mechanisms (Section 3.1) and analyze their downstream task suitability, providing clear architectural guidelines for next-generation systems; and (c) we provide a comprehensive, multi-dimensional taxonomy of frontier models across three appendices: representation models (Appendix D), pre-trained foundation models (Appendix F.1), and fine-tuned downstream models (Appendix F.2). This decoupled lens offers actionable blueprints for modular audio architectures.

Proceedings of the 1st Workshop on Multi-modality and Multi-lingual Language Learning, pages 100–130
November 9, 2026 ©2026 Association for Computational Linguistics

2 Audio Representation Learning

An ideal universal audio representation efficiently generalizes across diverse downstream tasks under limited supervision (Turian et al., 2021). To systematically deconstruct this field, this chapter analyzes: (1) **audio modeling domains** of raw signals (Section 2.1), (2) **representation form** of audio features (Section 2.2), (3) **architectural designs** of mainstream audio representation models (Section 2.3), and (4) data-driven **pre-training paradigms** (Section 2.4). Additionally, the appendix provides: (5) **evaluation methods** (Section B) and (6) a summary of mainstream **audio representation models** (Section D).

2.1 Audio Modeling Domain

2.1.1 Time Domain

Time-domain modeling refers to directly ingesting sampled raw acoustic waveforms into neural networks without any transform-domain operations. By preserving complete physical signals, such as phase and transient details, this approach is highly suited for tasks requiring sample-level fidelity and phase-sensitive reconstruction such as speech separation (Bahmaninezhad et al., 2019; Luo and Mesgarani, 2019; Luo et al., 2020), high-fidelity audio generation (van den Oord et al., 2016), and universal representation extraction (Yuksel et al., 2025). Eliminating transform-related latency also natively supports ultra-low latency applications (Défossez et al., 2023). However, raw waveforms result in extremely long sequences, posing severe computational challenges for global attention mechanisms. Furthermore, lacking explicit harmonic inductive biases, pure time-domain networks often face optimization difficulties in signal-to-signal tasks (Nossier et al., 2020; Lin et al., 2024). To address this, feature extraction front-ends commonly leverage deep strided 1D-CNNs (Défossez et al., 2023) or parameterized physical filterbanks (Ravanelli and Bengio, 2018) to reduce dimensionality and capture local temporal continuity.

2.1.2 Time-Frequency Domain

To capture temporal evolution, mainstream audio tasks rely on time-frequency (T-F) modeling, where the Short-Time Fourier Transform (STFT) projects 1D waveforms into 2D spectrograms. T-F representations compress the temporal dimension, reducing computational complexity (Radford et al., 2023; Chen et al., 2023), and allow the seamless inheritance of computer vision architectures using the "audio-as-image" paradigm (Gong et al., 2021; Huang et al., 2022; Tuncay et al., 2025). Backed by strong frequency inductive biases, T-F modeling is well-suited for spectrum-dependent tasks like Audio Event Detection (Hertel et al., 2016; Kong et al., 2020b), Speech Emotion Recognition (Penumajji, 2025; Cao et al., 2025), and Audio Deepfake Detection (Pachon and Ballesteros, 2026; Hamza et al., 2022). Nevertheless, T-F modeling faces physical and mathematical limitations. First, STFT suffers from a rigid time-frequency resolution trade-off, typically mitigated by multi-resolution STFT losses (Kong et al., 2020a; Yuan et al., 2025) or learnable front-ends (Zeghidour et al., 2021; Leiber et al., 2025). Second, phase wrapping within $(-\pi, \pi]$ disrupts gradient continuity (Engel et al., 2019; Ai and Ling, 2023). Consequently, standard models abandon phase modeling (Radford et al., 2023; Huang et al., 2022), or overcome this bottleneck through adversarial training with time-domain discriminators (Zeghidour et al., 2022; Défossez et al., 2023; Zhang et al., 2024) or direct prediction of the complex spectrum's real and imaginary parts (Feng et al., 2025; Wang et al., 2023).

2.2 Audio Representation Form

2.2.1 Continuous Representation

Continuous representations consist of high-dimensional dense float vector sequences extracted by an audio encoder at each time step. By preserving fine-grained acoustic details, they systematically excel on speech understanding benchmarks (Wang et al., 2025a). Consequently, mainstream autoregressive understanding models (e.g., Qwen2-Audio (Chu et al., 2024), SALMONN (Tang et al., 2024), Audio Flamingo 3 (Ghosh et al., 2025)) and diffusion language models (e.g., the DIFFA series (Zhou et al., 2026a,b)) widely adopt them. Integrating these continuous sequences into modern large language models (LLMs) typically requires custom adapter modules and alignment training for feature mapping and downsampling. On the generation side, continuous representations naturally align with continuous-space diffusion and flow matching algorithms (Le et al., 2023; Hung et al., 2026; Tian et al., 2026d). For autoregressive generation, systems can leverage regression-based objectives to predict continuous target vectors directly, which subsequently condition non-autoregressive models (Liu et al., 2024b; Rouard et al., 2026). Alterna-

tively, methods can map generated discrete tokens into continuous vectors to guide subsequent acoustic reconstruction (Du et al., 2024c, 2025).

2.2.2 Discrete Representation

Discrete representations compress high-dimensional continuous signals into sequences of discrete symbols. These are primarily categorized into acoustic tokens, generated by neural audio codecs (van den Oord et al., 2017; Kumar et al., 2023; Pepino et al., 2025; Gong et al., 2026a) via residual vector quantization (RVQ) for high-fidelity reconstruction, and semantic tokens, derived from applying K-means to self-supervised representations like HuBERT (Hsu et al., 2021) and wav2vec 2.0 (Baevski et al., 2020) to capture phonetic information (Borsos et al., 2023a,b). For downstream modeling, discrete tokens enable autoregressive language models to perform next-token prediction via vocabulary expansion (Borsos et al., 2023a; Agostinelli et al., 2023; Shi et al., 2025; Arora et al., 2025), or can be modeled via classification-based discrete diffusion (Borsos et al., 2023b; Yang et al., 2023b; Long et al., 2026). This discretization unifies text and audio processing, endowing speech synthesis systems with remarkable zero-shot and in-context capabilities (Chen et al., 2025c, 2024a). To optimize the compression ratio, recent architectures apply semantic distillation during codec training to preserve semantic content and acoustic quality at low frame rates (Défossez et al., 2024).

2.3 Audio Representation Models and Downstream Components

Figure 1 summarizes audio representation architectures and downstream components.

2.3.1 Encoder-only

With no learnable decoder, this architecture focuses on generating dense, decoupled, discriminative latent representations, and its downstream head is typically a lightweight linear classifier. Representative Encoder-only models include AST (Gong et al., 2021), HuBERT (Hsu et al., 2021), BEATs (Chen et al., 2023), MERT (Li et al., 2024c), USAD (Chang et al., 2025), and CAT (Han et al., 2026). Furthermore, this paradigm extends to multimodal scenarios, where models such as CLAP (Elizalde et al., 2023) and SLAP (Siamese) (Guinot et al., 2025) employ a dual-encoder structure to achieve cross-modal alignment.

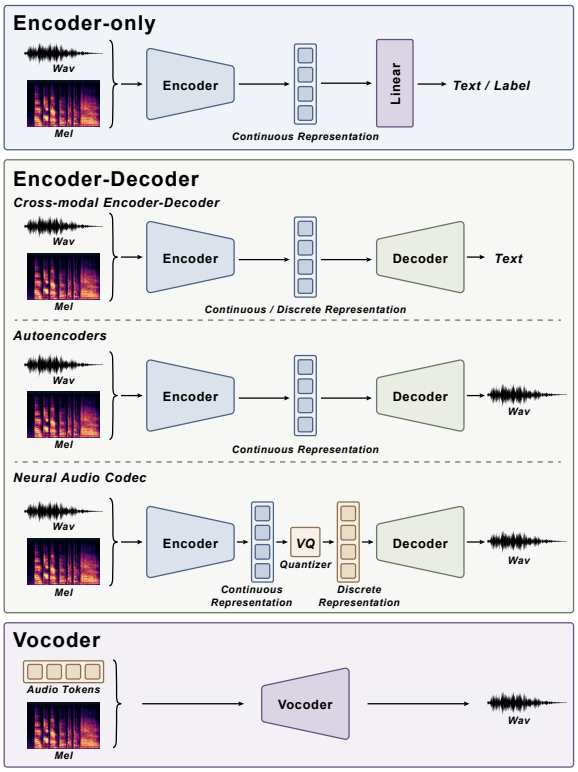


Figure 1: Architectures of Mainstream Audio Representation Models and Vocoder.

2.3.2 Encoder-Decoder

An encoder maps input signals to latent representations, which a decoder converts into waveforms, spectrograms, or text. Depending on the input/output modalities and bottleneck properties, this architecture features three subcategories:

Cross-modal Encoder-Decoders align and translate across heterogeneous modalities. Whisper (Radford et al., 2023) and OWSM v3.1 (Peng et al., 2024) employ native Transformer decoders for tasks like automatic speech recognition. StableToken (Song et al., 2026) extends Whisper with a quantization module for robust speech tokens, while SLAP (Scalable) (Mei et al., 2026) adopts an audio-text dual-encoder combined with a single text decoder for efficient multimodal retrieval and classification.

Autoencoders extract intrinsic representations by mapping an input $\mathbf{x}$ to a low-dimensional latent space $\mathbf{z}$ and minimizing the reconstruction error. Structurally, *Standard Autoencoders* such as DWM (Niizumi et al., 2026), DashengTokenizer (Dinkel et al., 2026a), and CARL (Sridhar and Cartwright, 2025) maintain symmetric frameworks. *Asymmetric Autoencoders*, including EnCodecMAE (Pepino et al., 2025), Audio-JEPA (Tuncay et al., 2025),

MATPAC (Quelennec et al., 2025), and WavJEPA (Yuksel et al., 2025), utilize lightweight decoders for pre-training reconstruction that are discarded during inference to yield highly discriminative features. Additionally, *Diffusion Autoencoders* like LatentFT (Wang and Huang, 2026) incorporate diffusion models into decoding to enhance latent capacity and signal restoration.

Neural Audio Codecs augment traditional autoencoders with quantizers. They impose a stringent information bottleneck to compress high-dimensional raw audio into low-dimensional tokens, followed by high-fidelity waveform restoration via a generative decoder. Representative models include MOSS-Audio-Tokenizer (Gong et al., 2026a), UniCodec (Jiang et al., 2025), SpeechTokenizer (Zhang et al., 2024), DAC (Kumar et al., 2023), EnCodec (Défossez et al., 2023), and SoundStream (Zeghidour et al., 2022).

2.3.3 Vocoders

Vocoders typically map acoustic intermediate representations, such as Mel-spectrograms, back into raw time-domain waveforms. Prominent examples of this paradigm include BemaGANv2 (Park et al., 2025), PeriodWave (Lee et al., 2025), Vocos (Siuzdak, 2024), and WaveNeXt (Okamoto et al., 2023). Furthermore, recent advancements have extended this framework beyond continuous spectrograms; for instance, T-Mimi (Wu et al., 2026b) trains the vocoder to reconstruct high-fidelity audio conditioned entirely on discrete codec tokens generated by the Mimi encoder.

2.4 Audio Representation Model Pre-training Paradigms

2.4.1 Supervised Pre-training

Audio representation models heavily rely on massive training datasets, yet acquiring clean human annotations is both challenging and expensive. Consequently, weakly supervised learning utilizing large-scale noisy data has emerged as one of the dominant paradigms to overcome this bottleneck. In large-scale weakly supervised pre-training, models such as Whisper (Radford et al., 2023), PANNs (Kong et al., 2020b), Contrastive-init/Captioning-scratch (Tseng et al., 2026), OWSM v3.1 (Peng et al., 2024), and StableToken (Song et al., 2026) directly leverage massive web data with weak labels, significantly enhancing their zero-shot generalization capabilities. Additionally, weakly supervised knowledge distillation—adopted by SLAP (Siamese)(Guinot et al., 2025), USAD (Chang et al., 2025), LLM-Codec (Yang et al., 2024), and CED (Dinkel et al., 2024)—utilizes soft labels generated by powerful teacher models to guide the learning process, enabling the efficient transfer of deep semantic features. On the generation side, neural vocoders including WaveNeXt (Okamoto et al., 2023), Vocos (Siuzdak, 2024), BemaGANv2 (Park et al., 2025), and T-Mimi (Wu et al., 2026b) still fundamentally depend on real physical waveforms or spectrograms as strong supervised regression targets to ensure high-fidelity audio reconstruction. In addition, PACE (Li et al., 2026a) proposes a continuous learning framework to address catastrophic forgetting in the pre-training of audio representation models.

2.4.2 Self-Supervised Pre-training (SSL)

Self-Supervised Learning (SSL) has become the dominant training paradigm for audio representation. It learns robust features from unlabeled audio through proxy tasks, deriving supervision directly from the data's structure. SSL objectives are primarily categorized into three types, often complemented by joint pre-training:

Predictive Objectives. These losses compel models to infer masked targets or distill teacher-generated latent features to capture high-level semantics. Specifically, HuBERT (Hsu et al., 2021), BEATs (Chen et al., 2023), and DWM (Niizumi et al., 2026) predict discrete labels, whereas data2vec (Baevski et al., 2022), EAT (Chen et al., 2024b), and Audio-JEPA (Tuncay et al., 2025) leverage Teacher-Student architectures for latent feature distillation. Additionally, LST (Lu et al., 2026b) employs autoregressive sequence prediction.

Contrastive Objectives. By pulling positive samples closer and pushing negatives apart, these losses filter out low-level acoustic noise to extract discriminative, abstract semantics. CLAP (Elizalde et al., 2023) and SLAP (Scalable) (Mei et al., 2026) utilize this mechanism for cross-modal audio-text alignment. Furthermore, wav2vec 2.0 (Baevski et al., 2020) combines contrastive learning with masked prediction over quantized representations, while AaSP (Yamamoto and Okusa, 2026) applies contrastive regularization to enhance representation discriminability.

Reconstructive Objectives. After an information bottleneck, these losses minimize reconstruction errors in signal or latent space to flawlessly

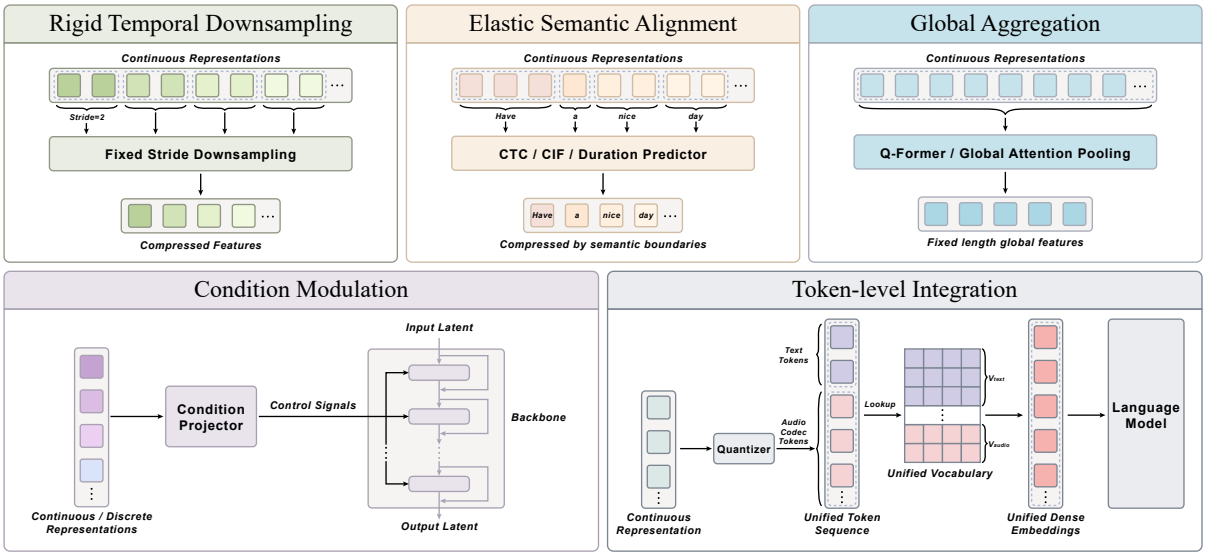


Figure 2: A Comparison of Audio Representation Integration Mechanisms for Downstream Models.

restore acoustic details. For example, Audio-MAE (Huang et al., 2022) and MAE-AST (Baade et al., 2022) directly reconstruct masked spectrograms. Similarly, neural codecs like VQ-VAE (van den Oord et al., 2017), EnCodec (Défossez et al., 2023), and DAC (Kumar et al., 2023) strictly rely on reconstruction objectives to guarantee high-fidelity audio generation.

3 Modeling for Audio Tasks

This chapter explores downstream workflows following audio representation extraction. Specifically, we investigate **integration mechanisms** into downstream models(Section 3.1) and subsequent downstream **modeling paradigms** (Section 3.2). Furthermore, the appendix introduces a comprehensive **taxonomy of audio tasks** (Section C), outlines **training paradigms** (Section E), and summarizes **representative audio models** (Section F).

3.1 Audio Representation Integration Mechanisms for Downstream Models

As illustrated in Figure 2, we systematically categorize these mechanisms into five paradigms: Rigid Temporal Downsampling, Elastic Semantic Alignment, Global Aggregation, Condition Modulation and Token-level Integration.

3.1.1 Rigid Temporal Downsampling

Rigid temporal downsampling applies fixed-ratio compression to extract acoustic envelopes. Though computationally efficient and strictly order-preserving, it risks over-smoothing and losing fine-grained high-frequency details. Common adapters for fixed-stride compression include 1D-CNNs (Xie and Wu, 2024; Chen et al., 2026; Tian et al., 2026b) and strided linear pooling (Xu et al., 2025; StepFun Audio Team, 2025; Tian et al., 2026a; Ghosh et al., 2026a). To mitigate the information loss inherent in uniform compression, window-level Q-Former (Tang et al., 2024; Wang et al., 2025d) and window-level self-attention (Jia et al., 2025) introduce adaptive weighting within fixed temporal windows.

3.1.2 Elastic Semantic Alignment

Addressing uneven audio density, elastic semantic alignment reshapes the time axis along semantic boundaries, mitigating the sequence length mismatch between acoustic frames and language model inputs while preserving semantics. Representative mechanisms include CTC-guided compression (Deng et al., 2026; Zhou et al., 2025), which suppresses blank frames via posterior peaks for variable-ratio compression; Continuous Integrate-and-Fire (CIF)(Gao et al., 2022; Deng et al., 2025a), which accumulates frame weights to emit fused representations upon reaching a threshold; and duration predictors(Le et al., 2023), which explicitly estimate frame counts per phoneme to guide feature pooling.

