Recent Advances in Speech Language Models: A Survey

Wenqian Cui[1]*, **Dianzhi Yu**[1], **Xiaoqi Jiao**[2], **Ziqiao Meng**[3], **Guangyan Zhang**[2], **Qichao Wang**[4], **Yiwen Guo**[5], **Irwin King**[1]†
[1]The Chinese University of Hong Kong, [2]LIGHTSPEED STUDIOS,
[3]National University of Singapore, [4]Tencent, [5]Independent Researcher

Abstract

Text-based Large Language Models (LLMs) have recently gained significant attention, primarily for their capabilities in text-based interactions. However, natural human interaction often relies on speech, highlighting the need for voice-based models. In this context, Speech Language Models (SpeechLMs)—foundation models designed to understand and generate speech—emerge as a promising solution for end-to-end speech interaction. This survey offers a comprehensive overview of recent approaches to building SpeechLMs, outlining their core architectural components, training methodologies, evaluation strategies, and the challenges and potential directions for future research in this rapidly advancing field.[1]

1 Introduction

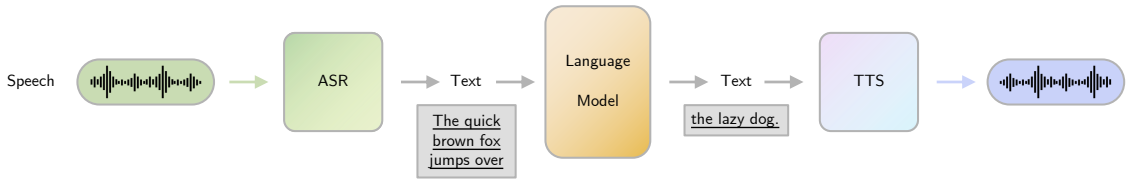
Text-based Large Language Models (LLMs) have demonstrated remarkable capabilities in generating text and performing a wide array of natural language processing tasks (Zhao et al., 2023; Minaee et al., 2024), serving as powerful foundation models for AI language understanding and generation. Their success has also spurred numerous applications in other domains, yet the reliance solely on text-based modalities presents a significant limitation. This leads to the development of speech-based generative models, which allow to interact with humans more naturally and intuitively.

Given the extensive mutual information between text and speech, it is natural to modify existing LLMs to enable speech interaction capabilities. A straightforward approach is to adopt an "Automatic Speech Recognition (ASR) + LLM + Text-to-Speech (TTS)" framework (Figure 1a) (Huang et al., 2024b; Shen et al., 2024). In this setup, the user's spoken input is first converted into text by the ASR module, the LLM then generates a text response based on this transcription, and the TTS module transforms the text response back into speech. However, this naive solution mainly suffers from the following three problems. 1) **Information loss.** Speech signals not only contain semantic information (i.e., the meaning of the speech) but also paralinguistic information (e.g., pitch, timbre, tonality, etc.). Putting a text-only LLM in the middle will cause the complete loss of paralinguistic information in the input speech (Zhang et al., 2023a). 2) **Significant latency.** Combining ASR, LLM, and TTS leads to considerable delays due to their complex structural designs. (Xie and Wu, 2024a; Défossez et al., 2024; Fang et al., 2024). For instance, ASR often includes an additional text decoder (Radford et al., 2023; Le et al., 2020), which increases computational demands. 3) **Cumulative error.** A staged approach like this can easily lead to cumulative errors throughout the pipeline, particularly during the transition between ASR and LLM (Fathullah et al., 2024; Tang et al., 2024).

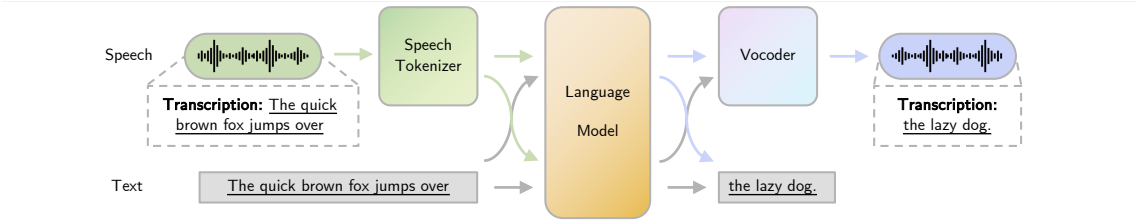
The limitations of this naive framework have led to the development of Speech Language Models (SpeechLMs, Figure 1b). Specifically, SpeechLMs encode speech waveforms directly into speech tokens (Section 3.1). This enables them to capture both semantic and paralinguistic information from audio, thereby minimizing information loss. SpeechLMs then model these tokens autoregressively (Section 3.2), without solely relying on text input. This allows them to use the additional paralinguistic information to generate more expressive and nuanced speech. Finally, the generated tokens are synthesized back to speech (Section 3.3). This integrated approach eliminates the need to chain three separate modules, significantly reducing latency. By working directly with the encoded speech tokens, SpeechLMs effectively mitigate the cumulative errors, as their training is integrated with

*Project lead. Email: wenqian.cui@link.cuhk.edu.hk
†Corresponding author. Email: king@cse.cuhk.edu.hk
[1]The GitHub repository is available at https://github.com/dreamtheater123/Awesome-SpeechLM-Survey

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13943–13970
July 27 - August 1, 2025



(a) Illustration of the “ASR + LLM + TTS" framework.



(b) Illustration of the architecture of a SpeechLM.

Figure 1: Architectures of the “ASR + LLM + TTS" framework and a SpeechLM. SpeechLMs are designed with end-to-end speech interaction capabilities, complemented by optional cross-modality interaction capabilities.

the speech encoding, whereas the training of the three modules is completely independent in the naive framework. Furthermore, SpeechLMs have the potential to enable real-time voice interactions, allowing the model to be interrupted by users or to speak simultaneously with them, mimicking natural human conversation patterns more closely.

In this survey, we provide the first thorough overview of recent advancements in SpeechLMs and introduce a comprehensive taxonomy (Figure 2). We explore the various components that constitute their architecture (Section 3) and the training recipes (Section 4) involved in their development. we aim to elucidate the current state of the field by analyzing these models from the above perspectives. We then classify metrics to evaluate SpeechLMs (Section 5) and discuss the challenges and future directions in this area, hoping to provide valuable insights that could drive further advancements in SpeechLM technology (Section 6).

2 Problem Formulation

This section provides the formal definition of Speech Language Models (SpeechLMs). SpeechLMs are autoregressive foundation models that perform **end-to-end speech interaction**[2]. They leverage contextual understanding for coherent sequence generation, enabling various downstream tasks through speech modality. While SpeechLMs are required to perform speech interactions (speech-in-speech-out), they can also integrate text, supporting cross-modal operations such as speech-in-text-out and vice versa. This distinguishes them from traditional text-based language models (TextLMs), where the only modality being processed within the model is text.

We offer a unified framework in which SpeechLMs can process and generate speech data, text data, or even interleaved speech and text data. Specifically, an audio waveform $\mathbf{a} = (a_1, a_2, \ldots, a_Q)$ consists of a sequence of audio samples $a_i \in \mathbb{R}$ of length Q, where $1 \leq q \leq Q$. Similarly, a text span $\mathbf{t} = (t_1, t_2, \ldots, t_K)$ consists of a sequence of text tokens t_j (word, subword, character, etc.) of length K. Let $\mathbf{M} = (M_1, M_2, \ldots, M_N)$ denote a multimodal sequence of length N, where each element $M_i \in \{a_i, t_j\}$. We define $\mathbf{M}^{\text{in}} = (M_1^{\text{in}}, M_2^{\text{in}}, \ldots, M_{N_{\text{in}}}^{\text{in}})$ as the input multimodal sequence and $\mathbf{M}^{\text{out}} = (M_1^{\text{out}}, M_2^{\text{out}}, \ldots, M_{N_{\text{out}}}^{\text{out}})$ as the output multimodal sequence, where $N_{\text{in}} \geq 0$ and $N_{\text{out}} \geq 0$. Then, A SpeechLM parameterized by θ can then be represented as: $\mathbf{M}^{\text{out}} = SpeechLM(\mathbf{M}^{\text{in}}; \theta)$.

3 Components in SpeechLM

SpeechLM consists of three key components: speech tokenizer, language model, and token-to-speech synthesizer (vocoder) (see Figure 1). The speech tokenizer converts audio waveforms into tokens, which the language model uses for next-token prediction. The vocoder then converts these predicted tokens back into audio waveforms. This three-stage architecture enables autoregressive speech modeling using traditional language model

[2]We are aware of similar concepts such as Vision Language Models (VLMs), which typically refer to models that process images as input and generate text as output. In contrast, we define SpeechLMs specifically as models capable of both receiving and generating speech.

architectures like decoder-only transformers. Table 1 shows common choices for the three SpeechLM components in various SpeechLM papers.

3.1 Speech Tokenizer

Speech tokenizer is the first component in SpeechLMs, which converts audio waveforms into tokens. It processes audio segment by segment, producing either **discrete tokens** (using indices) or **continuous tokens** (using embeddings). Both token types can serve as input for language models in autoregressive modeling, with the main goal of capturing essential audio features while reducing dimensionality. This section classifies speech tokenizers by how they model raw audio.

3.1.1 Semantic Understanding Objective

Speech tokenizers with a semantic understanding objective convert speech waveforms into tokens that accurately capture the meaning of the speech. These tokenizers extract semantic features from the waveforms, which enhances tasks like ASR.

A semantic understanding speech tokenizer typically comprises a speech encoder and a quantizer. The speech encoder ($f_E(\cdot)$) transforms waveform input into continuous embeddings ($\mathbf{v} = f_E(\mathbf{a}; \theta_{f_E})$), where $\mathbf{v} = (v_1, v_2, \ldots, v_P)$. A quantizer ($d(\cdot)$) can be added to convert these embeddings into discrete indexes. The speech tokens $\mathbf{s} = (s_1, s_2, \ldots, s_P)$ can be derived either from the original waveform or the embeddings: $\mathbf{s} = d(\mathbf{v}; \theta_d)$ or $\mathbf{s} = d(\mathbf{a}; \theta_d)$ for discrete tokens, or $\mathbf{s} = \mathbf{v}$ for continuous tokens. These tokens can then be used as target labels for training the tokenizer or subsequent language models.

The key design choices lie in how to effectively encode (and quantize) speech into tokens. Wav2vec 2.0 (Baevski et al., 2020b) combines convolutional encoding with product quantization (Jegou et al., 2010) for waveform discretization, using masked contrastive learning. W2v-BERT (Chung et al., 2021) extends this by adding the Masked Language Modeling (MLM) loss (Devlin et al., 2019). HuBERT (Hsu et al., 2021) uses k-means clustering to derive speech units and uses MLM loss in training. Google USM (Zhang et al., 2023b) incorporates text-injection loss (Chen et al., 2022b) to better align text and speech representations. WavLM (Chen et al., 2022a) introduces speech denoising during pre-training, proving beneficial for both semantic tasks (ASR, TTS) and non-semantic tasks (speaker verification, speech separation).

3.1.2 Acoustic Generation Objective

Speech tokenizers with an acoustic generation objective focus on preserving acoustic features for high-quality speech synthesis. Their architecture consists of an encoder $f_E(\cdot)$, quantizer $d(\cdot)$, and decoder $f_D(\cdot)$. The encoder and quantizer convert waveforms into tokens, while the decoder reconstructs these tokens back into speech waveforms, expressed as $\hat{\mathbf{a}} = f_D(\mathbf{s}; \theta_{f_E})$. This approach prioritizes acoustic characteristics over semantic content, optimizing for speech (re)synthesis tasks.

Neural audio codecs primarily serve as acoustic generation speech tokenizers (Zeghidour et al., 2021; Défossez et al., 2023). They use deep neural networks to compress audio into discrete tokens. Their encoder-quantizer-decoder structure works by: (1) encoding audio into latent representations, (2) discretizing these typically through Vector Quantization (VQ) (Van Den Oord et al., 2017) or Residual Vector Quantization (RVQ) (Zeghidour et al., 2021), and (3) decoding tokens back to audio. Therefore, the encoder and quantizer components function as the speech tokenizer.

3.1.3 Mixed Objective

Speech tokenizers with a mixed objective aim to balance both semantic understanding and acoustic generation. Rather than creating new architectures, most implementations modify acoustic tokenizers to incorporate semantic information. For example, SpeechTokenizer (Zhang et al., 2024e) uses RVQ-GAN (Défossez et al., 2023; Zeghidour et al., 2021) architecture and distills HuBERT's (Hsu et al., 2021) semantic tokens into its first RVQ layer, while Mimi (Défossez et al., 2024) distills WavLM tokens (Chen et al., 2022a) into a single VQ layer alongside the RVQ module.