3.1.3 Global Aggregation

By discarding fine-grained temporal information, global aggregation mechanisms condense core acoustic features from variable-length audio into a fixed-length summary. Representative techniques

Cat.	Task	Benchmark & Metrics	Model	Downstream Inference Data Flow Pipeline	Performance
Und	ASR	**LibriSpeech** test-clean WER↓|test-other WER↓	CJST (Zhou et al., 2025)	$[\mathcal{M}\rightarrow\text{VGG}^{\text{RTD}}\rightarrow\text{Conformer}^{\text{ESA}}\rightarrow H_{\text{A}}]\rightarrow$Dec-only Tr$_{\text{AR}}\rightarrow\mathcal{T}$	2.09% | 4.71%
			MDM-ASR (Yen et al., 2026)	$[\mathcal{M}\rightarrow\text{FastConformer}^{\text{RTD}}\rightarrow H_{\text{A}}^{[\text{M}]}]\ \|\ [Z_{\text{T}}]\rightarrow$Non-causal Tr Dec$_{\text{iNAR}}\rightarrow\mathcal{T}$	1.80% | 3.60%
			Qwen3-Omni-30B-A3B-Instruct (Qwen Team, 2025)	$[\mathcal{M}\rightarrow\text{AuT}^{\text{RTD}}\rightarrow H_{\text{A}}]\oplus[Z_{\text{T}}]\rightarrow\text{Qwen3}_{\text{AR}}\rightarrow\mathcal{T}$	**1.22%** | **2.48%**
	AQA	**MMAR** ACC↑	DiFFA-2 (Zhou et al., 2026b)	$[\mathcal{M}\rightarrow$Whisper Encoder$\rightarrow H_{\text{A}}\rightarrow$[CNN$\rightarrow$Linear$^{\text{RTD}}\rightarrow H_{\text{A}}]\oplus$[Q-Former$^{\text{DA}}\rightarrow H_{\text{A}}]]\ \|\ [Z_{\text{T}}]\rightarrow\text{LLaDA}_{\text{iNAR}}\rightarrow\mathcal{T}$	50.80%
			AF-Next-Think (Ghosh et al., 2026a)	$[\mathcal{M}\rightarrow$AF-Whisper Encoder$\rightarrow H_{\text{A}}\rightarrow$Stride-2 Pool$^{\text{RTD}}\rightarrow$MLP$\rightarrow H_{\text{A}}]\oplus[\mathcal{T}]\rightarrow\text{Qwen2.5}_{\text{AR}}\rightarrow\mathcal{T}$	61.00%
			Audio-DeepThinker (He et al., 2026)	$[\mathcal{M}]\ \|\ [\mathcal{T}]\rightarrow$Qwen3-Omni-30b-A3B$_{\text{AR}}\rightarrow\mathcal{T}$	**74.00%**
Gen	TTS	**Seed-TTS-Eval** test-zh CER↓|test-zh SIM↑|test-en WER↓|test-en SIM↑	OmniVoice (Zhu et al., 2026)	$[\mathcal{W}\rightarrow$Higgs-Audio Tok$\rightarrow Z_{\text{A}}^{[]}]\oplus[\mathcal{T}]\rightarrow\text{Qwen3}_{\text{iNAR}}\rightarrow Z_{\text{A}}\rightarrow$Higgs-Audio Detok$\rightarrow\mathcal{W}$	1.60% | 0.741 | **0.84%** | **0.777**
			CosyVoice 3 (Du et al., 2025)	$[\mathcal{W}\rightarrow\text{FSQ-MinMo}\rightarrow Z_{\text{A}}^{[]}\oplus[\mathcal{T}]\rightarrow\text{Tr}_{\text{AR}}\rightarrow Z_{\text{A}}^{[\text{M}]}\rightarrow\text{DiT}_{\text{iNAR}}\rightarrow\mathcal{M}\rightarrow\text{Vocoder}\rightarrow\mathcal{W}$	**0.71%** | **0.775** | 1.45% | 0.695
			VibeVoice (Peng et al., 2026b)	$[\mathcal{W}\rightarrow$Acoustic Tok$^{\text{RTD}}\rightarrow H_{\text{A}}]\oplus[\mathcal{T}]\rightarrow\text{Qwen2.5}_{\text{AR}}\rightarrow H_{\text{A}}^{\text{LM}}\rightarrow$Diffusion$_{\text{iNAR}}\rightarrow H_{\text{A}}\rightarrow$Acoustic Dec$\rightarrow\mathcal{W}$	1.16% | 0.744 | 3.04% | 0.689
	TTA	**AudioCaps** $\text{FD}_{\text{L3}}\downarrow$| $\text{KL}_{\text{P}}\downarrow$| FD ↓| FAD ↓| KL ↓| CLAP ↑	TangoFlux (Hung et al., 2026)	$[\mathcal{W}\rightarrow$Stable Audio Open VAE$\rightarrow H_{\text{A}}^{[\text{M}]}]\oplus[\mathcal{T}\rightarrow\text{FLAN-T5}\rightarrow H_{\text{T}}^{[\text{M}]}]\rightarrow\text{DiT}_{\text{iNAR}}\rightarrow H_{\text{A}}\rightarrow$Stable Audio Open VAE Dec$\rightarrow\mathcal{W}$	83.58 | 0.95 | 1.57 | 2.34 | 0.52 | **0.385**
			AudioLDM 2 (Liu et al., 2024b)	$\mathcal{T}\rightarrow$[CLAP]$\ |\ $[FLAN-T5]$\rightarrow H_{\text{T}}\rightarrow\text{GPT-2}_{\text{AR}}\rightarrow H_{\text{T}}^{\text{LM}}\rightarrow$ Transformer-UNet$_{\text{iNAR}}\rightarrow H_{\text{A}}\rightarrow$VAE Dec$\rightarrow\mathcal{M}\rightarrow$HiFi-GAN$\rightarrow\mathcal{W}$	87.74 | 1.01 | 1.59 | 2.63 | 0.57 | 0.252
			AudioStory (Guo et al., 2025)	$[\mathcal{M}\rightarrow$Whisper Encoder$^{\text{RTD}}\rightarrow H_{\text{A}}]\oplus[\mathcal{T}_p]\rightarrow\text{Qwen2.5}_{\text{AR}}\rightarrow[\mathcal{T}\ \|\ Z_{\text{A}}^{[\text{M}]}]\rightarrow\text{TangoFlux}_{\text{iNAR}}\rightarrow H_{\text{A}}\rightarrow$VAE Dec$\rightarrow\mathcal{W}$	**83.39** | **0.91** | **1.52** | **2.29** | **0.51** | 0.383
Conv	VC	**Seed-TTS-Eval** test-zh CER↓|test-zh SIM↑|test-en WER↓|test-en SIM↑	MeanVC (Ma et al., 2026a)	$[\mathcal{W}\rightarrow\text{ECAPA-TDNN}^{\text{GA}}\rightarrow H_{\text{A}}^{[\text{M}]}\oplus[[[\mathcal{W}\rightarrow\text{Fast-U2++}\rightarrow H_{\text{A}}]\ \|\ [\mathcal{M}]]\rightarrow$Timbre Enc$\rightarrow H_{\text{A}}^{[\text{M}]}]\rightarrow\text{DiT}_{\text{1-NAR}}\rightarrow\mathcal{M}\rightarrow\text{Vocos}\rightarrow\mathcal{W}$	3.89 | **0.73** | — | —
			X-VC (Zheng et al., 2026)	$[\mathcal{W}\rightarrow\text{SAC Enc}^{\text{RTD}}\rightarrow H_{\text{A}}]\ \|\ [[\mathcal{W}\rightarrow\text{ERes2Net}\rightarrow H_{\text{A}}^{[\text{M}]}]\oplus[\mathcal{M}^{[\text{M}]}]\rightarrow\text{Tr}_{\text{1-NAR}}\rightarrow H_{\text{A}}\rightarrow\text{SAC Dec}\rightarrow\mathcal{W}$	1.99 | **0.73** | 2.83 | **0.63**
			Seed-VC tiny (Liu, 2024)	$[\mathcal{M}\rightarrow$Whisper Encoder$\rightarrow H_{\text{A}}^{[\text{M}]}]\oplus[\mathcal{W}\rightarrow\text{CAM++}\rightarrow H_{\text{A}}^{[\text{M}]}]\oplus[\mathcal{W}\rightarrow\text{RMVPE}\rightarrow H_{\text{A}}^{[\text{M}]}]\rightarrow\text{U-DiT}_{\text{iNAR}}\rightarrow\mathcal{M}\rightarrow\text{BigVGANv2}\rightarrow\mathcal{W}$	**1.79** | 0.60 | **2.24** | 0.41
Int	FSD	**Full-Duplex-Bench** DMOS↓ | SSIM↑ | LAL↓ | GPT4o↑	PersonaPlex (Roy et al., 2026)	$\mathcal{W}\rightarrow\text{Moshi}_{\text{AR}}\rightarrow\mathcal{W}$	3.90 | **0.57** | **0.070** | **4.210**
			Moshi (Défossez et al., 2024)	$[\mathcal{W}\rightarrow\text{Mimi Tok}^{\text{RTD}}\rightarrow Z_{\text{A}}^{[]}]\rightarrow\text{RQ-Tr}_{\text{AR}}\rightarrow[Z_{\text{T}}\ \|\ Z_{\text{A}}]\rightarrow\text{Mimi Detok}\rightarrow\mathcal{W}$	**3.11** | 0.10 | 0.265 | 0.765
			Freeze-Omni (Wang et al., 2025c)	$[\mathcal{M}\rightarrow\text{Speech Enc}\rightarrow\text{Adapter}^{\text{RTD}}\rightarrow H_{\text{A}}]\rightarrow\text{Qwen2-7B}_{\text{AR}}\rightarrow[H_{\text{A}}\ \|\ Z_{\text{T}}]\rightarrow\text{Speech Dec}_{\text{AR}}\rightarrow Z_{\text{A}}\rightarrow\text{TiCodec}\rightarrow\mathcal{W}$	3.51 | 0.05 | 0.953 | 3.615

$\mathcal{W}$: Waveform | $\mathcal{M}$: Mel-spec | $\mathcal{T}$: Text | H_A/H_T: Audio/Text Continuous Representations | Z_A/Z_T: Audio/Text Discrete Tokens
||: Parallel branches | $\oplus$: Feature fusion or concatenation | **Tr**: Transformer | **Enc/Dec**: Encoder/Decoder | **Tok/Detok**: Tokenizer/Detokenizer
RTD: Rigid Temporal Downsampling | CM: Condition Modulation | ESA: Elastic Semantic Alignment | GA: Global Aggregation | TI: Token-level Integration
AR: Autoregressive generation | iNAR: Iterative Non-Autoregressive | 1-NAR: One-shot Non-Autoregressive

Table 1: Overview of representative audio and speech models, their inference pipelines, and performance. Performance for TTA, VC, and FSD tasks are sourced from AudioStory (Guo et al., 2025), X-VC (Zheng et al., 2026), and PersonaPlex (Roy et al., 2026), respectively. For all other models, the performance are obtained directly from their respective original publications. Bold indicates the best result, underline denotes the second-best result.

include Q-Former (Ghosh et al., 2024; Biswas et al., 2026; Lu et al., 2026a), which uses a set of learnable queries to attend to the entire acoustic sequence, yielding compact, length-agnostic representations for downstream LLMs. Global Attention Pooling (Popov et al., 2022; Wang et al., 2026d) derives this macroscopic summary by computing a dynamically weighted sum of all audio frames using learned attention scores.

3.1.4 Condition Modulation

Condition modulation guides generation by dynamically reshaping backbone feature distributions using audio representation as independent variables. Three core mechanisms prevail: FiLM (Liu et al., 2024c; Gupta et al., 2026) predicts channel-wise affine parameters for coarse-grained global injection, offering minimal computational overhead in CNNs; AdaLN (Chen et al., 2025c; Liu, 2024; Dror et al., 2026) maps conditions to scaling and shifting parameters, injecting global styles while stabilizing gradients—now standard in DiTs and flow-matching architectures; and cross-attention (Vyas et al., 2023; Evans et al., 2025; Peng et al., 2026b) leverages dynamic similarity addressing, using backbone features as queries to achieve fine-grained alignment with variable-length condition sequences.

3.1.5 Token-level Integration

Token-level integration quantizes continuous audio into symbol sequences, reducing complex cross-modal modeling to a unified next-token prediction paradigm. Current approaches fall into three categories: (1) training dedicated language models from scratch without pre-trained backbones (Borsos et al., 2023a; Chen et al., 2024a; Anastassiou et al., 2024); (2) expanding the vocabularies of pre-trained large language models (Rubenstein et al., 2023; Liu et al., 2026a) or discrete diffusion models (Long et al., 2026); and (3) employing independent audio and text vocabularies for decoupled prediction (Défossez et al., 2024; Chien et al., 2026).

3.2 Downstream Task Modeling Paradigm

3.2.1 Discriminative

Discriminative modeling directly learns mappings from audio representations to specific targets, outputting classification or regression results via lightweight neural networks. Discrete targets aim to map inputs into a finite, mutually exclusive label space, widely used in audio classification (Gupta et al., 2026; Li et al., 2026a). Conversely, continuous targets project features into a continuous real-valued space, exemplified by predicting continuous time-frequency masks for speech separation and enhancement (Wang et al., 2023; Liu et al., 2024c), and similarity-based anomalous sound detection (Wilkinghoff et al., 2026).

3.2.2 Autoregressive

Autoregressive modeling generates causal sequences via the chain rule, constrained by continuous or discrete representations. Continuous representations map to LLM latent spaces via

lightweight adapters for audio understanding (Chu et al., 2024; Xu et al., 2025; Qwen Team, 2025; Ghosh et al., 2025, 2026a). However, predicting continuous matrices is intractable, restricting them to inputs; furthermore, the lack of shared symbols creates a modality gap that impedes chain-of-thought integration (Wu et al., 2026a). Conversely, discrete tokens seamlessly align with AR paradigms, sharing embedding spaces and cross-entropy loss with text for audio generation (Borsos et al., 2023a; Chen et al., 2025c; Gong et al., 2026b). Nevertheless, discretization incurs information loss, and inherently lengthy audio sequences necessitate mitigation via dynamic token compression (Lu et al., 2026b).

3.2.3 Iterative NAR Generation

Iterative generation performs parallel, synchronous sequence updates using global context. In discrete spaces, discrete diffusion recovers real tokens from uniform noise, enabling audio generation via vocabulary expansion (Borsos et al., 2023b; Long et al., 2026), or audio understanding by predicting text tokens from concatenated continuous audio and text prompts (Zhou et al., 2026a,b). In continuous spaces, traditional diffusion models denoise Gaussian noise on latent (Evans et al., 2025) or Mel-spectrogram (Popov et al., 2022) manifolds. To optimize evolution trajectories, flow matching directly regresses ODE vector fields, significantly easing network fitting and permitting large inference steps to balance fidelity and efficiency (Le et al., 2023; Liu, 2024).

3.2.4 One-shot NAR Generation

Assuming conditional independence, one-shot non-autoregressive (NAR) generation yields complete sequences in a single forward pass with $O(1)$ latency. To bridge cross-modal length mismatches, it typically employs elastic semantic alignment via CTC (Altinok, 2025), CIF (Gao et al., 2022), or duration predictors (Ren et al., 2021; Kim et al., 2021a). For strictly time-aligned scenarios, it employs deterministic feature mapping in continuous latent spaces (Zheng et al., 2026) or parallel slot-filling label prediction in discrete spaces (Mu et al., 2026). Consistency Models (Ye et al., 2023; Li et al., 2024b) distill iterative diffusion trajectories into single-step mappings for high-fidelity continuous generation.

3.2.5 Hybrid Modeling Paradigms

Hybrid paradigms balance sequence coherence and generation efficiency through three evolutionary paths. First, integrating AR and Iterative NAR: Seed-TTS (Anastassiou et al., 2024) and CosyVoice3 (Du et al., 2025) cascade AR planning with NAR acoustic refinement, whereas DiTAR (Jia et al., 2025) interleaves global AR with local intra-patch diffusion. Second, synergizing AR with One-shot NAR: VALL-E2(Chen et al., 2024a) and CALM (Rouard et al., 2026) leverage AR for long-range logic while employing ultra-fast single-step NAR to decode multi-layer or continuous concurrent features. Third, cascading discriminative and generative models: Speech World Model (Zhou et al., 2026c) uses discriminative graph networks to predict audio tags, serving as prompts to guide controlled AR generation.

3.3 Representative Systems and Data Flow Pipelines

Table 1 summarizes over 20 representative downstream systems across four core domains (understanding, generation, conversion, and interaction) evaluated on standard benchmarks (e.g., LibriSpeech (Panayotov et al., 2015), MMAR (Ma et al., 2026b), Seed-TTS-Eval (Anastassiou et al., 2024), AudioCaps (Kim et al., 2019), and Full-Duplex-Bench (Lin et al., 2025a)). Our notation describes workflows through input modalities, integration adapters, and downstream modeling paradigms. Reported performance is not directly comparable because model scales, pre-training data, and evaluation settings differ. The table illustrates how pipeline choices—continuous versus discrete representations and autoregressive versus non-autoregressive architectures—shape downstream capabilities.

4 Comparative Discussion and Future Directions

4.1 Time and Time-Frequency Trade-off

Selecting between time-domain (T) and time-frequency (T-F) representations involves a trade-off between signal preservation and computational efficiency. Time-domain modeling preserves raw transients and phase, favoring tasks requiring waveform reconstruction or low-latency execution, such as speech separation (Luo and Mesgarani, 2019), telecommunication (Défossez et al., 2023), and super-resolution (Yuan et al., 2025). Conversely,

T-F modeling compresses sequence length and introduces harmonic priors, making it suitable for semantic comprehension and classification tasks like event detection (Kong et al., 2020b) and emotion recognition (Penumajji, 2025). For downstream tasks demanding both frequency inductive biases and phase details, hybrid modeling is adopted. For example, OWL (Biswas et al., 2026) extracts STFT and phase features simultaneously to support spatial audio understanding.

4.2 Continuous and Discrete Representation Trade-off

Selecting between continuous and discrete representations fundamentally trades off acoustic fidelity and text-audio modality alignment. Continuous representations preserve dense acoustic details, making them suited for discriminative speech understanding (Wang et al., 2025a) and high-fidelity continuous diffusion (Hung et al., 2026). Conversely, discrete representations introduce quantization loss but successfully unify audio and text tokens within a single vocabulary. Thus, discrete tokens are preferred for interactive speech-to-speech agents and in-context voice synthesis (Chen et al., 2025c; Défossez et al., 2024), which require unified autoregressive generation. Recent systems combine both representation types hierarchically (Du et al., 2024c, 2025), or use continuous representations for high-quality, end-to-end reasoning and generation (Rouard et al., 2026).

4.3 Comparative Discussion on Integration Mechanisms

Selecting an integration adapter depends heavily on the downstream task's demand for temporal resolution versus global stylization. For sequence-preserving speech understanding and translation, rigid downsampling (Xie and Wu, 2024; Tian et al., 2026a; Tang et al., 2024) and elastic semantic alignment (Deng et al., 2026; Gao et al., 2022; Le et al., 2023) are commonly adopted to align temporal resolution while preserving local ordering. Conversely, global classification and retrieval tasks prefer global aggregation (Biswas et al., 2026; Popov et al., 2022) to extract length-agnostic semantic summaries. For high-fidelity generation, condition modulation mechanisms like AdaLN (Chen et al., 2025c) and cross-attention (Vyas et al., 2023) are standard for injecting stylistic constraints, while token-level integration (Borsos et al., 2023a; Défossez et al., 2024; Roy et al., 2026) is well-suited for unified, conversational speech-to-speech modeling.

4.4 Downstream Modeling Paradigm Selection and Evolving Trends

Downstream modeling depends on representations and adapters. Discriminative modeling, paired with rigid downsampling or global aggregation, is suited for classification and mask prediction (Wang et al., 2023; Gupta et al., 2026). Conversely, target-centric generation utilizes iterative or one-shot NAR paradigms, paired with condition modulation and elastic alignment, to achieve low-latency synthesis (Ren et al., 2021; Le et al., 2023). For conversational and reasoning tasks, AR modeling remains preferred to unify text-audio sequence learning (Borsos et al., 2023a). The current trend emphasizes hybrid cascades combining AR semantic planning with NAR acoustic refinement (Du et al., 2025), while pushing toward continuous autoregressive architectures (Rouard et al., 2026) and audio-interleaved reasoning (Lin et al., 2025b; Wu et al., 2026a) to eliminate quantization bottlenecks.

5 Conclusion

This survey presents a decoupled framework separating the audio processing pipeline into representation extraction and downstream modeling. We systematically review audio domains, representations, and architectures, while categorizing five integration mechanisms and downstream modeling paradigms based on task suitability and trade-offs. The appendix catalogs representative models. These taxonomies offer guidelines for designing next-generation audio foundation models.

Limitations

This survey covers literature up to early 2026 and may omit recent developments. Differences in datasets, hardware, and evaluation protocols limit direct comparisons, especially for subjective generation tasks. Our analysis remains primarily qualitative because controlled comparisons across design choices are scarce and incomplete public documentation restricts assessment of training costs and efficiency.

Declaration on Generative AI

Generative AI was used to assist with language editing and text refinement, including rephrasing sentences and removing redundant expressions.

References

Ossama Abdel-Hamid, Abdel rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Yang Ai and Zhen-Hua Ling. 2023. Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Tony Alex, Sara Ahmed, Armin Mustafa, Muhammad Awais, and Philip J. B. Jackson. 2024. DTF-AT: Decoupled time-frequency audio transformer for event classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 17647–17655.

Duygu Altinok. 2025. Boosting ctc-based asr using llm-based intermediate loss regularization. In *International Conference on Text, Speech, and Dialogue*, pages 275–288. Springer.

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin M. Hasani, Saniya Karwa, Yuri Khrustalev, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel Loo, and 14 others. 2025. LFM2 technical report. *arXiv preprint arXiv:2511.23404*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. *Advances in Neural Information Processing Systems*, 37:56898–56918.

Hagai Aronowitz, Zvi Kons, Avihu Dekel, George Saon, and Ron Hoory. 2026. Speaker attributed automatic speech recognition using speech aware LLMs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19077–19081.

Siddhant Arora, Jinchuan Tian, Hayato Futami, Jeeweon Jung, Jiatong Shi, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. 2025. Chain-of-thought training for open E2E spoken dialogue systems. In *Proceedings of the 26th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA.

Alan Baade, Puyuan Peng, and David Harwath. 2022. MAE-AST: Masked autoencoding audio spectrogram transformer. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2438–2442.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1298–1312.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Fahimeh Bahmaninezhad, Jian Wu, Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Meng Yu, and Dong Yu. 2019. A comprehensive study of speech separation: Spectrogram vs waveform separation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4574–4578.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D. Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2025. AudioSetCaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2817–2829.

Nenad Banfic, David Fan, Kunal Vaishnavi, Sam Kemp, Sunghoon Choi, Rui Ren, Sayan Shaw, and Meng Tang. 2026. Pushing the limits of on-device streaming ASR: A compact, high-accuracy english model for low-latency inference. *arXiv preprint arXiv:2604.14493*.

Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. 2026. OWL: Geometry-aware spatial reasoning for audio large language models. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023a. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. SoundStorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

Ronghe Cao, Yunxing Wang, Xiaolong Wu, Shuang Jin, and Huiling Niu. 2025. A lightweight tri-stream feature fusion network for speech emotion recognition. *IEEE Access*.

Heng-Jui Chang, Saurabhchand Bhati, James R. Glass, and Alexander H. Liu. 2025. USAD: Universal speech and audio representation via distillation. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Gongyu Chen, Xiaoyu Zhang, Zhenqiang Weng, Junjie Zheng, Da Shen, Chaofan Ding, Wei-Qiang Zhang, and Zihao Chen. 2025a. YingMusic-SVC: Real-world robust zero-shot singing voice conversion with flow-grpo and singing-specific inductive biases. *arXiv preprint arXiv:2512.04793*.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021a. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proceedings of Interspeech 2021*, pages 3670–3674. ISCA.

Hongjie Chen, Zehan Li, Yaodong Song, Wenming Deng, Yitong Yao, Yuxin Zhang, Hang Lv, Xuechao Zhu, Jian Kang, Jie Lian, Jie Li, Chao Wang, Shuangyong Song, Yongxiang Li, Zhongjiang He, and Xuelong Li. 2025b. GOAT-SLM: A spoken language model with paralinguistic and speaker characteristic awareness. *arXiv preprint arXiv:2507.18119*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. VGGSound: A large-scale audio-visual dataset. In *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. 2021b. WaveGrad: Estimating gradients for waveform generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024a. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025c. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5178–5193.

Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. 2024b. EAT: Self-supervised pre-training with efficient audio transformer. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3807–3815.

William Chen, Prem Seetharaman, Rithesh Kumar, Oriol Nieto, Shinji Watanabe, Justin Salamon, and Zeyu Jin. 2026. AudioChat: Unified audio storytelling, editing, and understanding with transfusion forcing. *arXiv preprint arXiv:2602.17097*.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung yi Lee, and Lijuan Wang. 2026. STITCH: Simultaneous thinking and talking with chunked reasoning for spoken language models. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Chung-Ming Chien, Manu Orsini, Eugene Kharitonov, Neil Zeghidour, Karen Livescu, and Alexandre Défossez. 2026. MoshiRAG: Asynchronous knowledge retrieval for full-duplex speech language models. *arXiv preprint arXiv:2604.12928*.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. 2022. Self-supervised learning with random-projection quantizer for speech recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3915–3924.

Iwona Christop, Tomasz Kuczyński, and Marek Kubis. 2025. ClonEval: An open voice cloning benchmark. *Preprint*, arXiv:2504.20581.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Luca Comanducci, Paolo Bestagini, and Stefano Tubaro. 2025. FakeMusicCaps: A dataset for detection and attribution of synthetic music generated via text-to-music models. *Journal of Imaging*, 11(7):242.

Seamless Communication, Loic Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth ... Dalmia, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. LibriMix: An open-source dataset for generalizable speech separation. *Preprint*, arXiv:2005.11262.

Courtenay V. Cotton and Daniel P. W. Ellis. 2011. Spectral vs. spectro-temporal features for acoustic event detection. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 69–72.

Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2024. Emphassess: A prosodic benchmark on assessing emphasis transfer in speech-to-speech models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 495–507.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Keqi Deng, Wenxi Chen, Xie Chen, and Phil Woodland. 2025a. SimulS2S-LLM: Unlocking simultaneous inference of speech LLMs for speech-to-speech translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16718–16734, Vienna, Austria. Association for Computational Linguistics.

Keqi Deng, Ruchao Fan, Bo Ren, Yiming Wang, and Jinyu Li. 2026. Speech llms are contextual reasoning transcribers. *Preprint*, arXiv:2604.00610.

Keqi Deng, Guangzhi Sun, and Phil Woodland. 2025b. Wav2Prompt: End-to-end speech prompt learning and task-based fine-tuning for text-based LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6940–6956.

Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. 2026. Mellow: A small audio language model for reasoning. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: A multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. 2025. SoundMind: RL-incentivized logic reasoning for audio-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 528–540. Association for Computational Linguistics.

Heinrich Dinkel, Xingwei Sun, Gang Li, Jiahao Mei, Yadong Niu, Jizhong Liu, Xiyang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, and Jian Luan. 2026a. DashengTokenizer: One layer is enough for unified audio understanding and generation. *CoRR*, abs/2602.23765.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. 2024. CED: Consistent ensemble distillation for audio tagging. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 291–295.

Heinrich Dinkel, Zhiyong Yan, Tianzi Wang, Yongqing Wang, Xingwei Sun, Yadong Niu, Jizhong Liu, Gang Li, Junbo Zhang, and Jian Luan. 2026b. Glap: General contrastive audio-text pretraining across domains and languages. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 14737–14741. IEEE.

Tali Dror, Iftach Shoham, Moshe Buchris, Oren Gal, Haim H. Permuter, Gilad Katz, and Eliya Nachmani. 2026. Token-based audio inpainting via discrete

diffusion. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Hui-Peng Du, Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. 2024a. APNet2: High-quality and high-efficiency neural vocoder with direct prediction of amplitude and phase spectra. In *Proceedings of the Man-Machine Speech Communication Conference*, pages 66–80. Springer.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024b. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, and 3 others. 2025. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024c. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Mehrsa Golestaneh, and Robert Aichner. 2024. ICASSP 2023 Deep Noise Suppression Challenge. *IEEE Open Journal of Signal Processing*, 5:725–737.

Berkin Durmus, Blaise Munyampirwa, Eduardo Pacheco, Atila Orhon, and Andrey Leonov. 2025. SDBench: A comprehensive benchmark suite for speaker diarization. In *Proceedings of Interspeech 2025*, pages 1598–1602.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP learning audio concepts from natural language supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jesse H. Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. GANSynth: Adversarial neural audio synthesis. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable audio open. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiajun Fan, Roger Ren, Jingyuan Li, Rahul Pandey, Prashanth Gurunath Shivakumar, Ivan Bulyko, Ankur Gandhe, Ge Liu, and Yile Gu. 2026. Incentivizing consistent, effective and scalable reasoning capability in audio LLMs via reasoning process rewards. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Ruchao Fan, Bo Ren, Yuxuan Hu, Rui Zhao, Shujie Liu, and Jinyu Li. 2025. Alignformer: Modality matching can achieve better zero-shot instruction-following speech-llm. *IEEE Journal of Selected Topics in Signal Processing*, 19(7):1329–1337.

Tao Feng, Zhiyuan Zhao, Yifan Xie, Yuqi Ye, Xiangyang Luo, Xun Guan, and Yu Li. 2025. STFT-Codec: High-fidelity audio compression through time-frequency domain representation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Li Fu, Yu Xin, Sunlu Zeng, Lu Fan, Youzheng Wu, and Xiaodong He. 2026. PAC: Pronunciation-aware contextualized large language model-based automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16377–16381.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proc. Interspeech 2022*, pages 2063–2067.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. 2026a. Audio Flamingo Next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael

Valle, and Bryan Catanzaro. 2025. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2026b. Music flamingo: Scaling music understanding in audio language models. In *International Conference on Learning Representations (ICLR)*.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2026c. Music Flamingo: Scaling music understanding in audio language models. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S. Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313, Miami, Florida, USA. Association for Computational Linguistics.

Yitian Gong, Kuangwei Chen, Zhaoye Fei, Xiaogui Yang, Ke Chen, Yang Wang, Kexin Huang, Mingshu Chen, Ruixiao Li, Qingyuan Cheng, Shimin Li, and Xipeng Qiu. 2026a. MOSS-Audio-Tokenizer: Scaling audio tokenizers for future audio foundation models. *CoRR*, abs/2602.10934.

Yitian Gong, Botian Jiang, Yiwei Zhao, Yucheng Yuan, Kuangwei Chen, Yaozhou Jiang, Cheng Chang, Dong Hong, Mingshu Chen, Ruixiao Li, Yiyang Zhang, Yang Gao, Hanfu Chen, Ke Chen, Songlin Wang, Xiaogui Yang, Yuqian Zhang, Kexin Huang, ZhengYuan Lin, and 7 others. 2026b. MOSS-TTS technical report. *arXiv preprint arXiv:2603.18090*.

Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: Audio spectrogram transformer. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 571–575.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Petr Grinberg and Hassan Shahmohammadi. 2026. ALARM: Audio-language alignment for reasoning models. *arXiv preprint arXiv:2603.09556*.

Julien Guinot, Alain Riou, Elio Quinton, and George Fazekas. 2025. SLAP: Siamese language-audio pretraining without negative samples for music understanding. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 382–390.

Tianyu Guo, Hongyu Chen, Hao Liang, Meiyi Qiang, Bohan Zeng, Linzhuang Sun, Bin Cui, and Wentao Zhang. 2026a. Brace: A benchmark for robust audio caption quality evaluation. *Advances in Neural Information Processing Systems*, 38.

Yuxin Guo, Teng Wang, Yuying Ge, Shijie Ma, Yixiao Ge, Wei Zou, and Ying Shan. 2025. AudioStory: Generating long-form narrative audio with large language models. *arXiv preprint arXiv:2508.20088*.

Zhihan Guo, Wenqian Cui, Guan-Ting Lin, Daxin Tan, Jingyao Li, Qiyong Zheng, Dingdong Wang, Jing Xiong, Han Shi, Jiaya Jia, and Irwin King. 2026b. A survey of Audio Reasoning in Multimodal Foundation Models. *Preprint*, arXiv:2605.21008.

Shubham Gupta, Adarsh Arigala, B. R. Dilleswari, and Sri Rama Murty Kodukula. 2026. Shared representation learning for reference-guided targeted sound detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 14547–14551.

Alexandros Haliassos, Rodrigo Mira, and Stavros Petridis. 2026. Pay attention to CTC: Fast and robust pseudo-labelling for unified speech recognition. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Ameer Hamza, Abdul Rehman Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S. Almadhor, Zunera Jalil, and Rouba Borghol. 2022. Deepfake audio detection via MFCC features using machine learning. *IEEE Access*, 10:134018–134028.

Bing Han, Chushu Zhou, Yifan Yang, Wei Wang, Chenda Li, Wangyou Zhang, and Yanmin Qian. 2026. Representation-regularized convolutional audio transformer for audio understanding. *Preprint*, arXiv:2601.21612.

Xiang He, Chenxing Li, Jinting Wang, Yan Rong, Tianxin Xie, Wenfu Wang, Li Liu, and Dong Yu. 2026. Audio-DeepThinker: Progressive reasoning-aware reinforcement learning for high-quality chain-of-thought emergence in audio language models. *arXiv preprint arXiv:2604.18187*.

John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35.

Lars Hertel, Huy Phan, and Alfred Mertins. 2016. Comparing time and frequency domain for audio event recognition using deep learning. In *Proceedings of*

the 2016 International Joint Conference on Neural Networks (IJCNN), pages 3407–3411. IEEE.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97.

Yueran Hou, Peilei Jia, Zihan Sun, Qihang Lu, Wenbing Yang, Yingming Gao, Ya Li, and Jun Gao. 2026. Hello-Chat: Towards realistic social audio interactions. *arXiv preprint arXiv:2602.23387*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tiansheng Huang, Virat Shejwalkar, Oscar Chang, Milad Nasr, and Ling Liu. 2025. Rebellion: Noise-robust reasoning training for audio reasoning models. *arXiv preprint arXiv:2511.09682*.

Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, and Tomoki Toda. 2023. The singing voice conversion challenge 2023. In *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. 2026. TangoFlux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2207–2211.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, and Yuxuan Wang. 2025. DiTAR: Diffusion transformer autoregressive modeling for speech generation. In *Forty-second International Conference on Machine Learning*.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703, Marseille, France. European Language Resources Association.

Yidi Jiang, Qian Chen, Shengpeng Ji, Yu Xi, Wen Wang, Chong Zhang, Xianghu Yue, Shiliang Zhang, and Haizhou Li. 2025. UniCodec: Unified audio codec with single domain-adaptive codebook. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 19112–19124.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2415–2424.

Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. 2022. ISTFTNET: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6207–6211.

Harunori Kawano and Takeshi Sasaki. 2026. A human-inspired decoupled architecture for efficient audio representation learning. *Preprint*, arXiv:2603.26098.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021a. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International conference on machine learning*, pages 5530–5540. PMLR.

Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, and Seong-Whan Lee. 2021b. Fre-GAN: Adversarial frequency-consistent audio synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2197–2201.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. LibriTTS-R: A restored multi-speaker text-to-speech corpus. In *Proceedings of Interspeech 2023*, pages 5496–5500. ISCA.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020a. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 17022–17033.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020b. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A versatile diffusion model for audio synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. 2022. Efficient training of audio transformers with patchout. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2753–2757.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, and 1 others. 2026a. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 22688–22697.

Sonal Kumar, Prem Seetharaman, Ke Chen, Oriol Nieto, Jiaqi Su, Zhepei Wang, Rithesh Kumar, Dinesh Manocha, Nicholas J. Bryan, Zeyu Jin, and Justin Salamon. 2026b. TAC: Timestamped audio captioning. *arXiv preprint arXiv:2602.15766*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Ryan Langman, Ante Jukic, Kunal Dhawan, Nithin Rao Koluguri, and Jason Li. 2025. Spectral codecs: Improving non-autoregressive speech synthesis with spectrogram-based audio codecs. *Preprint*, arXiv:2406.05298.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. BigVGAN: A universal neural vocoder with large-scale training. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sang-Hoon Lee, Ha-Yeong Choi, and Seong-Whan Lee. 2025. PeriodWave: Multi-period flow matching for high-fidelity waveform generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Maxime Leiber, Yosra Marnissi, Axel Barrau, Sylvain Meignen, and Laurent Massoulié. 2025. Optimal adaptive time-frequency representation via differentiable short-time fourier transform. *IEEE Transactions on Signal Processing*, 73:5047–5059.