3.2 Language Model

Following TextLMs' success (Achiam et al., 2023; Dubey et al., 2024), SpeechLMs typically adopt transformer or decoder-only architectures for autoregressive speech generation. Formally, a text-based decoder-only transformer language model comprises: (1) An embedding matrix $E_t \in \mathbb{R}^{|V_t| \times h}$, where $|V_t|$ is vocabulary size and h is hidden dimension. (2) L transformer decoder blocks $\mathbf{De} = \{De_1, De_2, \ldots, De_L\}$. (3) Output embedding matrix $E'_t \in \mathbb{R}^{h \times |V_t|}$. Then, the language model can be expressed as $\mathbf{t}^{\text{out}} \sim \text{LM}(\mathbf{t}^{\text{in}}, (E_t, \mathbf{De}, E'_t))$.

To adapt the language model for speech generation, the text tokenizer can be replaced with a

speech tokenizer. For **discrete tokens**, the text embedding matrix $E_t \in \mathbb{R}^{|V_t| \times h}$ becomes speech embedding matrix $E_s \in \mathbb{R}^{|V_s| \times h}$, where $|V_s|$ is the speech tokenizer vocabulary size. Similarly, output embedding changes from $E'_t \in \mathbb{R}^{h \times |V_t|}$ to $E'_s \in \mathbb{R}^{h \times |V_s|}$. The language model is thus represented as $\mathbf{s}^{\text{out}} \sim \text{LM}(\mathbf{s}^{\text{in}}, (E_s, \mathbf{De}, E'_s))$.

Since SpeechLMs inherit the language model architecture from TextLMs, it is possible for them to handle both text and speech modalities. The common approach is to expand the TextLM's vocabulary to include both text and speech tokens by appending the speech embedding matrix to the text embedding matrix, creating $E_m \in \mathbb{R}^{(|V_t|+|V_s|) \times h}$. This allows the language model to process combined token sequences $\mathbf{m}$ as $\mathbf{m}^{\text{out}} \sim \text{LM}(\mathbf{m}^{\text{in}}, (E_j, \mathbf{De}, E'_j))$, enabling joint text-speech applications. Alternatively, when using **continuous tokens**, the speech embeddings are fed directly into the language model without changing the embedding layer architecture.

Researchers predominantly favor the joint modeling approach for two main reasons. First, integrating both speech and text tokens enables the model to leverage pre-trained text language models, effectively transferring text-based knowledge and capabilities to speech-related tasks. Second, maintaining the ability to process text tokens ensures that the model can still perform text generation, which is crucial for developing omni-modal LLMs such as VITA (Fu et al., 2024) and EMOVA (Chen et al., 2024a). However, there are also motivations for developing models that solely include speech tokens. Such models aim to develop textless speech language models, which focus on building speech intelligence without relying on textual supervision or guidance.

3.3 Token-to-Speech Synthesizer (Vocoder)

The token-to-speech synthesizer (vocoder) converts generated speech tokens back into audible waveforms. This process reverses the speech tokenization and can be expressed as $\mathbf{a} = Vo(\mathbf{s}; \theta_{V_o})$, where Vo is the vocoder model with parameters θ_{V_o}.

SpeechLM vocoder can operate through two pipelines: direct synthesis and input-enhanced synthesis. **Direct synthesis** converts speech tokens straight into audio waveforms. For example, (Polyak et al., 2021) adapts the HiFi-GAN (Kong et al., 2020) architecture and takes speech tokens as inputs. **Input-enhanced synthesis** uses an additional module to transform tokens into continuous latent representations before vocoding (Anastassiou et al., 2024; Betker, 2023). For example, CosyVoice (Du et al., 2024b) uses Conditional Flow-Matching (CFM) to convert tokens to mel-spectrograms before HiFi-GAN vocoding. While direct synthesis offers simplicity and speed advantages, the choice between these pipelines largely depends on the input token type: acoustic tokens work well with direct synthesis due to their sufficient acoustic information, whereas semantic tokens, lacking fine acoustic details (especially in higher frequencies), benefit from being enhanced into acoustic-rich representations like mel-spectrograms before final synthesis.

Vocoders can be categorized by architectural choice. Below, we summarize those commonly used in SpeechLM development. See Appendix B for additional types.

3.3.1 GAN-based Vocoder

Generative Adversarial Networks (GANs) are widely used as SpeechLM vocoders for their fast, high-quality speech synthesis (Kumar et al., 2019; Kong et al., 2020; Polyak et al., 2021). GANs consist of a generator that produces realistic audio from noise or input features and a discriminator that assesses the authenticity of the generated audio against real samples. GAN-based vocoders incorporate inductive biases for generating audio waveforms. MelGAN (Kumar et al., 2019) uses residual blocks with dilations in the generator to capture long-range correlations in audio and introduces a multi-scale discriminator to handle different audio frequency ranges. HiFi-GAN (Kong et al., 2020) extends this with a multi-period discriminator to model diverse periodic patterns in audio waveforms. Fre-GAN (Kim et al., 2021b) employs the Discrete Wavelet Transform (DWT) to downsample and learn spectral distributions across frequency bands, offering an efficient alternative to Average Pooling (AP) by decomposing signals into low-frequency and high-frequency sub-bands. BigVGAN (Lee et al., 2023) introduces a snake activation function and anti-aliased representation to minimize high-frequency artifacts in synthesized audio. For common loss functions for GAN-based Vocoders, readers can refer to Appendix A due to space limits.

4 Training Recipes

This section summarizes common training recipes in recent SpeechLM papers, covering the features

modeled, techniques employed in each training stage, and different speech generation paradigms.

4.1 Features Modeled

This section discusses the different features outputted by speech tokenizers and modeled by SpeechLMs. These features represent different aspects of speech waveforms and determine the performance of SpeechLMs.

4.1.1 Discrete Features

Discrete features (or discrete tokens) are quantized speech representations in the form of distinct, countable tokens, derived through encoding and quantization. SpeechLMs commonly use these features because their modeling process is the same as that of text tokens in a TextLM.

Most SpeechLMs use **semantic tokens** (generated by semantic understanding tokenizers, Section 3.1.1) to represent speech, as they capture crucial contextual information. GSLM (Lakhotia et al., 2021), the first SpeechLM, evaluates three tokenizers—CPC (Oord et al., 2018), wav2vec 2.0 (Baevski et al., 2020b), and HuBERT (Hsu et al., 2021)—and concludes HuBERT performs best on tasks like speech resynthesis and generation. Many works adopt HuBERT as the speech tokenizer (Hassid et al., 2024; Nguyen et al., 2024; Zhang et al., 2023a). AudioPaLM (Rubenstein et al., 2023) compares w2v-bert (Chung et al., 2021), USM-v1 (Zhang et al., 2023b), and USM-v2 (Rubenstein et al., 2023), concluding USM-v2 excels in ASR and Speech Translation (ST) tasks.