Chang Li, Kanglei Zhou, and Liyuan Wang. 2026a. PACE: Pretrained audio continual learning. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Hanzhao Li, Liumeng Xue, Haohan Guo, Xinfa Zhu, Yuanjun Lv, Lei Xie, Yunlin Chen, Hao Yin, and Zhifei Li. 2024a. Single-Codec: Single-codebook speech codec towards high-performance speech generation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, and Zhizheng Wu. 2025a. DualCodec: A low-frame-rate, semantically-enhanced neural audio codec for speech generation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Kang Li, Yan Song, Li-Rong Dai, Ian McLoughlin, Xin Fang, and Lin Liu. 2023a. AST-SED: An effective sound event detection method based on Audio Spectrogram Transformer. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tianhua Li, Chenda Li, Wei Wang, Xin Zhou, Xihui Chen, Jianqing Gao, and Yanmin Qian. 2026b. SLM-SS: Speech language model for generative speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 15432–15436.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. 2025b. Baichuan-Audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*.

Xiang Li, FanBu FanBu, Ambuj Mehrish, Yingting Li, Jiale Han, Bo Cheng, and Soujanya Poria. 2024b. Cm-tts: Enhancing real time text-to-speech synthesis efficiency through weighted samplers and consistency models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3777–3794.

Yadong Li, Guoxin Wu, Haiping Hou, and Biye Li. 2026c. UAF: A unified audio front-end LLM for full-duplex speech interaction. *arXiv preprint arXiv:2604.19221*.

Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. 2023b. HiFTNet: A fast high-quality neural vocoder with harmonic-plus-noise filter and inverse short time fourier transform. *Preprint*, arXiv:2309.09493.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger B. Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. 2024c. MERT: Acoustic music understanding model with large-scale self-supervised training. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ziwei Li, Lukuang Dong, Saierdaer Yusuyin, Xianyu Zhao, and Zhijian Ou. 2026d. Phonemes vs. projectors: An investigation of speech-language interfaces for LLM-based ASR. *arXiv preprint arXiv:2604.09332*.

Guan-Ting Lin, Chen Chen, Zhehuai Chen, and Hung-yi Lee. 2026a. Full-Duplex-Bench-v3: Benchmarking tool use for full-duplex voice agents under real-world disfluency. *arXiv preprint arXiv:2604.04847*.

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung yi Lee. 2026b. Full-Duplex-Bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 27–36.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-yi Lee. 2025a. Full-Duplex-Bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Xin Lin, Yang Zhang, and Shiyuan Wang. 2024. Mixed T-domain and TF-domain magnitude and phase representations for GAN-based speech enhancement. *Scientific Reports*, 14(1).

Yueqian Lin, Zhengmian Hu, Jayakumar Subramanian, Qinsi Wang, Nikos Vlassis, Hai Li, and Yiran Chen. 2025b. AsyncVoice agent: Real-time explanation for LLM planning and reasoning. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–4. IEEE.

Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D. Plumbley. 2024a. SemantiCodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing*, 18(8):1448–1461.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. 2024b. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*.

Tianqiao Liu, Xueyi Li, Hao Wang, Haoxuan Li, Zhichao Chen, Weiqi Luo, and Zitao Liu. 2026a. From text to talk: Audio-language model needs non-autoregressive joint training. *arXiv preprint arXiv:2509.20072*.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D. Plumbley, and Wenwu Wang. 2024c. Separate anything you describe. *IEEE Transactions on Audio, Speech, and Language Processing*, 33:458–471.

Yisi Liu, Nicholas Lee, and Gopala Anumanchipalli. 2026b. StyleStream: Real-time zero-shot voice style conversion. *arXiv preprint arXiv:2602.20113*.

Zefang Liu, Chenyang Zhu, Sangwoo Cho, and Shi-Xiong Zhang. 2026c. ReHear: Iterative pseudo-label refinement for semi-supervised speech recognition via audio large language models. *arXiv preprint arXiv:2602.18721*.

Zhijun Liu, Shuai Wang, Sho Inoue, Qibing Bai, and Haizhou Li. 2024d. Autoregressive diffusion transformer for text-to-speech synthesis. *arXiv preprint arXiv:2406.05551*.

LLM-Core Xiaomi. 2025. MiMo-Audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*.

Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, Chaoyou Fu, and 1 others. 2026. Omni-Diffusion: Unified multimodal understanding and generation with masked discrete diffusion. In *International Conference on Machine Learning (ICML)*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, and 1 others. 2026a. DeSTA2.5-Audio: Toward General-Purpose Large Audio Language Model with Self-Generated Cross-Modal

Alignment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Lie Lu, Hong-Jiang Zhang, and Hao Jiang. 2002. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516.

Yen-Ju Lu, Yashesh Gaur, Wei Zhou, Benjamin Muller, Jesus Villalba, Najim Dehak, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Srini Iyer, and Duc Le. 2026b. Latent speech-text transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yi Luo, Zhuo Chen, and Takuya Yoshioka. 2020. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50.

Yi Luo and Nima Mesgarani. 2019. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266.

Guobin Ma, Jixun Yao, Ziqian Ning, Yuepeng Jiang, Lingxin Xiong, Lei Xie, and Pengcheng Zhu. 2026a. Meanvc: Lightweight and streaming zero-shot voice conversion via mean flows. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17347–17351.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-CoT: Exploring chain-of-thought reasoning in large audio language model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–6. IEEE.

Ziyang Ma, Yinghao Ma, Yanqiao Ext Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2026b. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 564–572.

Potsawee Manakul, Woody Haosheng Gan, Martijn Bartelds, Guangzhi Sun, William Barr Held, and Diyi Yang. 2026. Scaling open discrete audio foundation models with interleaved semantic, acoustic, and text tokens. *CoRR*, abs/2602.16687.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alexander Smola. 2026. EmergentTTS-eval: Evaluating TTS models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *Advances in Neural Information Processing Systems*, 38.

Xinhao Mei, Gaël Le Lan, Haohe Liu, Zhaoheng Ni, Varun Nagaraja, Yang Liu, Yangyang Shi, and Vikas Chandra. 2026. SLAP: Scalable language-audio pretraining with variable-duration audio and multi-objective training. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 14517–14521.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2018. A multi-device dataset for urban acoustic scene classification. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pages 9–13.

Christoph Minixhofer, Ondřej Klejch, and Peter Bell. 2026. TTSDS2: Resources and benchmark for evaluating human-quality text to speech systems. In *International Conference on Learning Representations (ICLR)*.

Bingshen Mu, Xian Shi, Xiong Wang, Hexin Liu, Jin Xu, and Lei Xie. 2026. LLM-ForcedAligner: A non-autoregressive and accurate LLM-based forced aligner for multilingual and long-form speech. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25627–25638.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2023. Masked modeling duo: Learning representations by encouraging both networks to model the input. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Noboru Harada, and Nobutaka Ono. 2026. Rethinking masking strategies for masked prediction-based audio self-supervised learning. *Preprint*, arXiv:2603.23810.

Soha A. Nossier, Julie A. Wall, Mansour Moniri, Cornelius Glackin, and Nigel Cannings. 2020. A comparative study of time and frequency domain approaches to deep learning based speech enhancement. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Takuma Okamoto, Haruki Yamashita, Yamato Ohtani, Tomoki Toda, and Hisashi Kawai. 2023. WaveNeXt: ConvNeXt-based fast neural vocoder without ISTFT layer. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Cesar Pachon and Dora Ballesteros. 2026. FakeVoiceFinder: An open-source framework for synthetic and deepfake audio detection. *Big Data and Cognitive Computing*, 10(1):25.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Taesoo Park, Mungwi Jeong, Mingyu Park, Narae Kim, Junyoung Kim, Dohyun Park, Mujung Kim, Jisang Yoo, Hoyun Lee, Sanghoon Kim, and 1 others. 2025. BemaGANv2: A vocoder with superior periodicity capture for long-term audio generation. In *Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 0830–0834.

Jing Peng, Chenghao Wang, Yi Yang, Lirong Qian, Junjie Li, Yu Xi, Shuai Wang, and Kai Yu. 2026a. TASU2: Controllable CTC simulation for alignment and low-resource adaptation of speech LLMs. *arXiv preprint arXiv:2604.08384*.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, and 1 others. 2025a. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jeeweon Jung, and Shinji Watanabe. 2024. OWSM v3.1: Better and faster open whisper-style speech models based on E-Branchformer. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Eng Siong Chng. 2025b. FD-Bench: A full-duplex benchmarking pipeline designed for full duplex spoken dialogue systems. In *Proceedings of Interspeech 2025*. ISCA.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. 2026b. VibeVoice: Expressive podcast generation with next-token diffusion. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Niketa Penumajji. 2025. Deep learning for speech emotion recognition: A CNN approach utilizing Mel spectrograms. In *Proceedings of the 15th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 1–5. IEEE.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2025. EnCodecMAE: Leveraging neural codecs for universal audio representation learning. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Sergeevich Kudinov, and Jiansheng Wei. 2022. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621.

Aurian Quelennec, Pierre Chouteau, Geoffroy Peeters, and Slim Essid. 2025. Masked latent prediction and classification for self-supervised audio representation learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Qwen Team. 2025. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 28492–28518.

Alain Rakotomamonjy and Gilles Gasso. 2015. Histogram of gradients of time–frequency representations for audio scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):142–153.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. 2025. BirdSet: A large-scale dataset for audio classification in avian bioacoustics. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 2025, pages 29482–29520.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with SincNet. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, João Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993.

Pengyu Ren, Wenhao Guan, Kaidi Wang, Peijie Chen, Qingyang Hong, and Lin Li. 2025. ReFlow-VC: Zero-shot voice conversion based on rectified flow and speaker feature optimization. In *Proceedings of the 26th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations (ICLR)*.

Fabian Retkowski, Maike Züfle, Andreas Sudmann, Dinah Pfau, Shinji Watanabe, Jan Niehues, and Alex Waibel. 2025. Summarizing speech: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27263–27294.

Douglas A. Reynolds and Richard C. Rose. 1995. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83.

Simon Rouard, Manu Orsini, Axel Roebel, Neil Zeghidour, and Alexandre Défossez. 2026. Continuous audio language models. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hildegard Kuehne, Rogério Feris, and James R. Glass. 2025. Omni-R1: Do you really need audio to fine-tune your audio LLM? In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7.

Rajarshi Roy, Jonathan Raiman, Sang-Gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. 2026. PersonaPlex: Voice and role control for full duplex conversational speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16137–16141.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Kohei Saijo, Wangyou Zhang, Samuele Cornell, Robin Scheibler, Chenda Li, Zhaoheng Ni, Anurag Kumar, Marvin Sach, Yihui Fu, Wei Wang, and 1 others. 2025. Interspeech 2025 URGENT speech enhancement challenge. In *Proc. Interspeech 2025*, pages 858–862.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 2025, pages 84929–84964.

Eric D. Scheirer and Malcolm Slaney. 1997. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1331–1334.

Siavash Shams, Sukru Samet Dindar, Xilin Jiang, and Nima Mesgarani. 2024. SSAMBA: Self-supervised audio representation learning with Mamba state space model. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 1053–1059.

Maohao Shen, Tejas Jayashankar, Osama Hanna, Naoyuki Kanda, Yancheng Wang, Kateřina Žmolíková, Ruiming Xie, Niko Moritz, Anfeng Xu, Yashesh Gaur, Gregory Wornell, Qing He, and Jilong Wu. 2026. GSRM: Generative speech reward model for speech RLHF. *arXiv preprint arXiv:2602.13891*.

Zhichao Sheng, Shilin Zhou, Chen Gong, and Zhenghua Li. 2026. Think smart, not hard: Difficulty adaptive reasoning for large audio language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 32771–32790.

Yemin Shi, Yu Shu, Siwei Dong, Guangyi Liu, Jaward Sesay, Jingwen Li, and Zhiting Hu. 2025. Voila: Voice-language foundation models for real-time autonomous interaction and voice role-play. *arXiv preprint arXiv:2505.02707*.

Hubert Siuzdak. 2024. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. 2024. SNAC: Multi-scale neural audio codec. In *Proceedings of the NeurIPS Workshop on AI-Driven Speech, Music, and Sound Generation*.

Yuhan Song, Linhao Zhang, Chuhan Wu, Aiwei Liu, Wei Jia, Houfeng Wang, and Zhou Xiao. 2026. Stabletoken: A noise-robust semantic speech tokenizer for resilient speechLLMs. In *The Fourteenth International Conference on Learning Representations*.

Sripathi Sridhar and Mark Cartwright. 2025. Compositional audio representation learning. In *Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

StepFun Audio Team. 2025. Step-Audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao,

Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2025. Step-Audio-R1 technical report. *arXiv preprint arXiv:2511.15848*.

Jinchuan Tian, Sang gil Lee, Zhifeng Kong, Sreyan Ghosh, Arushi Goel, Chao-Han Huck Yang, Wenliang Dai, Zihan Liu, Hanrong Ye, Shinji Watanabe, Mohammad Shoeybi, Bryan Catanzaro, Rafael Valle, and Wei Ping. 2026a. UALM: Unified audio language model for understanding, generation and reasoning. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Jinchuan Tian, Haoran Wang, Bo-Hao Su, Chien yu Huang, Qingzheng Wang, Jiatong Shi, William Chen, Xun Gong, Siddhant Arora, Chin-Jou Li, Masao Someki, Takashi Maekaku, Keita Goto, Yusuke Shinohara, Jin Sakuma, Chao-Han Huck Yang, and Shinji Watanabe. 2026b. Bagpiper: Solving open-ended audio tasks via rich captions. *arXiv preprint arXiv:2602.05220*.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2026c. AudioX: A unified framework for anything-to-audio generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zeyue Tian, Binxin Yang, Zhaoyang Liu, Jiexuan Zhang, Ruibin Yuan, Hubery Yin, Qifeng Chen, Chen Li, Jing Lv, Wei Xue, and Yike Guo. 2026d. AudioOmni: Extending multi-modal understanding to versatile audio generation and editing. *arXiv preprint arXiv:2604.10708*.

Wei-Cheng Tseng, Xuanru Zhou, Mingyue Huo, Yiwen Shao, Hao Zhang, and Dong Yu. 2026. Revisiting audio-language pretraining for learning general-purpose audio representation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34244–34263.

Ludovic Tuncay, Etienne Labbe, Emmanouil Benetos, and Thomas Pellegrini. 2025. Audio-JEPA: Joint-embedding predictive architecture for audio representation learning. *Preprint*, arXiv:2507.02915.

Joseph Turian, Björn W. Schuller, Dorien Herremans, Katrin Kirchhoff, L. Paola García-Perera, and Philippe Esling, editors. 2021. *HEAR: Holistic Evaluation of Audio Representations*, volume 166. PMLR.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, and 1 others. 2022. HEAR: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. In *Proceedings of the ISCA Speech Synthesis Workshop (SSW)*, page 125.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6306–6315.

Lester Phillip Violeta, Wen-Chin Huang, and Tomoki Toda. 2025. Serenade: A singing style conversion framework based on audio infilling. In *Proceedings of the 33rd European Signal Processing Conference (EUSIPCO)*, pages 411–415. IEEE.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, and 5 others. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

Zixiang Wan, Guochang Zhang, Yifeng He, and Jianqiang Wei. 2025. SpecTokenizer: A lightweight streaming codec in the compressed spectrum domain. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and massively multilingual speech translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. 2026a. Closing the modality reasoning gap for speech large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18821–18835.

Dingdong Wang, Junan Li, Mingyu Cui, Dongchao Yang, Xueyuan Chen, and Helen Meng. 2025a. Speech discrete tokens or continuous features? a comparative analysis for spoken language understanding in speechllms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24924–24935.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen M. Meng. 2026b. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen M. Meng. 2026c. EmotionThinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

He Wang, Linhan Ma, Dake Guo, Xiong Wang, Lei Xie, Jin Xu, and Junyang Lin. 2025b. ContextASR-Bench: A massive contextual speech recognition benchmark. *Preprint*, arXiv:2507.05727.

Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. 2026d. Listening between the frames: Bridging temporal gaps in large audio-language models. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, pages 26233–26241. AAAI Press.

Hui Wang, Cheng Liu, Junyang Chen, Haoze Liu, Yuhang Jia, Shiwan Zhao, Jiaming Zhou, Haoqin Sun, Hui Bu, and Yong Qin. 2026e. TTA-Bench: A comprehensive benchmark for evaluating text-to-audio models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33512–33520.

Hui Wang, Jinghua Zhao, Junyang Cheng, Cheng Liu, Yuhang Jia, Haoqin Sun, Jiaming Zhou, and Yong Qin. 2026f. AudioEval: Automatic dual-perspective and multi-dimensional evaluation of text-to-audio-generation. *Preprint*, arXiv:2510.14570.

Jieyi Wang, Yazhe Niu, Dexuan Xu, and Zhongyu Wei. 2026g. Listen, pause, and reason: Toward perception-grounded hybrid reasoning for audio understanding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 35653–35671.

Mason Long Wang and Cheng-Zhi Anna Huang. 2026. Latent fourier transform. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, and 10 others. 2026h. ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95:101825.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2025c. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *International Conference on Machine Learning (ICML)*, pages 63345–63354. PMLR.

Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. 2023. TF-GridNet: Integrating full- and sub-band modeling for speech separation. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 31:3221–3236.

Ziqian Wang, Xianjun Xia, Xinfa Zhu, and Lei Xie. 2025d. U-SAM: An audio language model for unified speech, audio, and music understanding. In *Proceedings of the 26th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. SUPERB: Speech processing universal PERformance benchmark. In *Proc. Interspeech 2021*.

Shu wen Yang, Ming Tu, Andy T. Liu, Xinghua Qu, Hung yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2026. ParaS2S: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *International Conference on Learning Representations (ICLR)*.

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. 2026. AudSemThinker: Enhancing audio-language models through reasoning over semantics of sound. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Kevin Wilkinghoff, Sarthak Yadav, and Zheng-Hua Tan. 2026. Temporal pooling strategies for training-free anomalous sound detection with self-supervised audio embeddings. *arXiv preprint arXiv:2603.04605*.