While semantic tokens generate semantically meaningful speech, they lack expressive elements like prosody and timbre (Nguyen et al., 2023a, 2024). To address this, **paralinguistic tokens** can be added to capture expressive features. pGSLM (Kharitonov et al., 2022) integrates prosody features like fundamental frequency (F0) and unit duration along with HuBERT tokens, using a multi-stream transformer to predict all tokens. Similarly, SPIRIT-LM (Nguyen et al., 2024) complements HuBERT tokens with pitch and style tokens (Duquenne et al., 2023), improving expressiveness without compromising semantic understanding.

Acoustic tokens, derived from neural audio codec models (Section 3.1.2), aim to capture acoustic features for high-fidelity speech reconstruction. For instance, Viola (Wang et al., 2024c) handles ASR, TTS, and Machine Translation using codec tokens, while Parrot (Meng et al., 2024) leverages VQ-VAE (Van Den Oord et al., 2017) tokens to model dual-channel spoken dialogue.[3]

4.1.2 Continuous Features

Continuous features are unquantized, real-valued speech representations on a continuous scale, such as spectral representations (e.g., mel-spectrograms) or latent representations from neural networks. Spectron (Nachmani et al., 2024) predicts spectrograms frame-by-frame for speech continuation. Mini-Omni (Xie and Wu, 2024a) and SLAM-Omni (Chen et al., 2024b) use intermediate representations from a frozen Whisper encoder, while LauraGPT (Du et al., 2023) employs a co-trained audio encoder and language model to extract latent speech representations. Continuous features capture fine-grained speech details often lost in discretization but require modifying traditional text-based language model pipelines. Additionally, they demand more storage than discrete features.

4.2 Training Stages

Training a SpeechLM involves three components: speech tokenizer, language model, and vocoder. This section focuses on the primary techniques used in training the language model, as the language model plays a crucial role in generating speech continuations, which are central to SpeechLM. We divide the SpeechLM language model training process into three stages: pre-training, instruction tuning, and post-alignment.

4.2.1 Language Model Pre-training

Pre-training in SpeechLMs is crucial for generating coherent and contextually relevant speech. It involves training the model to autoregressively predict the next token using a large corpus of speech data, thereby learning statistical patterns and dependencies to generate speech based on context.

Training data. SpeechLMs pre-training utilizes large-scale open-source speech datasets, including those for ASR (Panayotov et al., 2015; Kahn et al., 2020; Wang et al., 2021a), TTS (Zen et al., 2019), ST (Jia et al., 2022; Wang et al., 2021a), podcasts (Clifton et al., 2020), and dialogues (Cieri et al., 2004). Some datasets contain only speech, while others include both speech and text transcripts, enabling models to learn the relationship between spoken and written language. Table 2 lists popular datasets used in SpeechLMs pre-training.

[3] More discussions on discrete features are in Appendix C.

Cold Initialization. Some SpeechLMs use cold initialization during pre-training, where model parameters are randomly initialized. The first SpeechLM, GSLM (Lakhotia et al., 2021) trained a transformer (Vaswani et al., 2017) from scratch as the language model and compared tokens from different speech tokenizers. They found that HuBERT (Hsu et al., 2021) outperformed CPC (Oord et al., 2018) and wav2vec 2.0 (Baevski et al., 2020b) in speech understanding and generation. SUTLM (Chou et al., 2023) also employed a transformer and explored joint modeling of speech and text using four methods: speech-only, text-only, concatenated speech-text, and alternating (interleaving) speech-text (Table 3). Alternating speech-text performs the best in their cross-modal evaluations.

Some models use non-standard transformer architectures and are usually trained from scratch when they differ significantly from standard transformers or TextLM architectures. For instance, pGSLM (Kharitonov et al., 2022) introduces a multi-stream transformer language model (MS-TLM) to simultaneously generate speech units, duration, and pitch embeddings. dGSLM (Nguyen et al., 2023b) presents a dialogue transformer language model (DLM) for jointly modeling dialogue speech data from two speakers. LSLM (Ma et al., 2024) integrates a streaming self-supervised learning (SSL) encoder with an autoregressive token-based TTS model to enable SpeechLMs to listen while speaking.

Continued Pre-Training. Continued pre-training starts with pre-trained TextLM weights, adapting them for speech modeling. This leverages linguistic knowledge in TextLMs for more efficient SpeechLM training. TWIST (Hassid et al., 2024) showed that TextLMs like OPT (Zhang et al., 2022b) and LLaMA (Touvron et al., 2023a) improve convergence and speech understanding, outperforming cold initialization. AudioPaLM (Rubenstein et al., 2023) demonstrated that larger TextLM checkpoints (e.g., PaLM, PaLM-2 (Chowdhery et al., 2023; Anil et al., 2023a)) and datasets further enhance SpeechLM performance.

Aligning text and speech modality representations can further boost the performance of SpeechLMs. One approach aligns text and speech in a **single sequence**. SPIRIT-LM (Nguyen et al., 2024) demonstrated that interleaving text and speech tokens during pretraining improves performance on speech tasks and increases text-speech feature similarity. SpeechGPT (Nachmani et al., 2024) aligns representations by enabling the SpeechLM to answer step-by-step: transcribing input speech to text, predicting text responses, and synthesizing speech. Another method uses **multi-sequence** alignment, where text and speech sequences are generated simultaneously. Mini-Omni (Xie and Wu, 2024a) produces one text token sequence and seven acoustic token sequences, aligned at the sentence level. Similarly, Moshi (Défossez et al., 2024) generates one text token sequence, one semantic token sequence, and seven acoustic token sequences, aligned at the word level. Further discussion on speech-text representation alignment is in Appendix D.

4.2.2 Language Model Instruction-Tuning

Instruction-tuning fine-tunes SpeechLMs to follow instructions for various tasks. This phase is crucial for enhancing the pre-trained model's generalization capabilities and making it more adaptable to diverse applications.

Effective instruction-following datasets are crucial in instruction-tuning. SpeechGPT (Zhang et al., 2023a) and SpeechGPT-Gen (Zhang et al., 2024b) propose a two-stage instruction-tuning process: (1) cross-modal instruction fine-tuning, where instructions are appended to paired ASR data, asking the model to convert speech into text (or vice versa). (2) chain-of-modality fine-tuning, where text-based instruction datasets are synthesized into speech-in-speech-out datasets using TTS. To more closely resemble spoken language pattern, Llama-Omni (Fang et al., 2024) synthesizes text-based instruction data by reformatting text prompts to mimic natural speech, generating responses via a TextLM, and synthesizing prompt-response pairs using TTS.