Daiqing Wu, Xuan Zhang, Dongbao Yang, Jiashu Yao, Longfei Chen, Qingsong Liu, Sicheng Zhao, Can Ma, Yangyang Kang, and Yu Zhou. 2026a. Echo: Towards advanced audio comprehension via audio-interleaved reasoning. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Haibin Wu, Ho-Lam Chung, Yi-Cheng Lin, Yuan-Kuei Wu, Xuanjun Chen, Yu-Chi Pai, Hsiu-Hsuan Wang, Kai-Wei Chang, Alex Liu, and Hung-yi Lee. 2024. Codec-SUPERB: An in-depth analysis of sound codec models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10330–10348. Association for Computational Linguistics.

Haibin Wu, Bach Viet Do, Naveen Suda, Julian Chan, Madhavan CR, Gene-Ping Yang, Yi-Chiao Wu, Naoyuki Kanda, Yossef Adi, Xin Lei, and 1 others. 2026b. T-Mimi: A transformer-based mimi decoder for real-time on-phone TTS. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16992–16996.

Shih-Lun Wu, Ge Zhu, Juan-Pablo Caceres, Cheng-Zhi Anna Huang, and Nicholas J. Bryan. 2026c. Stemphonic: All-at-once flexible multi-stem music generation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 15067–15071.

Shilong Wu. 2025. M3SD: Multi-modal, multi-scenario and multi-language speaker diarization dataset. *Preprint*, arXiv:2506.14427.

Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. 2025a. FireRedTTS-2: Towards long conversational speech generation for podcast and chatbot. *arXiv preprint arXiv:2509.02020*.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025b. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 764–791.

Yuan Xie, Jiaqi Song, Guang Qiu, Xianliang Wang, Ming Lei, Jie Gao, and Jie Wu. 2026. Rethinking entropy allocation in LLM-based ASR: Understanding the dynamics between speech encoders and LLMs. *arXiv preprint arXiv:2604.08003*.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025c. Audio-Reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 23829–23851. Association for Computational Linguistics.

Zhifei Xie and Changqiao Wu. 2024. Mini-Omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Jilan Xu, Carl Thomé, Danijela Horak, Weidi Xie, and Andrew Zisserman. 2026. Scaling audio-text retrieval with multimodal large language models. *arXiv preprint arXiv:2602.18010*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). Technical report, University of Edinburgh, The Centre for Speech Technology Research (CSTR).

Kohei Yamamoto and Kosuke Okusa. 2026. AaSP: Aliasing-aware self-supervised pre-training for audio spectrogram transformers. *IEEE Transactions on Audio, Speech and Language Processing*. Early access.

Yongpeng Yan, Yanan Li, Qiyang Xiao, and Yanzhen Ren. 2026. PRoADS: Provably secure and robust audio diffusion steganography with latent optimization and backward euler inversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13597–13601.

Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. 2024. UniAudio 1.5: Large language model-driven audio codec is a few-shot audio task learner. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023a. HiFi-Codec: Group-residual vector quantization for high fidelity audio codec. *Preprint*, arXiv:2305.02765.

Dongchao Yang, Yuxin Xie, Yuguo Yin, Zheyu Wang, Xiaoyu Yi, Gongxi Zhu, Xiaolong Weng, Zihan Xiong, Yingzhe Ma, Dading Cong, Jingliang Liu, Zihang Huang, Jinghan Ru, Rongjie Huang, Haoran Wan, Peixu Wang, Kuoxi Yu, Helin Wang, Liming Liang, and 10 others. 2026a. HeartMuLa: A family of open sourced music foundation models. *arXiv preprint arXiv:2601.10547*.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023b. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733.

Fei Yang, Xuanfan Ni, Renyi Yang, Jiahui Geng, Qing Li, Chenyang Lyu, Yichao Du, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2026b. LongSpeech: A scalable benchmark for transcription, translation and understanding in long speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 15547–15551.

Hao Yang, Minghan Wang, Tongtong Wu, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. 2026c. Resurfacing paralinguistic awareness in large audio language models. *arXiv preprint arXiv:2603.11947*.

Ke Yang, Bolin Chen, Yuejie Li, Yueying Hua, Jianhao Nie, Yueping He, Bowen Li, and Chengjun Mao. 2026d. Attention-guided evidence grounding for spoken question answering. *arXiv preprint arXiv:2603.16292*.

Xiaoyu Yang, Yifan Yang, Zengrui Jin, Ziyun Cui, Wen Wu, Baoxiang Li, Chao Zhang, and Phil Woodland. 2026e. SPEAR: A unified SSL framework for learning speech and audio representations. *Preprint*, arXiv:2510.25955.

Zhengdong Yang, Shuichiro Shimizu, Yahan Yu, and Chenhui Chu. 2025. When large language models meet speech: A survey on integration approaches. *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20298–20315.

Zhen Ye, Wei Xue, Xu Tan, Jie Chen, Qifeng Liu, and Yike Guo. 2023. Comospeech: One-step speech and singing voice synthesis via consistency model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1831–1839.

Hao Yen, Pin-Jui Ku, Ante Jukić, and Sabato Marco Siniscalchi. 2026. MDM-ASR: Bridging accuracy and efficiency in ASR with diffusion-based non-autoregressive decoding. *arXiv preprint arXiv:2602.18952*.

Jun-Wei Yeow, Ee-Leng Tan, Santi Peksi, and Woon-Seng Gan. 2025. Environmental acoustic intelligence

through sound event localization and detection: a review. *npj Acoustics*, 1(1):31.

Xiaofeng Yu, Jiaheng Dong, Jean Honorio, Abhirup Ghosh, Hong Jia, and Ting Dang. 2026. Disentangling reasoning in large audio-language models for ambiguous emotion prediction. *arXiv preprint arXiv:2603.08230*.

Jiajun Yuan, Xiaochen Wang, Yuhang Xiao, Yulin Wu, Chenhao Hu, and Xueyang Lv. 2025. STSR: High-fidelity speech super-resolution via spectral-transient context modeling. *Preprint*, arXiv:2509.03913.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Zhuo Le, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Brook Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger Dannenberg, Wenhu Chen, Gus Xia, and 6 others. 2023. MARBLE: Music audio representation benchmark for universal evaluation. In *Advances in Neural Information Processing Systems*, volume 36, pages 39626–39647. Curran Associates, Inc.

Goksenin Yuksel, Pierre Guetschel, Michael Tangermann, Marcel van Gerven, and Kiki van der Heijden. 2025. WavJEPA: Semantic learning unlocks robust audio foundation models for raw waveforms. *arXiv preprint arXiv:2509.23238*.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. SoundStream: An end-to-end neural audio codec. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. 2021. LEAF: A learnable frontend for audio classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Proc. Interspeech 2019*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Dan Zhang, Yishu Lei, Jing Hu, Shuwei He, Songhe Deng, Xianlong Luo, Danxiang Zhu, Shikun Feng, Rui Liu, Jingzhou He, Yu Sun, Hua Wu, and Haifeng Wang. 2026a. Eureka-Audio: Triggering audio intelligence in compact language models. *arXiv preprint arXiv:2602.13954*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773. Association for Computational Linguistics.

Honglei Zhang, Yuting Chen, Chenpeng Hu, Siyue Zhang, and Yilei Shi. 2026b. ReasonAudio: A benchmark for evaluating reasoning beyond matching in text-audio retrieval. *Preprint*, arXiv:2605.03361.

Junbo Zhang, Heinrich Dinkel, Yadong Niu, Chenyu Liu, Si Cheng, Anbei Zhao, and Jian Luan. 2025. X-ARES: A comprehensive framework for assessing audio encoder performance. In *Proc. Interspeech 2025*, pages 4868–4872.

Qian Zhang, Yuqin Cao, Yixuan Gao, and Xiongkuo Min. 2026c. VidAudio-Bench: Benchmarking V2A and VT2A generation across four audio categories. *Preprint*, arXiv:2604.10542.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. SpeechTokenizer: Unified speech tokenizer for speech language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yizhou Zhang, Yuan Gao, Wangjin Zhou, Zicheng Yuan, Keisuke Imoto, and Tatsuya Kawahara. 2026d. SONAR: Self-distilled continual pre-training for domain adaptive audio representation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 14752–14756.

Zhisheng Zhang, Xiang Li, Yixuan Zhou, Jing Peng, Shengbo Cai, Guoyang Zeng, and Zhiyong Wu. 2026e. UniSRCodec: Unified and low-bitrate single codebook codec with sub-band reconstruction. *arXiv preprint arXiv:2601.02776*.

Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhen-Hua Ling, and Tomoki Toda. 2020. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. In *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pages 80–98.

Qixi Zheng, Yuxiang Zhao, Tianrui Wang, Wenxi Chen, Kele Xu, Yikang Li, Qinyuan Chen, Xipeng Qiu, Kai Yu, and Xie Chen. 2026. X-VC: Zero-shot streaming voice conversion in codec space. *arXiv preprint arXiv:2604.12456*.

Jiaming Zhou, Hongjie Chen, Shiwan Zhao, Jian Kang, Jie Li, Enzhi Wang, Yujie Guo, Haoqin Sun, Hui Wang, Aobo Kong, Yong Qin, and Xuelong Li. 2026a. DIFFA: Large language diffusion models can listen and understand. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, pages 35112–35120. AAAI Press.

Jiaming Zhou, Xuxin Cheng, Shiwan Zhao, Yuhang Jia, Cao Liu, Ke Zeng, Xunliang Cai, and Yong Qin. 2026b. DIFFA-2: A practical diffusion large language model for general audio understanding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 4789–4805.

Wei Zhou, Junteng Jia, Leda Sari, Jay Mahadeokar, and Ozlem Kalinli. 2025. Cjst: Ctc compressor based joint speech and text training for decoder-only asr. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xuanru Zhou, Jiachen Lian, Henry Hong, Xinyi Yang, and Gopala Anumanchipalli. 2026c. Speech world model: Causal state-action planning with explicit reasoning for speech. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Xuanru Zhou, Yiwen Shao, Wei-Cheng Tseng, and Dong Yu. 2026d. Unlocking strong supervision: A data-centric study of general-purpose audio pretraining methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24780–24791.

Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. 2026. OmniVoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*.

Zhiyuan Zhu, Yu Zhang, Wenxiang Guo, Changhao Pan, and Zhou Zhao. 2025. Asaudio: A survey of advanced spatial audio research. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 417–442.

A Overview of Taxonomy Framework

To provide a structured overview of the current research landscape, we summarize the classification framework proposed in this paper as shown in Figure 3. This systematic categorization organizes existing literature across key design dimensions, including audio representation learning, integration mechanisms, and downstream task modeling paradigms.

B Evaluation Methods for Audio Representation Models

Evaluating audio representations centers on assessing their latent space features. Models' representational and generalization capabilities are typically quantified based on feature independence, transferability, and acoustic fidelity across three mainstream paradigms:

B.1 Frozen Feature Extraction (FFE)

FFE freezes the encoder as a static feature extractor. It encompasses two methods: linear probing (LP), which trains a shallow classifier to test the linear separability of original features (wen Yang et al., 2021; Turian et al., 2022; Wu et al., 2024); and zero-shot (ZS) evaluation, which performs direct inference via cross-modal alignment without downstream training, verifying the representations' intrinsic semantic understanding and generalization (Elizalde et al., 2023; Radford et al., 2023; Dinkel et al., 2026b; Mei et al., 2026).

B.2 Fine-tuning (FT)

FT unfreezes the pre-trained model for joint optimization on specific downstream tasks (Shams et al., 2024; Alex et al., 2024). By utilizing the model as an initialization backbone adapted to target data distributions, this approach explores the transferability limits and performance upper bounds, intuitively reflecting its practical deployment potential in complex tasks.

B.3 Direct Reconstruction (DR)

DR quantifies feature fidelity by restoring original waveforms from compressed representations via a decoder. Targeting reconstruction models like autoencoders (Wang and Huang, 2026; Dinkel et al., 2026a), neural codecs (Jiang et al., 2025; Wan et al., 2025), and vocoders (Lee et al., 2025; Park et al., 2025), it evaluates information compression efficiency and acoustic integrity by measuring the preservation of underlying physical attributes like timbre and phase.

C Summary of Audio Downstream Tasks

Audio downstream tasks and mainstream benchmarks are summarized in Table 2

C.1 Understanding Tasks

Audio understanding maps audio inputs to textual outputs, aligning condensed acoustic representations with textual semantic spaces to reflect human cognition. In multimodal paradigms, these tasks leverage text or video *prompts* as auxiliary instructions to guide specific auditory comprehension. These tasks are typically evaluated using diverse benchmarks, such as LibriSpeech (Panayotov et al., 2015) for automatic speech recognition, MMSU (Wang et al., 2026b) for audio question answering, and Clotho (Drossos et al., 2020) for automated captioning. Classification (Cls) fundamentally encompasses tasks such as acoustic scene classification (ASC), sound event detection (SED),

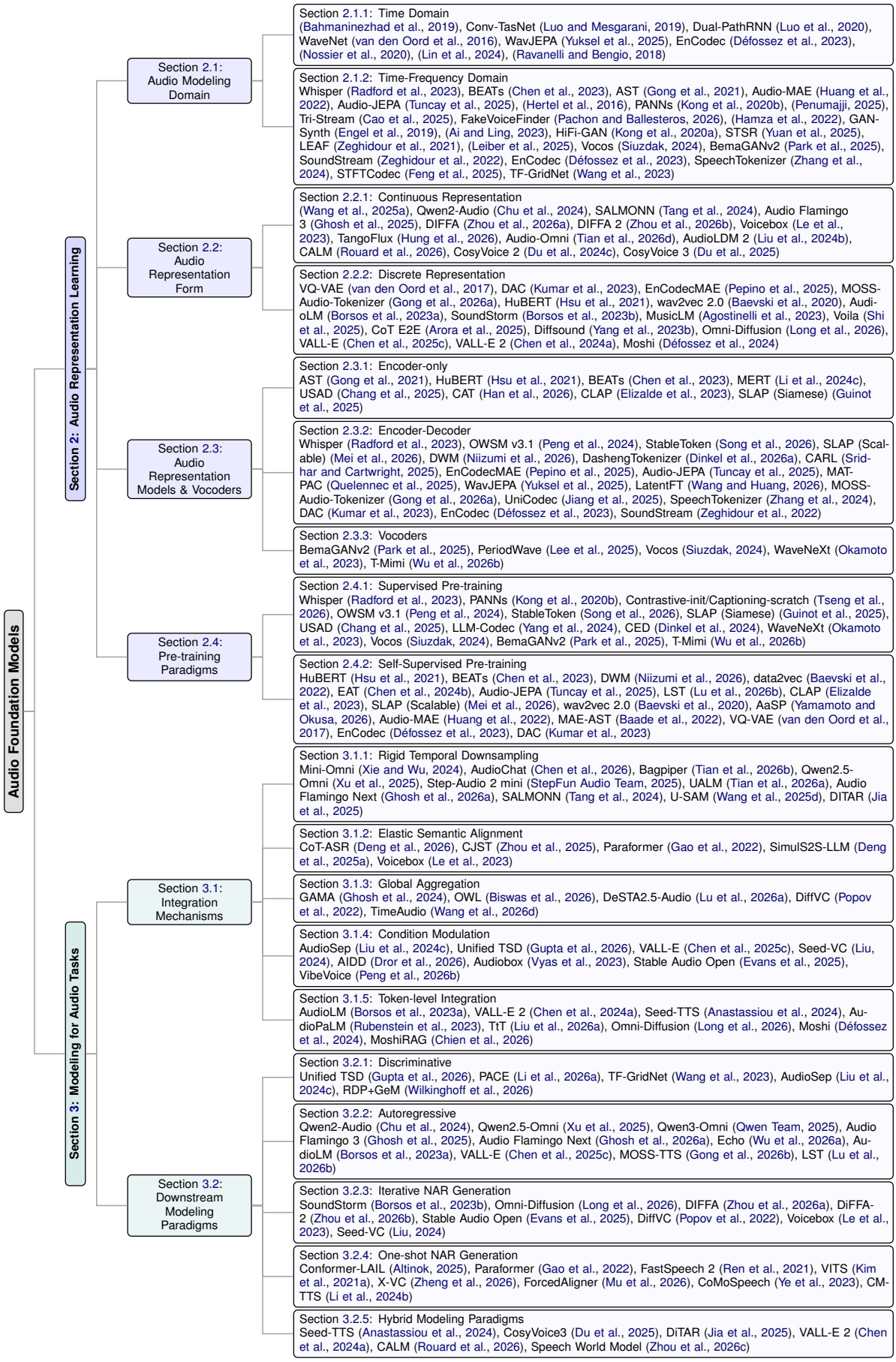


Figure 3: Taxonomy of Audio Representation Learning and Downstream Task Modeling Paradigms.