4.2.3 Language Model Post-Alignment

Post-alignment refines a language model's behavior to align with human preferences, ensuring safe and reliable outputs. As the final training phase, it often uses preference alignment algorithms like Reinforcement Learning from Human Feedback (RLHF) (e.g., Proximal Policy Optimization (PPO) (Schulman et al., 2017)) or Direct Preference Optimization (DPO) (Rafailov et al., 2023).

Post-alignment in SpeechLMs addresses unique challenges in the speech interaction pipeline. Align-SLM (Lin et al., 2024) identifies that SpeechLMs often generate inconsistent semantic content. They propose to use a TextLM to select preferred responses from SpeechLMs (transcribed via ASR)

and align preferences with DPO. SpeechAlign (Zhang et al., 2024a) argues that sub-optimal speech tokens generated by SpeechLM degrade the generated audio quality. They improve it by aligning model-generated tokens to the "golden" token distribution. Despite its importance, the post-alignment of SpeechLMs remains under-explored. Future research should prioritize identifying and addressing the unique safety challenges posed by SpeechLMs (see Section 6.3).

4.3 Speech Interaction Paradigm

Most earlier SpeechLM approaches follow the **traditional speech interaction paradigm**, which involves taking a complete input sequence and generating a complete response. However, this approach does not mirror natural conversations, where participants may interrupt one another. Therefore, some studies attempt to equip SpeechLM with **real-time interaction** ability.

Real-time interaction of SpeechLMs involves the advanced handling of conversation data from two or more people, and it can be understood through several progressive stages. The initial stage is the adoption of **streaming tokenizers and vocoders**, which eliminate the need for the language model to wait for complete speech encoding before processing. This architecture enables immediate, low-latency responses to user queries, marking a significant improvement over the traditional interaction paradigm. Nonetheless, while this streaming approach supports basic real-time interaction, it remains insufficient for capturing the more sophisticated interaction patterns observed in natural conversation. The next frontier is **full-duplex modeling**, which allows SpeechLMs to support simultaneous bidirectional communication—specifically, the ability to handle interruptions initiated by either the user or the model. It mainly includes two features: 1) User interruption, where models can be interrupted and respond appropriately to new instructions during a conversation, and 2) Simultaneous response, enabling models to process input and generate output concurrently. Achieving this requires the joint modeling of both user and model audio streams. dGSLM (Nguyen et al., 2023b) employs a separate transformer for each participant in two-speaker dialogues, with cross-attention layers capturing speaker interactions. Most methods, however, rely on a single language model. Parrot (Meng et al., 2024) employs a "next-token-pair prediction" approach with a decoder-only Transformer to predict tokens for both channels. Moshi (Défossez et al., 2024) concatenates user input and model response channels data, using an RQ-Transformer to process the data together. LSLM (Ma et al., 2024) focuses on modeling one speaker's speech using a decoder-only Transformer, integrating a streaming SSL encoder to fuse listening and speaking channel embeddings.

5 Evaluations

Similar to TextLMs, SpeechLMs possess diverse capabilities, making model comparisons difficult. Therefore, evaluating them from multiple perspectives is crucial. This section reviews common evaluation methods and benchmarks (Table 10) for SpeechLMs, categorized into automatic and human assessments with distinct aspects.

5.1 Automatic (Objective) Evaluation

Automatic evaluation methods are crucial for quick, consistent assessments of SpeechLMs, using quantitative metrics computed without needing humans. Common techniques are outlined below.

Representation Evaluation. Representations (embeddings) are hidden vectors that represent input/output data in a lower-dimensional space. It lays a foundation for the understanding and generation abilities of the models. For SpeechLMs, representation evaluation measures how effectively the model converts speech features into meaningful vectors. The *between-speaker ABX score*, used by GSLM (Lakhotia et al., 2021), assesses phonetic category separation by comparing three sound samples: two from one category and one from another. Additionally, speech resynthesis evaluation (Lakhotia et al., 2021) involves encoding speech into tokens, reconstructing it back to speech, and measuring the word or character error rates between original and reconstructed versions.

Linguistic Evaluation. Linguistic evaluation methods, covering lexical, syntactic, and semantic aspects, test a model's ability to understand and generate words, sentences, and meaningful content. Benchmark datasets include sWUGGY (Nguyen et al., 2020) for distinguishing real words from non-words, sBLIMP (Nguyen et al., 2020) for identifying grammatically correct sentences, and Spoken StoryCloze (Hassid et al., 2024) for selecting the semantically-coherent story ending.

Paralinguistic Evaluation. Paralinguistic evaluation examines non-verbal aspects of communi-

cation in speech. SpeechLMs that integrate paralinguistic tokens into their modeling can be evaluated at the token level. For instance, pGSLM (Kharitonov et al., 2022) assesses prosodic tokens based on accuracy (using minimum MAE), consistency (using Pearson correlation), and expressiveness (using standard deviation) by comparing the generated token to the reference token. At the perceptual level, SPIRIT-LM (Nguyen et al., 2024) introduced the STSP benchmark, which measures how well sentiment is preserved in generated speech or text based on the prompt. This approach could also be adapted to evaluate other paralinguistic features, such as timbre or prosody.

Generation Quality and Diversity. Quality and diversity are two crucial aspects of model generation. Typically, there is a trade-off between these dimensions when sampling model responses at different temperatures, so GSLM (Lakhotia et al., 2021) suggests using the Area Under the Curve (AUC) with various temperature values. Specifically, AUC on perplexity and VERT are employed to assess these factors, where VERT represents the geometric mean of the ratio of k-grams in the generated speech that repeat at least once. Additionally, the ChatGPT score can be utilized to evaluate the quality of the generated speech. In this process, the generated speech is transcribed using state-of-the-art ASR models and then sent to ChatGPT for quality (and diversity) assessment.

Real-time Interaction Evaluation. Real-time interaction evaluation assesses SpeechLMs' ability to interact in real-time, essential for full-duplex communication. Existing research focuses on assessing the naturalness and usefulness of the real-time interaction speech. dGSLM (Nguyen et al., 2023b) measures naturalness by analyzing turn-taking events (e.g., speech segments, pauses, gaps, overlaps), with speech being more natural if these statistics resemble human dialogues. Another approach evaluates speech continuation, where naturalness depends on alignment between turn-taking statistics of prompts and continuations. Usefulness is also critical in real-time interaction scenarios. The Parrot model (Meng et al., 2024) introduces reflective pauses (SpeechLM remains silent while the user speaks) and interruptions (SpeechLM stops when interrupted) to assess real-time interaction quality. Additionally, there are two benchmarks specifically designed for evaluating full duplex modeling of SpeechLMs. Arora et al. (Arora et al., 2025) introduce a predictive model-based evaluation where SpeechLM quality is determined by alignment between generated turn-taking events and those predicted by a specialized evaluation model. Full-Duplex-Bench (Lin et al., 2025) advances the field by evaluating four critical aspects of conversational dynamics: pause handling, backchanneling, turn-taking, and interruption management—each with dedicated metrics that provide nuanced assessment.