Task	Benchmarks
Understanding Tasks	
Cls *Classification*	AudioSet(Gemmeke et al., 2017), TUT Urban Acoustic Scenes 2018(Mesaros et al., 2018), X-ARES(Zhang et al., 2025), BirdSet(Rauch et al., 2025), ASVspoof 5(Wang et al., 2026h)
ASR *Automatic Speech Recognition*	Librispeech(Panayotov et al., 2015), GigaSpeech(Chen et al., 2021a), FLEURS(Conneau et al., 2023), ContextASR-Bench(Wang et al., 2025b)
S2TT *Speech-to-Text Translation*	MuST-C(Di Gangi et al., 2019), CoVoST 2(Wang et al., 2021), LongSpeech(Yang et al., 2026b)
AQA *Audio Question Answering*	MMAR(Ma et al., 2026b), MMSU(Wang et al., 2026b), MMAU-Pro(Kumar et al., 2026a), MMAU(Sakshi et al., 2025)
AAC *Automated Audio Captioning*	AudioCaps(Kim et al., 2019), AudioSetCaps(Bai et al., 2025), Clotho(Drossos et al., 2020), BRACE(Guo et al., 2026a)
ATR *Audio-Text Retrieval*	ReasonAudio(Zhang et al., 2026b), MARBLE(Yuan et al., 2023)
SD *Speaker Diarization*	SDBench(Durmus et al., 2025), M3SD(Wu, 2025)
Generation Tasks	
TTS *Text to Speech*	TTSDS2(Minixhofer et al., 2026), EmergentTTS-Eval(Manku et al., 2026), Seed-TTS-Eval(Anastassiou et al., 2024), LibriTTS(Zen et al., 2019), LibriTTS-R(Koizumi et al., 2023)
TTA *Text to Audio*	TTA-Bench(Wang et al., 2026e), AudioEval(Wang et al., 2026f)
VTA *Video to Audio*	Vggsound(Chen et al., 2020), VidAudio-Bench(Zhang et al., 2026c)
TTM *Text to Music*	FakeMusicCaps(Comanducci et al., 2025), MusicCaps(Agostinelli et al., 2023)
ZVC *Zero-shot Voice Cloning*	ClonEval(Christop et al., 2025), VCTK(Yamagishi et al., 2019)
T2ST *Text-to-Speech Translation*	–
Conversion Tasks	
VC *Voice Conversion*	VCC2020(Zhao et al., 2020), SVCC2023(Huang et al., 2023)
SE *Speech Enhancement*	DNS Challenge(Dubey et al., 2024), URGENT SE Challenge(Saijo et al., 2025)
SS *Speech Separation*	WSJ0-mix(Hershey et al., 2016), LibriMix(Cosentino et al., 2020)
AI *Audio Inpainting*	–
S2ST *Speech-to-Speech Translation*	CVSS(Jia et al., 2022), Emphassess(de Seyssel et al., 2024)
Interaction Tasks	
FSD *Full-duplex Spoken Dialogue*	Full-Duplex-Bench(Lin et al., 2025a), Full-Duplex-Bench-v2(Lin et al., 2026b), Full-Duplex-Bench-v3(Lin et al., 2026a), FD-Bench(Peng et al., 2025b)
SDS *Spoken Dialogue System*	ParaS2S(wen Yang et al., 2026), SD-Eval(Ao et al., 2024)

Table 2: Overview of Speech and Audio Tasks and Representative Benchmarks.

speaker identification (SID), speech emotion recognition (SER), and other tasks that output classification labels.

C.2 Generation Tasks

Conditioned on textual, visual, or acoustic signals, audio generation tasks employ deep generative models (*e.g.*, autoregressive or diffusion) to synthesize novel, high-fidelity, diverse, and highly controllable audio content, encompassing human speech, natural sound effects, and music. Key evaluation benchmarks include Seed-TTS-Eval (Anastassiou et al., 2024) for text-to-speech, TTA-Bench (Wang et al., 2026e) for text-to-audio generation, and VidAudio-Bench (Zhang et al., 2026c) for video-to-audio synthesis. Additionally, T2ST and S2TT can share benchmarks.

C.3 Conversion Tasks

Audio conversion tasks leverage source audio as a semantic and temporal backbone. Guided by text, audio conditions, or implicit objectives, they perform feature disentanglement and attribute reconstruction within acoustic latent spaces, yielding signals that preserve source content fidelity while successfully capturing target attributes. Common tasks like voice conversion, speech enhancement, and speech-to-speech translation are evaluated on benchmarks such as SVCC2023 (Huang et al., 2023), DNS Challenge (Dubey et al., 2024), and CVSS (Jia et al., 2022). Particularly, Audio inpainting lacks dedicated benchmarks, relying on synthetic gaps across general datasets.

C.4 Interactive Tasks

Audio interaction tasks facilitate real-time, bidirectional, and multi-turn human-machine communication via continuous audio streams. Driven by natural speech and enriched with paralinguistic and emotional features, they leverage full-duplex mechanisms for seamless listening and speaking, ultimately realizing fluent spoken dialogues. To assess interactive capabilities, researchers employ spoken dialogue system benchmarks like ParaS2S (wen Yang et al., 2026) and full-duplex evaluation suites such as Full-Duplex-Bench-v3 (Lin et al., 2026a).

D Summary of Audio Representation Models

D.1 Encoder-Only

Model	Scale	Audio Modeling Domain	Audio Rep. Form	Pre-training	Architecture	Data			Tasks	Eval
						SP	SO	MU		
PANNs (Kong et al., 2020b)	∼81M	T+TF	Cont.	SL	CNN	✓	✓	✓	CLS	LP, FT
PASE+ (Ravanelli et al., 2020)	–	T	Cont.	SSL (Pred+dist)	CNN+RNN	✓			ASR	LP, FT
wav2vec 2.0 (Baevski et al., 2020)	95M/317M	T	Cont.	SSL (Cont+Pred)	CNN+Transformer	✓			ASR, CLS	FT
AST (Gong et al., 2021)	86M	TF	Cont.	SL	ViT	✓	✓	✓	CLS	FT
HuBERT (Hsu et al., 2021)	95M/317M/964M	T	Cont.	SSL (Pred)	CNN+Transformer	✓			ASR	FT
PaSST (Koutini et al., 2022)	87M	TF	Cont.	SL	ViT	✓	✓	✓	CLS	FT
WavLM (Chen et al., 2022)	94.70M/316.62M	T	Cont.	SSL (Pred)	CNN+Transformer	✓			ASR, SS, CLS, SD	LP, FT
data2vec (Baevski et al., 2022)	∼90M/∼300M	T	Cont.	SSL (dist+Pred)	CNN+Transformer	✓			ASR, CLS	FT
BEST-RQ (Chiu et al., 2022)	0.6B	TF	Cont.	SSL (Pred)	Conformer	✓			ASR	FT
CLAP (Elizalde et al., 2023)	1.9B	TF	Cont.	SSL (Cont)	PANNs+BERT	✓	✓	✓	CLS	ZS, LP, FT
M2D (Niizumi et al., 2023)	86M	TF	Cont.	SSL (Pred+dist)	ViT	✓	✓	✓	CLS	LP+FT
BEATs (Chen et al., 2023)	∼90M	TF	Cont.	SSL (Pred)	ViT	✓	✓	✓	CLS	FT
MERT (Li et al., 2024c)	95M/330M	T	Cont.	SSL (Pred+Reco)	CNN+Transformer			✓	CLS	LP
CED (Dinkel et al., 2024)	5.5M/10M/22M/86M	TF	Cont.	SSL (Reco) + SL (dist)	ViT		✓		CLS	LP
DTF-AT (Alex et al., 2024)	69M	TF	Cont.	SL	ViT	✓	✓	✓	CLS	FT
SSAMBA (Shams et al., 2024)	7M/26M/99M	TF	Cont.	SSL (Cont+Reco)	Bi-Mamba	✓	✓	✓	CLS	FT
SLAP (Siamese) (Guinot et al., 2025)	193M	TF	Cont.	SSL (self-dist)	HTS-AT			✓	ATR, CLS	LP, ZS
USAD (Chang et al., 2025)	24/94/330M	TF	Cont.	SL (dist)	Transformer	✓	✓	✓	ASR, PR, SD, CLS	LP, FT
GLAP (Dinkel et al., 2026b)	∼90M	TF	Cont.	SSL (Cont)	Dasheng + Sonar + MLP	✓	✓	✓	ATR, CLS	ZS
SONAR (Zhang et al., 2026d)	∼90M	TF	Cont.	SSL (Pred+self-dist)	BEATs iter3+	✓	✓	✓	CLS	LP, FT
SPEAR (Yang et al., 2026e)	94M/327M/600M	TF	Cont.	SSL (Pred) + SL (dist)	Zipformer	✓	✓	✓	ASR, PR, CLS, SE, SS, VC	LP, FT
Contrastive-init/Captioning-scratch (Tseng et al., 2026)	∼70M	TF	Cont.	SL	Zipformer-M	✓	✓	✓	CLS, ATR, ACC, AQA	LP
CAT (Han et al., 2026)	–	T	Cont.	SSL (self-dist)	Transformer	✓	✓	✓	CLS, ACC	FT
UTS-Multi-Task (Zhou et al., 2026d)	–	TF	Cont.	SL	Zipformer-M	✓	✓	✓	CLS, ATR, ACC, AQA	LP
HEAR (Kawano and Sasaki, 2026)	15M	TF	Cont.	SSL (Pred)	Transformer		✓		CLS	LP, FT

Audio Rep. Form: **Cont.**: Continuous, **Disc.**: Discrete.
Pre-training: **SSL**: Self-supervised learning, **SL**: Supervised learning, **USL**: Unsupervised learning. **Pred**: Predictive, **Cont**: Contrastive, **Reco**: Reconstructive, **dist**: distillation, **self-dist**: self-distillation.
Data: **SP**: Speech, **SO**: Sound, **MU**: Music.
Tasks: **CLS**: Classification, **ASR**: Automatic Speech Recognition, **SD**: Speaker Diarization, **PR**: Phoneme Recognition, **AQA**: Audio Question Answering, **ATR**: Audio-Text Retrieval, **AAC**: Automated Audio Captioning, **SE**: Speech Enhancement, **SS**: Speech Separation, **VC**: Voice Conversion.
Eval: **LP**: Linear Probing, **ZS**: Zero-shot, **FT**: Fine-tuning.

Table 3: Overview of Encoder-only Audio Representation Learning Models

D.2 Encoder-Decoder

Model	Scale	Audio Modeling Domain	Audio Rep. Form	Pre-training	Architecture			Data			Tasks	Eval
					Encoder	Quantizer	Decoder	SP	SO	MU		
Cross-modal Encoder-Decoder												
Whisper (Radford et al., 2023)	39M/74M/244M/769M/1.55B	TF	Cont	SL	Transformer	-	Transformer	✓			ASR, S2TT	ZS
OWSM v3.1 (Peng et al., 2024)	101M/367M/1.02B	TF	Cont	SL	E-Branchformer	-	Transformer	✓			ASR, S2TT, SLU, CLS	LP
StableToken (Song et al., 2026)	320M	TF	Disc	SL	Transformer	Voting-LFQ	Transformer	✓			ASR, CLS, TTS	FT, DR
SLAP (Scalable) (Mei et al., 2026)	∼2.6b	TF	Cont	SSL (Cont+Pred)	Audio Encoder + Text Encoder	-	Transformer		✓		ATR, CLS, AAC	ZS, LP, FT
Autoencoders												
MAE-AST (Baade et al., 2022)	-	TF	Cont	SSL (Reco)	ViT	-	ViT	✓	✓	✓	CLS	FT
Audio-MAE (Huang et al., 2022)	22M/86M/304M	TF	Cont	SSL (Reco)	ViT	-	Transformer	✓	✓	✓	CLS, REC	FT, DR
EnCodecMAE (Pepino et al., 2025)	43.6M/86.6/261.6M	TF	Cont	SSL (dist+Pred)	Transformer	-	Transformer	✓	✓	✓	CLS, ASR	LP
EAT (Chen et al., 2024b)	88M	TF	Cont	SSL (dist)	ViT + CNN	-	CNN	✓	✓	✓	CLS	LP, FT
CARL (Sridhar and Cartwright, 2025)	-	TF	Cont	SL/SSL (Reco)	AudioMAE + Slot Transformer	-	MLP + CNN + Transformer		✓		CLS	LP
MATPAC (Quelennec et al., 2025)	∼86M	TF	Cont	SSL (Pred+dist)	AST	-	Transformer	✓	✓	✓	CLS	LP
Audio-JEPA (Tuncay et al., 2025)	85.4M	TF	Cont	SSL (dist)	ViT	-	ViT	✓	✓	✓	CLS	LP
WavJEPA (Yuksel et al., 2025)	∼90/300M	T	Cont	SSL (dist)	Wav2Vec 2.0 + ViT	-	ViT	✓	✓	✓	CLS	LP
AaSP (Yamamoto and Okusa, 2026)	87M	TF	Cont	SSL (Pred+Cont)	ViT	-	Cross-attention Blocks	✓	✓	✓	CLS	LP, FT
DashengTokenizer (Dinkel et al., 2026a)	∼800M	TF	Cont	SSL (Reco)	MiDasheng + 2D CNN	-	Vocos	✓	✓	✓	CLS, REC, TTA, SE	LP, DR
DWM (Niizumi et al., 2026)	∼86M	TF	Cont	SSL (Pred)	ViT	-	Transformer	✓	✓	✓	CLS	LP, FT
LatentFT (Wang and Huang, 2026)	-	T/F	Cont	SSL (Reco)	DAC / 1D U-Net / MLP	-	1D U-Net			✓	MG	DR
Neural Audio Codec												
VQ-VAE (van den Oord et al., 2017)	-	T	Disc	SSL (Reco)	CNN	VQ	CNN	✓			REC, CLS	LP, DR
SoundStream (Zeghidour et al., 2022)	8.4M	T	Disc	SL	CNN	RVQ	CNN	✓		✓	REC, SE	DR
EnCodec (Défossez et al., 2023)	-	T	Disc	SSL (Reco)	CNN + LSTM	RVQ	CNN + LSTM	✓	✓	✓	REC	DR
HiFi-Codec (Yang et al., 2023a)	-	T	Disc	SSL (Reco)	CNN + LSTM	Group-RVQ	CNN + LSTM	✓			REC	DR
DAC (Kumar et al., 2023)	76M	T	Disc	SSL (Reco)	CNN	RVQ	CNN	✓	✓	✓	REC	DR
SpeechTokenizer (Zhang et al., 2024)	∼100M	T	Disc	SSL (Reco+dist)	CNN + BiLSTM	RVQ	CNN + BiLSTM	✓			TTS	DR
SemantiCodec (Liu et al., 2024a)	75M	TF	Disc	SSL (Reco)	AudioMAE + BiLSTM	VAE	Diffusion + Vocoder	✓	✓	✓	CLS	LP, DR
Single-Codec (Li et al., 2024a)	-	TF	Disc	SSL (Reco)	Conformer + CNN	VQ	BiLSTM	✓			TTS	DR
Spectral Codecs (Langman et al., 2025)	65M	TF	Disc	SSL (Reco)	CNN	RVQ	HiFi-GAN	✓			TTS	DR
LLM-Codec (Yang et al., 2024)	160M	T	Disc	USL (GAN) + SL (dist)	CNN	RVQ	Transformer	✓	✓		CLS, TTS, SE	ZS, DR
SNAC (Siuzdak et al., 2024)	54.5M/19.8M	T	Disc	SSL (Reco)	CNN	RVQ	CNN	✓	✓	✓	REC	DR
UniCodec (Jiang et al., 2025)	274M	T	Disc	SSL (Pred)	ConvNet + Transformer + MoE	SimVQ	Transformer	✓	✓	✓	REC, CLS	DR
STFTCodec (Feng et al., 2025)	36M∼42M	TF	Disc	SSL (Reco)	ConvNeXt V2 + Self-Attention	RVQ	ConvNeXt V2 + Self-Attention	✓			REC	DR
DualCodec (Li et al., 2025a)	∼628M	T	Disc	SSL (Reco)	CNN + ResNet	RVQ	CNN	✓			TTS	DR
SpecTokenizer (Wan et al., 2025)	450K/70M	TF	Disc	SSL (Reco)	CNN	RVQ	RNN	✓			REC, ASR	DR
UniSRCodec (Zhang et al., 2026e)	-	TF	Disc	SSL (Reco)	Open-MagViT2	SimVQ	Open-MagViT2 + BigVGAN-v2	✓	✓	✓	CLS	LP, DR
MOSS-Audio-Tokenizer (Gong et al., 2026a)	1.6b	T	Disc	SSL (Reco)	Transformer	RVQ	Transformer	✓	✓	✓	TTS, ASR	DR, FT

Form: **Cont**: Continuous, **Disc**: Discrete.
Pre-training: **SSL**: Self-Supervised Learning, **SL**: Supervised Learning, **USL**: Unsupervised Learning, **Pred**: Predictive, **Cont**: Contrastive, **Reco**: Reconstructive, **dist**: distillation.
Data: **SP**: Speech, **SO**: Sound, **MU**: Music.
Tasks: **CLS**: Classification, **ASR**: Automatic Speech Recognition, **TTS**: Text-to-Speech, **TTA**: Text-to-Audio, **ATR**: Audio-Text Retrieval, **AAC**: Automated Audio Captioning, **MG**: Music Generation, **SE**: Speech Enhancement, **REC**: Reconstruction, **S2TT**: Speech-to-Text Translation, **SLU**: Spoken Language Understanding.
Eval: **LP**: Linear Probing, **ZS**: Zero-shot, **FT**: Fine-tuning, **DR**: Direct Reconstruction.

Table 4: Overview of Encoder-Decoder Audio Representation Learning Models

D.3 Vocoder

Model	Scale	Audio Modeling Domain	Input Rep. Form	Pre-training Method	Architecture	Data			Tasks	Eval
						SP	SO	MU		
WaveNet (van den Oord et al., 2016)	-	T → T	Disc	SL	CNN	✓		✓	REC, ASR	DR, FT
WaveRNN (Kalchbrenner et al., 2018)	-	TF → T	Cont	SL	RNN + CNN	✓			REC	DR
WaveGlow (Prenger et al., 2019)	-	TF → T	Cont	SL	CNN	✓			REC	DR
GANSynth (Engel et al., 2019)	-	TF → T	Cont	SL	CNN	✓			REC	DR
WaveGrad (Chen et al., 2021b)	15M/23M	TF → T	Cont	SL	UBlock + DBlock	✓			REC	DR
DiffWave (Kong et al., 2021)	2.64M/6.91M	TF → T	Cont	USL	Diffusion	✓			REC, UG	DR
HiFi-GAN (Kong et al., 2020a)	0.92M/1.46M/13.92M	TF → T	Cont	SL	CNN	✓			REC	DR
Fre-GAN (Kim et al., 2021b)	-	TF → T	Cont	SL	HiFi-GAN	✓			REC	DR
UnivNet (Jang et al., 2021)	4M/14.86M	TF → T	Cont	SL	MelGAN	✓			REC, TTS	DR, FT
iSTFTNet (Kaneko et al., 2022)	13.26M	TF → T	Cont	SL	CNN + DSP	✓			REC, TTS	DR
BigVGAN (Lee et al., 2023)	14M/112M	TF → T	Cont	SL	HiFi-GAN	✓			REC	DR
Vocos (Siuzdak, 2024)	13.5M	Token / TF → T	Disc/Cont	SL	ConvNeXt	✓			REC, TTS, MG	DR, ZS
HiFTNet (Li et al., 2023b)	17.7M	TF → T	Cont	SL	LSTM + CNN	✓			REC	DR
APNet2 (Du et al., 2024a)	-	TF → T	Cont	SL	ConvNeXt V2	✓			REC	DR
WaveNeXt (Okamoto et al., 2023)	-	TF → T	Cont	SL	ConvNeXt + Linear Layer	✓			REC	DR
PeriodWave (Lee et al., 2025)	29.7M	TF → T	Cont	SL	U-Net	✓			REC	DR
BemaGANv2 (Park et al., 2025)	∼14M	TF → T	Cont	SL	CNN	✓			REC	DR
T-Mimi (Wu et al., 2026b)	40.8M	Token → T	Disc	SL	Transformer	✓			REC	DR

Input Rep. Form: **Cont**: Continuous, **Disc**: Discrete. **Pre-training**: **SL**: Supervised Learning, **USL**: Unsupervised Learning.
Data: **SP**: Speech, **SO**: Sound, **MU**: Music.
Tasks: **REC**: Reconstruction, **ASR**: Automatic Speech Recognition, **TTS**: Text-to-Speech, **UG**: Unconditional Generation, **MG**: Music Generation.
Eval: **ZS**: Zero-shot, **FT**: Fine-tuning, **DR**: Direct Reconstruction.