Downstream Evaluation. Downstream evaluation assesses SpeechLMs' performance on specific tasks like ASR, TTS, and Speaker Identification, either through few-shot examples or direct instruction. Various benchmarks have emerged to provide comprehensive assessment: SUPERB (Yang et al., 2021) focuses on speech understanding, SD-Eval (Ao et al., 2024) tests paralinguistic comprehension, SALMON evaluates speech generation consistency, VoiceBench (Chen et al., 2024c) evaluates SpeechLM general capabilities, while Dynamic-SUPERB (Huang et al., 2024a), MMAU (Sakshi et al., 2024), AirBench (Yang et al., 2024c), and AudioBench (Wang et al., 2024a) incorporate speech, sound, and music-related tasks. However, these benchmarks primarily require text responses. To address this limitation, VoxEval (Cui et al., 2025) supports the evaluation of speech output, making it more suitable for end-to-end speech interaction evaluation.

5.2 Human (Subjective) Evaluation.

Human evaluation is vital for assessing SpeechLMs, as speech is meant to be heard and understood by humans. Below are common human evaluation methods for SpeechLMs.

Mean Opinion Score (MOS) is a key metric in speech evaluation that measures perceived speech quality through human listener ratings on a 1-5 scale, with variations like MMOS, PMOS, and SMOS (Kharitonov et al., 2022; Zhang et al., 2024b) focusing on naturalness, prosody, and timbre similarity respectively. MOS is the averaged human-rated score for each audio sample. While human evaluation is crucial, the challenges in recruiting participants and gathering evaluations have led researchers to increasingly adopt machine-based evaluations. For example, a naturalness prediction model (Mittag et al., 2021) can assess the naturalness of generated outputs.

6 Challenges and Future Directions

Research on SpeechLMs shows promise but remains in the early stages. This section explores challenges and future directions in the field.

6.1 Understanding Different Component Choices

Current research on SpeechLMs encompasses key components such as speech tokenizers, language models, and vocoders, each offering a diverse range of options. While some studies have compared various component choices—primarily focusing on speech tokenizers—the comparisons tend to be limited in scope and depth (Lakhotia et al., 2021; Rubenstein et al., 2023). Consequently, there remains a significant gap in understanding the advantages and disadvantages of different component selections. Therefore, studies aimed at comprehensively comparing these choices are essential. Such an investigation would yield valuable insights and serve as a guide for selecting more efficient components when developing SpeechLMs.

6.2 End-to-End Training

Although SpeechLMs can generate speech directly without relying on text signals, some studies train the three components separately. This separate optimization may hinder the model's overall potential. Consequently, it would be worthwhile to investigate whether training can be conducted in an end-to-end manner, allowing gradients to be back-propagated from the vocoder's output to the tokenizer's input. By exploring this fully end-to-end approach, we could potentially enable SpeechLMs to produce more coherent, contextually relevant, and high-fidelity speech outputs.

6.3 Safety Risks in SpeechLMs

Safety is a crucial topic in Machine Learning, especially for large-scale generative AI models. Unlike TextLMs, the safety concerns in SpeechLMs remain underexplored. SpeechLMs share some safety challenges with TextLMs but also have unique issues, as noted in OpenAI's report on GPT-4o's voice mode (OpenAI, 2024). Future research must address vulnerabilities in SpeechLMs to enhance their safety. Key safety concerns in SpeechLMs include **toxicity** and **privacy**. Toxicity involves generating harmful content, such as dangerous instructions or inappropriate speech like erotic audio (OpenAI, 2024). Privacy risks include revealing personal information from speech input, such as inferring a speaker's identity, ethnicity, or beliefs, even when insufficient information is available (OpenAI, 2024).

6.4 Performance on Rare Languages

SpeechLMs directly model speech data, which allows them to more effectively handle "low-resource" languages compared to TextLMs. "Low-resource" languages are those that lack extensive textual data, making it challenging for TextLMs to model them efficiently. In contrast, SpeechLM provides a better solution by modeling the speech data of these "low-resource" languages, which often have more available audio data than text (Lakhotia et al., 2021). Therefore, future research could focus on training SpeechLMs in "low-resource" languages or dialects to expand their capabilities.

7 Conclusions

This survey provides a comprehensive overview of recent advancements in Speech Language Models (SpeechLMs). We begin by addressing the limitations of the naive framework that combines ASR, LLM, and TTS for speech interactions. Following this, we highlight the key advantages offered by SpeechLMs. We then explore the various components of SpeechLM architectures and outline their training methodologies, including the features they model and their distinct training stages. Lastly, we summarize the evaluation techniques used for SpeechLMs and discuss the primary challenges and promising future research directions in this field. We hope this survey will illuminate the field and assist the research community in creating more powerful Speech Language Models.

8 Limitations

Due to space constraints, several aspects of SpeechLMs could not be fully addressed in the main body of this survey. Firstly, the comprehensive taxonomy of SpeechLMs, which provides a detailed classification and organization of various models and approaches, has been placed in the appendix. Secondly, while the main text focuses on the most commonly used vocoders in SpeechLMs, the introduction of less frequently employed vocoder types had to be omitted. Lastly, some detailed discussions, such as the comparison of different features and the downstream tasks of SpeechLMs, have also been moved to the appendix.

Readers are encouraged to consult the appendix for a more comprehensive treatment of these aspects of SpeechLM research and development.

Acknowledgments

The research presented in this paper was partially funded by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 2410072, RGC R1015-23).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ehab A AlBadawy, Andrew Gibiansky, Qing He, Jilong Wu, Ming-Ching Chang, and Siwei Lyu. 2022. Vocbench: A neural vocoder benchmark for speech synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 881–885. IEEE.

Robin Algayres, Yossi Adi, Tu Nguyen, Jade Copet, Gabriel Synnaeve, Benoît Sagot, and Emmanuel Dupoux. 2023. Generative spoken language model based on continuous word-sized audio tokens. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3008–3028, Singapore. Association for Computational Linguistics.

Robin Algayres, Adel Nabli, Benoît Sagot, and Emmanuel Dupoux. 2022. Speech sequence embeddings using nearest neighbors contrastive learning. In *Interspeech 2022*, pages 2123–2127.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023a. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023b. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

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. *arXiv preprint arXiv:2406.13340*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. 2025. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. *arXiv preprint arXiv:2503.01174*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-supervised learning of discrete speech representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

James Betker. 2023. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021a. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 6, pages 4376–4380. International Speech Communication Association.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. 2024a. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*.