Table 5: Overview of Vocoders

E Training Paradigms for Audio Downstream Task Models

Building a robust downstream audio system requires distinct training paradigms to align acoustic features with semantic goals. Depending on the architecture and data availability, these paradigms range from massive multimodal pre-training to training-free adaptations.

E.1 Pre-training

Pre-training refers to constructing a general-purpose foundation model with strong generalization capability using large-scale universal data. In multimodal audio modeling, it is divided into four primary approaches:

Native Multimodal End-to-End Training Initialized from scratch, this paradigm maps all modalities into a shared semantic space from the outset. Joint modeling under a unified objective enables native multimodal capability, exemplified by GPT-4o (Hurst et al., 2024) and Gemini2.5(Gemini Team, 2025). Despite prohibitive costs and severe convergence challenges, SODA (Manakul et al., 2026) unifies audio and text tokens to pre-train a Transformer backbone entirely from scratch.

Cross-modal Alignment End-to-End Pre-training Relying on pre-trained LLMs and audio extractors, this paradigm uses bridging mechanisms to map audio representations into the dominant text modality's continuous latent or discrete token space. It typically adopts multi-stage training: models like Qwen2.5-Omni (Xu et al., 2025) and Qwen3-Omni (Qwen Team, 2025) initially execute cross-modal alignment before conducting global end-to-end joint optimization. Alternatively, this paradigm can treat one modality as the generative target and the other as a guiding condition, as exemplified by the text-guided speech generation in DiTAR (Jia et al., 2025).

Modular Pre-training This paradigm decomposes complex systems into functionally independent sub-networks trained separately with local objectives, avoiding global gradient sharing. Successive modules are conditioned on preceding outputs to align multimodal capabilities with text. For instance, Seed-TTS (Anastassiou et al., 2024) independently pre-trains a speech tokenizer and vocoder, subsequently concatenating audio and text tokens to pre-train a cascaded Transformer-DiT backbone. Similarly, Moshi (Défossez et al., 2024) separately pre-trains a text backbone and an audio

codec, later leveraging extracted audio tokens to continually train the backbone, thereby unifying the text and audio spaces.

E.2 Post-training

After unsupervised distribution matching, post-training introduces explicit supervision and human preferences to directionally distort the model's generation trajectory for real-world interactions.

Supervised Fine-Tuning (SFT) employs high-quality instruction-output pairs for either full or parameter-efficient updates. It activates pre-trained knowledge, enabling models to strictly follow human instructions and master specific downstream tasks and interaction paradigms. Recent frameworks like MoshiRAG (Chien et al., 2026) leverage SFT to integrate retrieval-augmented generation for dialog systems, while CoT-ASR (Deng et al., 2026) utilizes it to incorporate multi-step reasoning into speech recognition.

Reinforcement Learning (RL) leverages human or AI feedback to construct reward signals via algorithms like RLHF, DPO, or GRPO. By maximizing expected returns, RL compensates for SFT's limitations in optimizing subjective criteria, thereby enhancing generation quality, ensuring human-value alignment, and effectively mitigating hallucinations. To unlock advanced cognitive abilities, systems such as Step-Audio-R1 (Tian et al., 2025) and Audio-DeepThinker (He et al., 2026) employ RL-based reasoning strategies to systematically enhance performance in complex audio question-answering tasks.

E.3 Supervised Training from Scratch

Supervised Training from Scratch optimizes a randomly initialized backbone for a specific downstream task using labeled data. However, its performance ceiling is inherently bounded by the scale of human annotations, resulting in poor transferability. Despite these transferability limits, it remains a robust choice for specialized generation and conversion tasks, as demonstrated in Stable Audio Open (Evans et al., 2025) for text-to-music generation and X-VC (Zheng et al., 2026) for voice conversion.

E.4 Training-free

Training-Free approaches yield decisions or generations without any gradient updates during downstream adaptation. For instance, RDP (Wilkinghoff et al., 2026) detects anomalous sounds by comparing feature distances between test and normal audio, whereas Audio-CoT (Ma et al., 2025) employs inference-time prompting to elicit the model's internalized reasoning capabilities.

F Summary of Audio Downstream Task Models

F.1 Overview of Pre-trained Models

Model	Scale	Audio Representation Model	Base Model	DeTokenizer / Vocoder	Cond. Method	Modeling Paradigm	Pre-training	Post-training	Tasks	Downstream
GSLM (Lakhotia et al., 2021)	-	CPC / wav2vec 2.0 / HuBERT + k-means	Transformer	Tacotron-2 + WaveGlow	TI	AR	Modular	-	Gen	ZVC
VALL-E (Chen et al., 2025c)	∼300M	EnCodec	Decoder-only Transformer	EnCodec Decoder	TI + CM	AR, Iter. NAR	Modular	-	Gen, Conv	TTS, ZVC, SE
AudioPaLM (Rubenstein et al., 2023)	8B	USM-v2 / USM-v1 / w2v-BERT	PaLM/PaLM-2	SoundStorm / SoundStream Decoder	RTD + TI	AR	CMA-E2E	SFT	Und, Gen	ASR, S2TT, S2ST, TTS
AudioLM (Borsos et al., 2023a)	∼0.9B	w2v-BERT + SoundStream	Decoder-only Transformer	SoundStream Decoder	TI + RTD	AR	Modular	-	Gen	Cls
Qwen-Audio (Chu et al., 2023)	∼8B	Whisper-large-v2 Encoder	Qwen-7B	-	RTD	AR	CMA-E2E	SFT	Und	ASR, S2TT, AAC, Cls
SpeechGPT (Zhang et al., 2023)	13B	HuBERT	Transformer	HiFi-GAN	TI	AR	CMA-E2E	SFT	Int, Und, Gen	ASR, SDS, TTS
MusicLM (Agostinelli et al., 2023)	∼1.29B	w2v-BERT + SoundStream + MuLan	Decoder-only Transformer	SoundStream Decoder	TI	AR	Modular	-	Gen	TTM
SoundStorm (Borsos et al., 2023b)	350M	w2v-BERT + SoundStream	Conformer	SoundStream Decoder	TI	Iter. NAR	Modular	-	Gen	TTS, ZVC
Audiobox (Vyas et al., 2023)	330M	EnCodec	Transformer	EnCodec Decoder	CM	Iter. NAR	Modular	SFT	Gen, Conv	TTS, TTA, ZVC, AI
Diffsound (Yang et al., 2023b)	∼160M	VQ-VAE	Transformer	MelGAN	CM	Iter. NAR	Modular	-	Gen	TTA
SeamlessM4T (Communication et al., 2025)	1.2B/2.3B	w2v-BERT 2.0	Seamless M4T-NLLB	HiFi-GAN	RTD + CM	AR	Modular	SFT	Und	ASR, S2TT, T2ST, S2ST
Voicebox (Le et al., 2023)	∼360M	Mel-Spectrogram	Transformer	HiFi-GAN	ESA + CM	Iter. NAR	Modular	-	Gen, Conv	TTS, ZVC, SE
LTU (Gong et al., 2024)	∼7B	AST	LLaMA-7B	-	TI	AR	CMA-E2E	SFT	Und	Cls, AAC, AQA
VALL-E 2 (Chen et al., 2024a)	-	EnCodec	Decoder-only Transformer	Vocos	TI + CM	AR, Iter. NAR	Modular	-	Gen, Conv	TTS, ZVC
Qwen2-Audio (Chu et al., 2024)	8.2B	Whisper-large-v3 Encoder	Qwen-7B	-	RTD	AR	CMA-E2E	SFT, RL	Und	ASR, S2TT, Cls
AudioLDM 2 (Liu et al., 2024b)	712M/346M	AudioMAE + VAE	GPT-2 + Transformer-UNet	HiFi-GAN	CM	AR + Iter. NAR	Modular	-	Gen	TTA, TTM, TTS
SALMONN (Tang et al., 2024)	13B	Whisper-large-v2 Encoder + BEATs	Vicuna-13B	-	RTD	AR	CMA-E2E	SFT	Und	ASR, S2TT, AAC
CosyVoice 2 (Du et al., 2024c)	-	FSQ-SenseVoice	Qwen2.5-0.5B + Transformer-UNet	HiFi-GAN	CM + TI	AR + Iter. NAR	Modular	SFT, RL	Gen	TTS, ZVC
Moshi (Défossez et al., 2024)	∼7.6B	Mimi	RQ-Transformer	Mimi Detokenizer	RTD + TI	AR	Modular	SFT	Int	FSD
Seed-TTS (Anastassiou et al., 2024)	-	Custom Speech Tokenizer	Transformer+DiT	Acoustic Vocoder	TI	AR + Iter. NAR	Modular	SFT, RL	Gen, Conv	TTS, VC
GLM-4-Voice (Zeng et al., 2024)	9B	Whisper-large-v3 Encoder + VQ	GLM-4-9B-Base	HiFi-GAN	TI	AR	Modular	SFT	Int, Gen	AQA, TTS, SDS
GAMA (Ghosh et al., 2024)	7B	AST	LLaMA-2-7B	-	GA	AR	Modular	SFT	Und	Cls, AQA, AAC
GPT-4o (Hurst et al., 2024)	-	-	Transformer	-	TI	AR	NM-E2E	-	Und, Gen, Int, Conv	-
Baichuan-Audio (Li et al., 2025b)	7B	Baichuan-Audio-Tokenizer	Baichuan 7B + Flow-matching Decoder	HiFi-GAN	TI	AR + Iter. NAR	CMA-E2E	SFT	Und, Gen, Int	ASR, S2TT, TTS, AQA
Qwen2.5-Omni (Xu et al., 2025)	7B	Qwen2 Audio Encoder	Qwen2.5-7B + Flow-matching Decoder	BigVGAN	TI + RTD	AR + Iter. NAR	CMA-E2E	SFT, RL	Und, Gen, Int	ASR, S2TT, AQA, Cls
Kimi-Audio (KimiTeam et al., 2025)	7B	GLM-4-Voice Tokenizer + Whisper Encoder	Qwen2.5-7B + Flow-matching Decoder	BigVGAN	TI + RTD	AR + Iter. NAR	CMA-E2E	SFT	Und, Gen, Int	ASR, AQA, Cls, AAC
Gemini 2.5 (Gemini Team, 2025)	-	-	-	-	TI	AR	NM-E2E	SFT, RL	Und, Gen, Int	ASR, S2TT, TTS, SDS
Step-Audio 2 mini (StepFun Audio Team, 2025)	7B	Qwen2 Audio Encoder	Qwen2.5-7B + Flow Matching Decoder	HiFi-GAN	RTD + TI	AR + Iter. NAR	CMA-E2E	SFT, RL	Und, Gen, Int	ASR, S2TT, AQA, TTS
FireRedTTS-2 (Xie et al., 2025a)	-	Whisper-based Encoder + RVQ	Qwen2.5	Vocos	TI	AR	Modular	SFT	Gen, Conv	ZVC, SDS, TTS
Qwen3-Omni (Qwen Team, 2025)	33B	AuT	Qwen3	Codec2Wav	RTD	AR	CMA-E2E	SFT, RL	Und, Gen, Int	ASR, S2TT, ATR, AQA
LFM2-Audio (Amini et al., 2025)	1.5B	FastConformer	LFM2-1.2B + RQ-Transformer	Vocos	RTD	AR	CMA-E2E	SFT, RL	Und, Gen, Int	ASR, TTS, SDS
Mimo-Audio (LLM-Core Xiaomi, 2025)	7B	Mimo-Audio-Tokenizer	Mimo-7B-Base	Mimo-Audio-Tokenizer / Vocos	RTD + TI	AR	CMA-E2E	SFT	Und, Gen, Conv, Int	ASR, TTS, AQA, VC
Audio Flamingo 3 (Ghosh et al., 2025)	7B	AF-Whisper Encoder	Qwen2.5-7B	Neural Audio Codec	RTD	AR	CMA-E2E	SFT	Und, Gen, Int	ASR, Cls, AAC
GOAT-SLM (Chen et al., 2025b)	7B	Whisper-small Encoder	TeleChat2-7B + Flow-matching Decoder	Vocoder	RTD	AR + Iter. NAR	CMA-E2E	SFT	Und, Gen, Int	AQA, SDS, TTS
CoT E2E (Arora et al., 2025)	1.7B	ESPnet-Codec + XEUS	SpeechLM	ESPnet-Codec	TI	AR	CMA-E2E	SFT	Int	SDS
CosyVoice 3 (Du et al., 2025)	0.5B, 1.5B	FSQ-MinMo	Transformer	DiT	TI + CM	AR + Iter. NAR	Modular	SFT, RL	Gen	TTS, ZVC
Voila (Shi et al., 2025)	-	Voila Tokenizer	Voice-language LLM backbone + Audio Transformer	Voila Detokenizer	TI	AR	CMA-E2E	SFT, RL	Und, Gen, Int, Conv	ASR, TTS, SDS, S2TT
DITAR (Jia et al., 2025)	0.1B/0.4B/0.6B/1.0B	VAE	Causal Transformer + LocDiT	BigVGAN	RTD + CM	AR + Iter. NAR	CMA-E2E	SFT	Gen	TTS
HeartMula (Yang et al., 2026a)	3.3B / 7B	HeartCodec Tokenizer	Llama 3.2	HeartCodec Detokenizer	TI	AR	Modular	SFT, RL	Gen	TTM
SODA (Manakul et al., 2026)	135M/600M/1.7B/4B	Mimi Tokenizer	Qwen3-based	Mimi Detokenizer	TI	AR	NM-E2E	SFT	Gen, Conv, Und	TTS, ZVC, ASR
MOSS-TTS (Gong et al., 2026b)	1.7B / 8B	MOSS-Audio-Tokenizer	Decoder-only LLM	MOSS-Audio-Tokenizer	TI	AR	Modular	-	Gen	TTS, ZVC
AudioX (Tian et al., 2026c)	2.4B	AudioLDM 2 Encoder	DiT	-	GA + CM	Iter. NAR	CMA-E2E	-	Gen, Conv	TTA, VTA, TTM
Audio Flamingo Next (Ghosh et al., 2026a)	7B	AF-Whisper Encoder + 2-layer MLP	Qwen2.5-7B	Streaming TTS	RTD	AR	CMA-E2E	SFT, RL	Und, Int	ASR, SD, S2TT, Cls
Eureka-Audio (Zhang et al., 2026a)	1.7B	Whisper-based Encoder	Qwen3-1.7B-base	-	RTD	AR	CMA-E2E	SFT	Und	ASR, AAC, AQA
UALM (Tian et al., 2026a)	1.5B / 7B	Qwen2 Audio Encoder	Qwen2.5-7B	X-codec Decoder + VAE Decoder	RTD	AR	CMA-E2E	SFT, RL	Und, Gen, Int	ASR, TTA, SDS, AQA
OmniVoice (Zhu et al., 2026)	0.8B	Higgs-audio Tokenizer	Qwen3-0.6B	Higgs-audio Detokenizer	TI	Iter. NAR	Modular	-	Gen, Conv	TTS, ZVC, SE
Audio-Omni (Tian et al., 2026d)	7.9B	VAE	Qwen2.5-Omni-3B + DiT	VAE Decoder	CM	AR + Iter. NAR	Modular	-	Gen, Conv	AQA, TTA, TTM, VTA
Omni-Diffusion (Long et al., 2026)	7B	SenseVoiceSmall	Qwen-7B	GLM-4-Voice Decoder	RTD	AR + Iter. NAR	CMA-E2E	SFT	Und, Gen, Int	ASR, TTS, AQA, SDS
Bagpiper (Tian et al., 2026b)	8B	Qwen3-Omni-Encoder + X-Codec	Qwen3-8B-Base	X-Codec Decoder	RTD + TI	AR	CMA-E2E	SFT	Und, Gen	ASR, AQA, TTS, TTA
AudioChat (Chen et al., 2026)	3.6B	DAC-VAE Encoder	Gemma 2 2B	DAC-VAE Decoder	RTD	AR + Iter. NAR	CMA-E2E	SFT	Und, Gen, Conv	ASR, AAC, SD, TTS
Hello-Chat (Hou et al., 2026)	∼8B	MiDashengLM	Qwen2.5-7B-Instruct	CosyVoice 2	RTD	AR + Iter. NAR	CMA-E2E	SFT	Und, Gen, Int	ASR, S2TT, Cls
DIFFA-2 (Zhou et al., 2026b)	8.77B	Whisper-large-v3 Encoder	LLaDA-8B-Instruct	-	RTD + GA	Iter. NAR	CMA-E2E	SFT, RL	Und	ASR, AQA
CALM (Rouard et al., 2026)	100M∼3B	VAE Encoder	Transformer / Helium-1	VAE Decoder	RTD +CM	AR + One-shot NAR	CMA-E2E	-	Gen	TTS, TTM
DIFFA (Zhou et al., 2026a)	8B	Whisper-small Encoder	LLaDA-8B-Instruct	-	RTD + GA	Iter. NAR	CMA-E2E	SFT	Und	ASR, AQA
VibeVoice (Peng et al., 2026b)	1.5B, 7B	Acoustic Tokenizer + Semantic Tokenizer	Qwen2.5	Acoustic Detokenizer	RTD + CM	AR + Iter. NAR	CMA-E2E	-	Gen	TTS, ZVC
TiT (Text-to-Talk) (Liu et al., 2026a)	1.5B, 3B	GLM-4-Voice Tokenizer	Qwen2.5-Base	GLM-4-Voice Decoder	TI	AR + Iter. NAR	CMA-E2E	SFT	Und, Gen, Int	AQA, ASR, AAC, TTS
OWL (Biswas et al., 2026)	7B	SAGE	LLaMA-2-7B	-	GA	AR	CMA-E2E	SFT	Und	Cls, AQA
DeSTA2.5-Audio (Lu et al., 2026a)	8.8B	Whisper-large-v3 Encoder	Llama3.1-8B-Instruct	-	GA	AR	Modular	SFT	Und	Cls, AQA, AAC, ASR
TangoFlux (Hung et al., 2026)	515M	Stable Audio Open VAE Encoder	MMDiT + DiT	Stable Audio Open VAE Decoder	CM	Iter. NAR	Modular	SFT, RL	Gen	TTA
SLM-SS (Li et al., 2026b)	-	WavLM-Large + EnCodec	Transformer	EnCodec Decoder	TI + CM	AR + One-shot NAR	Modular	SFT	Conv	SS, SE
UAF (Li et al., 2026c)	3B, 7B, 30B	-	Qwen3-Omni-30B-A3B-Instruct	-	-	AR	CMA-E2E	SFT	Und	Cls, ASR, AQA
LLM-ASR (Xie et al., 2026)	2.3B	Conformer	Qwen3-1.7B	-	RTD	AR	CMA-E2E	SFT	Und	ASR

Cond. Method: **GA**: Global Aggregation, **CM**: Condition Modulation, **TI**: Token-level Integration, **RTD**: Rigid Temporal Downsampling, **ESA**: Elastic Semantic Alignment.
Modeling Paradigm: **AR**: Autoregressive, **Iter. NAR**: Iterative Non-Autoregressive, **One-shot NAR**: One-shot Non-Autoregressive.
Pre-training: **CMA-E2E**: Cross-modal Alignment End-to-End Pre-training, **Modular**: Modular Pre-training, **NM-E2E**: Native Multimodal End-to-End Training.
Tasks: **Gen**: Generation, **Conv**: Conversion, **Und**: Understanding, **Int**: Interaction.
Understanding: **Cls**: Classification, **ASR**: Auto Speech Recognition, **S2TT**: Speech-to-Text Translation, **AQA**: Audio Question Answering, **AAC**: Automated Audio Captioning, **ATR**: Audio Text Retrieval, **SD**: Speaker Diarization.
Generation: **TTS**: Text to Speech, **ZVC**: Zero-shot Voice Cloning, **TTA**: Text to Audio, **VTA**: Video to Audio, **TTM**: Text to Music, **T2ST**: Text-to-Speech Translation.
Conversion: **VC**: Voice Conversion, **SE**: Speech Enhancement, **SS**: Speech Separation, **AI**: Audio Inpainting, **S2ST**: Speech-to-Speech Translation.
Interaction: **FSD**: Full-duplex Spoken Dialogue, **SDS**: Spoken Dialogue System.

Table 6: Overview of Audio Pre-training Models

F.2 Overview of Downstream Fine-tuning Models

Model	Scale	Audio Representation Model	Base Model	DeTokenizer / Vocoder	Cond. Mech.	Modeling Paradigm	Pre-training	Post-training	Tasks	Downstream
FastSpeech 2 (Ren et al., 2021)	~27M	Mel-Spectrogram	Feed-Forward Transformer	Parallel WaveGAN	ESA	One-shot NAR	Sup. Scratch		Gen	TTS
VITS (Kim et al., 2021a)	~12M	Mel-Spectrogram	Transformer	HiFi-GAN	ESA	One-shot NAR	Sup. Scratch		Gen	TTS
DiffVC (Popov et al., 2022)	-	Mel-Spectrogram	Transformer + U-Net	HiFi-GAN	GA	Iter. NAR	Sup. Scratch		Conv	VC
Paraformer (Gao et al., 2022)	41M	Mel-Spectrogram	Paraformer	-	ESA	One-shot NAR	Sup. Scratch		Und	ASR
MUSICGEN (Copet et al., 2023)	300M, 1.5B, 3.3B	EnCodec	Decoder-only Transformer	EnCodec	CM + TI	AR	Sup. Scratch		Gen	TTM
AST-SED (Li et al., 2023a)	-	AST	Bi-GRU	-	RTD	Disc.	-	SFT	Und	Cls
TF-GridNet (Wang et al., 2023)	2.1M - 14.5M	Mel-Spectrogram	TF-GridNet	iSTFT	-	Disc.	Sup. Scratch		Conv	SS, SE, SD
CoMoSpeech (Ye et al., 2023)	-	Mel-Spectrogram	FastSpeech-style FFT Transformer	HiFi-GAN	-	One-shot NAR	Sup. Scratch		Gen	TTS, TTM
Seed-VC (Liu, 2024)	~200M	Whisper-small Encoder / CAM++ / RMVPE	U-DiT	BigVGANv2	CM	Iter. NAR	Sup. Scratch		Conv	VC
AudioSep (Liu et al., 2024c)	238.6M	CLAP / CLIP	ResUNet	-	CM	Disc.	Sup. Scratch		Und	SS
Mini-Omni (Xie and Wu, 2024)	0.5B	Whisper-small Encoder	Qwen2-0.5B	SNAC	RTD	AR	-	SFT	Und, Gen, Int	ASR, TTS, AQA, SDS
Freeze-Omni (Wang et al., 2025c)	7B	CNN+Transformer	Qwen2-7B-Instruct	TiCodec	RTD	AR, One-shot NAR	-	SFT	Int	FSD, SDS, ASR, AQA
Tango 2 (Majumder et al., 2024)	866M	AudioLDM VAE	Tango	HiFi-GAN	CM	Iter. NAR	-	RL	Gen	TTA
CosyVoice (Du et al., 2024b)	~300M	S3 Tokenizer	Transformer + ResNet1D	HiFi-GAN	TI	AR + Iter. NAR	Sup. Scratch		Gen	TTS, ZVC
ARDiT (Liu et al., 2024d)	~130M	Contextual Mel Spectrogram Autoencoder	ARDiT	BigVGAN	RTD	AR + Iter. NAR	Sup. Scratch		Gen, Conv	TTS, ZVC, AI
Wav2Prompt (Deng et al., 2025b)	7B	Conformer	LLaMA-2 / BLOOMZ / Vicuna	-	ESA	AR	-	SFT	Und	ASR, S2TT, AQA
CM-TTS (Li et al., 2024b)	28.6M	Mel-Spectrogram	FastSpeech2-style Transformer	HiFi-GAN	-	One-shot NAR	Sup. Scratch		Gen	TTS
STSR (Yuan et al., 2025)	66.2M	MDCT Coefficient	U-Net	-	RTD	One-shot NAR	Sup. Scratch		Conv	SE
Tri-Stream (Cao et al., 2025)	-	Swin Transformer, HuBERT, openSMILE	GRU	-	GA	Disc.	Sup. Scratch		Und	Cls
Step-Audio-R1 (Tian et al., 2025)	32B	Qwen2 Audio Encoder	Qwen2.5-32B	-	RTD	AR	-	SFT, RL	Und, Int	AQA, SDS
Stable Audio Open (Evans et al., 2025)	1B	VAE	DiT	VAE Decoder	CM	Iter. NAR	Sup. Scratch		Gen	TTA, TTM
Music Flamingo (Ghosh et al., 2026b)	3B / 7B	-	Audio Flamingo 3	-	-	AR	-	SFT, RL	Und	ATR, ASR, AQA
PAC (FireRed-LLM) (Fu et al., 2026)	7B	-	FireRed-LLM-7B	-	-	AR	-	SFT, RL	Und	ASR
YingMusic-SVC (Chen et al., 2025a)	~200M	-	Seed-VC	BigVGAN	-	Iter. NAR	-	SFT, RL	Conv	VC
Omni-R1 (Rouditchenko et al., 2025)	7B	-	Qwen2.5-Omni-7B	-	-	AR	-	RL	Und	AQA
DAR-LALM (Sheng et al., 2026)	7B	-	Qwen2.5-Omni-7B / Qwen2-Audio-7B-Instruct	-	-	AR	-	SFT, RL	Und	AQA, Cls
Rebellion (Huang et al., 2025)	7B	-	Qwen2-Audio-7B	-	-	AR	-	SFT	Und	AQA
AsyncVoice Agent (Lin et al., 2025b)	-	-	GPT-4o	Azure TTS	-	-		Tr. Free	Int	AQA, FSD
AudioStory (Guo et al., 2025)	3B	Whisper-large-v3 Encoder	Qwen-2.5-3B-Instruct	TangoFlux	RTD + CM	AR + Iter. NAR	-	SFT	Gen, Und	TTA, AQA, AAC, VTA
AlignFormer (Fan et al., 2025)	4.3B	Conformer	Phi3.1-mini-instruct	HiFi-GAN	ESA	AR	-	SFT	Und	ASR, S2TT, AQA
Audio-Reasoner (Xie et al., 2025c)	8.4B	-	Qwen2-Audio-Instruct	-	-	AR	-	SFT	Und	AQA, Cls, S2TT
SoundMind-RL (Diao et al., 2025)	7B	-	Qwen2.5-Omni-7B	-	-	AR	-	SFT, RL	Und, Gen	AQA, TTA
Serenade (Violeta et al., 2025)	-	ContentVec, Prior Encoder, Style Encoder	1D U-Net	HIFIGAN, SIFIGAN	CM	Iter. NAR	Sup. Scratch		Conv	VC
Reflow-VC (Ren et al., 2025)	-	Average Voice Encoder, HuBERT-Soft, Speaker Encoder, VQ-VAE	U-Net	HiFi-GAN	CM	Iter. NAR	Sup. Scratch		Conv	VC
U-SAM (Wang et al., 2025d)	8B	Whisper Encoder, CED, MERT	LLaMA-3.1-8B	-	RTD	AR	-	SFT	Und	ASR, S2TT, AAC, Cls
Audio-CoT (Ma et al., 2025)	7B	-	Qwen2-Audio-7B-Instruct	-	-	AR		Tr. Free	Und	AQA, ASR, AAC
SimulS2S-LLM (Deng et al., 2025a)	7B/8B	Streaming Transformer	BLOOMZ-7B / Llama3-8B	HiFi-GAN	ESA	AR	-	SFT	Und, Conv	S2ST, S2TT
CJST (Zhou et al., 2025)	~110M	Conformer+VGG	Decoder-only Transformer	-	RTD + ESA	AR	-	SFT	Und	ASR
Conformer-LAIL (Altinok, 2025)	8.3B	Wav2Vec 2.0	Wav2Vec 2.0 Conformer Large	-	RTD	One-shot NAR	-	SFT	Und	ASR
HyPeR (Wang et al., 2026g)	-	Whisper Encoder	Qwen2-Audio-7B-Instruct	-	-	AR	-	SFT, RL	Und	AQA, S2TT, Cls, SD
Nemotron-0.6B INT4 k-quant (Banfic et al., 2026)	-	-	Nemotron-0.6B	-	-	AR	-	-	Und	ASR
Music Flamingo (Ghosh et al., 2026c)	7B	-	Audio Flamingo 3	-	-	AR	-	SFT, RL	Und	AQA, AAC, ATR
MDM-ASR (Yen et al., 2026)	180M / 1B	FastConformer	Non-causal Transformer Decoder	-	RTD + CM	Iter. NAR	-	-	Und	ASR
Audio-DeepThinker (He et al., 2026)	30B	-	Qwen3-Omni-30B-A3B-Instruct	-	-	AR	-	RL	Und	AQA
X-VC (Zheng et al., 2026)	539M	SAC Encoder, ERes2Net	Transformer	SAC Decoder	RTD + CM	One-shot NAR	Sup. Scratch		Conv	VC
MoshiRAG (Chien et al., 2026)	7B	-	Moshi	-	TI	AR	-	SFT	Und, Int	AQA, SDS, ATR
Speech World Model (SWM) (Zhou et al., 2026c)	~8B	WavLM	Qwen2-Audio, LLaMA3.1-8B	-	TI	Disc. + AR	-	SFT	Und	Cls
Granite-speech (Aronowitz et al., 2026)	8B	Conformer	Granite-3.3-8b-instruct	-	GA	AR	-	SFT	Und	ASR, SD, Cls
Projector-based LLM-ASR (Li et al., 2026d)	1.7B / 8B	Whistle, WavLM	Qwen3-Base	-	RTD + CM	AR	-	SFT	Und	ASR
Phoneme-based LLM-ASR (Li et al., 2026d)	1.7B / 8B	Whistle, WavLM	Qwen3-Base	-	ESA	AR	-	SFT	Und	ASR
STEMPHONIC (Wu et al., 2026c)	~1B	VAE	DiT	VAE	CM	Iter. NAR	-	SFT	Gen	TTM
TASU2 (Peng et al., 2026a)	~1.8B	SenseVoice-Small	Qwen2.5-1.5B	-	ESA	AR	-	SFT	Und	ASR, AAC
AEG (Yang et al., 2026d)	7B	-	Qwen2-Audio-7B	-	-	AR	-	SFT	Und	AQA
PE-FT LALMs (Yang et al., 2026c)	-	-	Qwen2.5-Omni / Kimi-Audio	-	-	AR	-	SFT	Und, Int	Cls, SDS
PROADS (Yan et al., 2026)	-	EzAudio Codec Encoder	EzAudio	EzAudio Codec Decoder	-	Iter. NAR	-	-	Gen	AI
ALARM-E (Grinberg and Shahmohammadi, 2026)	4B	Whisper Encoder, W2V-BERT-2.0, MuQ, SSLAM	Qwen3-4B-Thinking-2507	-	RTD	AR	-	SFT	Und	AQA
Ambiguity-Aware LALM (Yu et al., 2026)	7B	-	Qwen2-Audio-7B-Instruct	-	-	AR	-	SFT, RL	Und	Cls
RDP / RDP+GeM (Wilkinghoff et al., 2026)	-	OpenL3, BEATs, EAT, Dasheng	-	-	-	Disc.		Tr. Free	Und	Cls
StyleStream (Liu et al., 2026b)	-	HuBERT, WavLM	DiT	Vocos (Causal)	CM	Iter. NAR	Sup. Scratch		Conv	VC
ReHear (Liu et al., 2026c)	3B	-	Whisper-large-v3	-	-	AR	-	SFT	Und	ASR
AuroLA (Xu et al., 2026)	7B	-	Qwen2.5-Omni-7B	-	-	AR	-	SFT	Und	ATR
TAC (Kumar et al., 2026b)	-	-	Qwen2-Audio	-	-	AR	-	SFT	Und	AAC, Cls, AQA
AIDD (Dror et al., 2026)	90M	WavTokenizer	DiT	WavTokenizer	CM	Iter. NAR	Sup. Scratch		Conv	AI
GSRM (Shen et al., 2026)	7B	-	Qwen2.5-Omni-7B	-	-	AR	-	SFT	Und	AQA
Echo (Wu et al., 2026a)	7B	-	Qwen2.5-Omni-7B	-	-	AR	-	SFT, RL	Und	AQA
LLM-ForcedAligner (Mu et al., 2026)	0.9B	AuT	Qwen3-0.6B	-	RTD	One-shot NAR	-	SFT	Und	ASR
EmotionThinker (Wang et al., 2026c)	7B	-	Qwen2.5-Omni-7B	-	-	AR	-	SFT, RL	Und	Cls
TARS (Wang et al., 2026a)	7B	-	Qwen2.5-Omni-7B / Phi-4-MM-7B	-	-	AR	-	SFT, RL	Und	AQA, ASR
TimeAudio (Wang et al., 2026d)	7B	BEATs, Whisper Encoder	SALMONN-7B	-	GA	AR	-	SFT	Und	AAC, AQA
CESAR (Fan et al., 2026)	7B	-	Qwen2.5-Omni-7B	-	-	AR	-	RL	Und	AQA
USR 2.0 (Haliassos et al., 2026)	86M/171M/503M/953M	ResNet-18	USR	-	ESA	Disc. + AR	-	SFT	Und	ASR
AudSemThinker (Wijngaard et al., 2026)	7B	-	Qwen2.5-Omni	-	-	AR	-	SFT, RL	Und	AQA, AAC
STITCH (Chiang et al., 2026)	9B	-	GLM-4-Voice-9B	-	-	AR	-	SFT	Int	AQA, SDS
Mellow (Deshmukh et al., 2026)	167M	HTSAT	SmolLM2	-	RTD	AR	-	SFT	Und	AQA, AAC
Unified TSD (Gupta et al., 2026)	-	ConvNeXT	BiGRU	-	CM	Disc.	-	SFT	Und	Cls
SLM-SS (Li et al., 2026b)	~600M	Encodec, WavLM-large	Transformer	Encodec Decoder	CM + TI	AR + Iter. NAR	Sup. Scratch		Conv	SS
CoT-ASR (Deng et al., 2026)	3.8B	Conformer	Phi-4-mini-instruct	-	ESA	AR	-	SFT	Und	ASR
MeanVC (Ma et al., 2026a)	14M	ECAPA-TDNN, Timbre Encoder, Fast-U2++	DiT	Vocos	CM + GA	AR + One-shot NAR	Sup. Scratch		Conv	VC
PersonaPlex (Roy et al., 2026)	7B	-	Moshi	-	-	AR	-	SFT	Int	FSD, SDS, S2ST, ZVC

Cond. Method: **CM**: Condition Modulation, **TI**: Token-level Integration, **RTD**: Rigid Temporal Downsampling, **GA**: Global Aggregation, **ESA**: Elastic Semantic Alignment.
Modeling Paradigm: **AR**: Autoregressive, **Iter. NAR**: Iterative Non-Autoregressive, **One-shot NAR**: One-shot Non-Autoregressive, **Disc.**: Discriminative.
Pre-training: **Sup. Scratch**: Supervised training from scratch, **Tr. Free**: Training-free.
Tasks: **Und**: Understanding, **Gen**: Generation, **Conv**: Conversion, **Int**: Interaction.
Understanding: **Cls**: Classification, **ASR**: Auto Speech Recognition, **S2TT**: Speech-to-Text Translation, **AQA**: Audio Question Answering, **AAC**: Automated Audio Captioning, **ATR**: Audio Text Retrieval, **SD**: Speaker Diarization.
Generation: **TTS**: Text to Speech, **ZVC**: Zero-shot Voice Cloning, **TTA**: Text to Audio, **VTA**: Video to Audio, **TTM**: Text to Music, **T2ST**: Text-to-Speech Translation.
Conversion: **VC**: Voice Conversion, **SE**: Speech Enhancement, **SS**: Speech Separation, **AI**: Audio Inpainting, **S2ST**: Speech-to-Speech Translation.
Interaction: **FSD**: Full-duplex Spoken Dialogue, **SDS**: Spoken Dialogue System.

Table 7: Overview of Downstream Task Fine-tuning Models