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 9th International Conference on Learning Representations (ICLR)*.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, et al. 2025. Minmo: A multimodal large language model for seamless voice interaction. *arXiv preprint arXiv:2501.06282*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, et al. 2024b. Slam-omni: Timbre-controllable voice interaction system with single-stage training. *arXiv preprint arXiv:2412.15649*.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024c. Voicebench: Benchmarking llm-based voice assistants. *CoRR*, abs/2410.17196.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. 2022b. Maestro: Matched speech text representations through modality matching. In *Proceedings of Interspeech*, pages 4093–4097.

Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. 2023. Toward joint language modeling for speech units and text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6582–6593, Singapore. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, and Sebastian Gehrmann. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pages 1086–1090.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Ann Clifton, Aasish Pappu, Sravana Reddy, Yongze Yu, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. The spotify podcast dataset. *arXiv preprint arXiv:2004.04270*.

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025. Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. *arXiv preprint arXiv:2501.04962*.

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

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. Technical report, Kyutai.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenpeng Du, Yiwei Guo, Feiyu Shen, Zhijun Liu, Zheng Liang, Xie Chen, Shuai Wang, Hui Zhang, and Kai Yu. 2024a. Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17924–17932.

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

Zhihao Du, Jiaming Wang, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, Chang Zhou, Zhijie Yan, and Shiliang Zhang. 2023. LauraGPT: Listen, Attend, Understand, and Regenerate Audio with GPT. https://arxiv.org/abs/2310.04673v4.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024c. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2019. The zero resource speech challenge 2019: Tts without t. In *Interspeech 2019*, pages 1088–1092.

Paul-Ambroise Duquenne, Kevin Heffernan, Alexandre Mourachko, Benoît Sagot, and Holger Schwenk. 2023. Sonar expressive: Zero-shot expressive speech-to-speech translation. *Meta AI Research*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2024. Audiochatllama: Towards general-purpose speech abilities for llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Zhifu Gao, Shiliang Zhang, Ming Lei, and Ian McLoughlin. 2020. San-m: Memory equipped self-attention for end-to-end speech recognition. *arXiv preprint arXiv:2006.01713*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. 2024. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36.

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.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. 2024a. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *arXiv preprint arXiv:2411.05361*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024b. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Wen-Chin Huang, Yi-Chiao Wu, Hsin-Te Hwang, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, Tomoki Toda, Yu Tsao, and Hsin-Min Wang. 2019. Refined wavenet vocoder for variational autoencoder based voice conversion. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE.

Wenyong Huang, Zhenhe Zhang, Yu Ting Yeung, Xin Jiang, and Qun Liu. 2022. Spiral: Self-supervised perturbation-invariant representation learning for speech pre-training. *arXiv preprint arXiv:2201.10207*.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *Preprint*, arxiv:2401.04088.

Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Soyoon Kim, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Jung-Woo Ha, et al. 2024. Paralinguistics-aware speech-empowered large language models for natural conversation. *arXiv preprint arXiv:2402.05706*.

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 Interspeech*, pages 2197–2201.

Yuma Koizumi, Kohei Yatabe, Heiga Zen, and Michiel Bacchiani. 2023. Wavefit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 884–891. IEEE.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. Diffwave: A versatile diffusion model for audio synthesis. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, and Jay Mahadeokar. 2024. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. 2022. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

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 11th International Conference on Learning Representations (ICLR)*.

Jiaqi Li, Dongmei Wang, Xiaofei Wang, Yao Qian, Long Zhou, Shujie Liu, Midia Yousefi, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, et al. 2024. Investigating neural audio codecs for speech language model-based speech generation. *arXiv preprint arXiv:2409.04016*.

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

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. 2025. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi

Lee, and Ivan Bulyko. 2024. Align-slm: Textless spoken language models with reinforcement learning from ai feedback. *arXiv preprint arXiv:2411.01834*.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2024. Language model can listen while speaking. *arXiv preprint arXiv:2408.02622*.

Gallil Maimon, Amit Roth, and Yossi Adi. 2024. A suite for acoustic language model evaluation. *arXiv preprint arXiv:2409.07437*.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330. IEEE.

Ziqiao Meng, Qichao Wang, Wenqian Cui, Yifei Zhang, Bingzhe Wu, Irwin King, Liang Chen, and Peilin Zhao. 2024. Parrot: Autoregressive spoken dialogue language modeling with decoder-only transformers. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. 2024. Pslm: Parallel generation of text and speech with llms for low-latency spoken dialogue systems. *arXiv preprint arXiv:2406.12428*.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proceedings of Interspeech 2021*, pages 2127–2131.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2024. Spoken question answering and speech continuation using spectrogram-powered llm. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *Proceedings of the Workshop on Self-Supervised Learning for Speech and Audio Processing*. NeurIPS.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. 2023a. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Proceedings of INTERSPEECH 2023*, pages 4823–4827.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023b. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, et al. 2024. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2024. Gpt-4o system card. Online; Accessed on 6-September-2024.

OpenBMB. 2024. Minicpm-o 2.6: A gpt-4.0 level mllm for vision, speech, and multimodal live streaming on your phone. `https://huggingface.co/openbmb/MiniCPM-o-2_6-int4`.

Jing Pan, Jian Wu, Yashesh Gaur, Sunit Sivasankaran, Zhuo Chen, Shujie Liu, and Jinyu Li. 2023. Cosmic: Data efficient instruction-tuning for speech in-context learning. *arXiv preprint arXiv:2311.02248*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Se Jin Park, Chae Won Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeong Hun Yeo, and Yong Man Ro. 2024. Let's go real talk: Spoken dialogue model for face-to-face conversation. *arXiv preprint arXiv:2406.07867*.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12442–12462, Bangkok, Thailand. Association for Computational Linguistics.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. In *Proc. Interspeech 2021*, pages 3615–3619.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

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

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yong Ren, Tao Wang, Jiangyan Yi, Le Xu, Jianhua Tao, Chu Yuan Zhang, and Junzuo Zhou. 2024. Fewer-token neural speech codec with time-invariant codes. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12737–12741. IEEE.

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, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of pre-trained models for the japanese language. *arXiv preprint arXiv:2404.01657*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Tongyi SpeechTeam. 2024. FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs. *arXiv preprint arXiv:2407.04051*.

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 12th International Conference on Learning Representations (ICLR)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Christophe Veaux, Junichi Yamagishi, and Simon King. 2013. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on*

Asian spoken language research and evaluation (O-COCOSDA/CASLRE), pages 1–4. IEEE.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21390–21402, Miami, Florida, USA. Association for Computational Linguistics.

Maarten Versteegh, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2016. The zero resource speech challenge 2015: Proposed approaches and results. *Procedia Computer Science*, 81:67–72. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, and Juan Pino. 2021b. Covost 2 and massively multilingual speech-to-text translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2024c. Viola: Conditional language models for speech recognition, synthesis, and translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xiong Wang, Yangze Li, Chaoyou Fu, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024d. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*.

Zhifei Xie and Changqiao Wu. 2024a. Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming. *arXiv preprint arXiv:2408.16725*.

Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*. Version 2, revised 3 Jan 2025.

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. *arXiv preprint arXiv:2305.02765*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, and Xixin Wu. 2023b. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024c. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

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, et al. 2021. Superb: Speech processing universal performance benchmark. In *Proc. Interspeech*, pages 1194–1198.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. 2024. Salmonn-omni: A codec-free llm for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*.

Ze Yuan, Yanqing Liu, Shujie Liu, and Sheng Zhao. 2024. Continuous speech tokens makes llms robust multi-modality learners. *arXiv preprint arXiv:2412.04917*.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

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*, pages 1526–1530.

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

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. 2024b. Scaling speech-text pre-training with synthetic interleaved data. *arXiv preprint arXiv:2411.17607*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. 2022a. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE.

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

Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2024a. Speechalign: Aligning speech generation to human preferences. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. Poster presentation on December 11, 2024, from 11 a.m. to 2 p.m. PST.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024b. Speechgpt-gen: Scaling chain-of-information speech generation. *arXiv preprint arXiv:2401.13527*.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, and Chaohong Tan. 2024c. Omniflatten: An end-to-end gpt model for seamless voice conversation. *arXiv preprint arXiv:2410.17799*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xin Zhang, Xiang Lyu, Zhihao Du, Qian Chen, Dong Zhang, Hangrui Hu, Chaohong Tan, Tianyu Zhao, Yuxuan Wang, Bin Zhang, et al. 2024d. Intrinsicvoice: Empowering llms with intrinsic real-time voice interaction abilities. *arXiv preprint arXiv:2410.08035*.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024e. Speechtokenizer: Unified speech tokenizer for speech large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023b. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhisheng Zhong, Chengyao Wang, Yuqi Liu, Senqiao Yang, Longxiang Tang, Yuechen Zhang, Jingyao Li, Tianyuan Qu, Yanwei Li, Yukang Chen, et al. 2024. Lyra: An efficient and speech-centric framework for omni-cognition. *arXiv preprint arXiv:2412.09501*.

Yongxin Zhu, Dan Su, Liqiang He, Linli Xu, and Dong Yu. 2024. Generative pre-trained speech language model with efficient hierarchical transformer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1764–1775, Bangkok, Thailand. Association for Computational Linguistics.

A GAN-based Vocoder Loss Functions

To utilize GAN to synthesize high-fidelity speech, various training objectives are designed, focusing on different aspects. First, **GAN loss** is utilized as the fundamental objective for the operation of the generator and the discriminator. Specifically, the typical choice of the GAN loss for the generator (G) and discriminator (D) is to use the least squares loss function. The GAN loss for the generator ($\mathcal{L}_{\text{GAN}}(G)$) and the discriminator ($\mathcal{L}_{\text{GAN}}(D)$) are

$$\mathcal{L}_{\text{GAN}}(G) = \mathbb{E}_{ms}\left[(D(G(ms)) - 1)^2\right] \quad (1)$$

and

$$\mathcal{L}_{\text{GAN}}(D) = \mathbb{E}_{(x,ms)}\left[(D(x) - 1)^2 + (D(G(ms)))^2\right], \quad (2)$$

respectively. In these loss functions, x represents the ground truth audio and ms represents its mel-spectrogram. Second, most GAN-based vocoders synthesize speech waveform from mel-spectrograms, so **mel-spectrogram loss** is proposed to align the mel-spectrogram synthesized by the generator and the mel-spectrogram transformed from the ground-truth waveform, in order to improve the fidelity of the generated speech. Mel-spectrogram loss ($\mathcal{L}_{\text{Mel}}(G)$) works by minimizing the L1 distance between the two versions of mel-spectrograms mentioned above. Its formula is shown below:

$$\mathcal{L}_{\text{Mel}}(G) = \mathbb{E}_{(x,ms)}\left[\|\phi(x) - \phi(G(ms))\|_1\right], \quad (3)$$

where $\phi(\cdot)$ is the function to transform a waveform into the corresponding mel-spectrogram. Third, to further enhance the generation fidelity, **feature matching loss** ($\mathcal{L}_{FM}(G)$) is proposed to align the discriminator-encoded features of the ground truth sample and the generated sample with L1 distance, which has the following formula:

$$\mathcal{L}_{FM}(G) = \mathbb{E}_{(x,ms)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D^i(x) - D^i(G(ms))\right\|_1\right], \quad (4)$$

where $D^i(\cdot)$ and N_i denote the features and the number of features in the i-th layer of the discriminator, respectively.

B Other Vocoders

The variety of vocoders is not restricted to the ones mentioned in Section 3.3, as those are the ones commonly employed in SpeechLMs. This section briefly outlines other potential vocoder types that are seldom explored as a component in SpeechLMs.

GAN-based Neural Audio Codec. Given that many neural audio codecs employ a GAN architecture, they can be effectively discussed within the context of GAN-based vocoders. Similar to its role as a tokenizer, although the primary objective of neural audio codecs is for audio compression, the encoded compact token sequences capture the essential information buried in the audio waveforms and therefore the codec decoder can be leveraged as a vocoder in SpeechLMs to transform the tokens into speech waveforms. EnCodec (Défossez et al., 2023) uses a GAN architecture and proposes a novel generator including an encoder, a quantizer, and a decoder. The compressed audio representations are outputted by the quantizer by using Residual Vector Quantization (RVQ). Polyak *et al.* utilizes HiFi-GAN (Kong et al., 2020) as the vocoder backbone and proposes to disentangle the input features of a vocoder into distinct properties (Polyak et al., 2021), which include semantic tokens, pitch tokens, and speaker embeddings. Such a design choice enables the codec to better perform on pitch and speaker-related tasks such as voice conversion and F_0 manipulation.

Pure Signal Processing Vocoder. Pure signal processing vocoders are traditional methods that rely on deterministic algorithms rather than deep learning models to synthesize speech (Griffin and Lim, 1984; Morise et al., 2016). However, this kind of vocoders introduces noticeable artifacts in the synthesized audio and is thus rarely used.

Autoregressive Vocoder. Autoregressive vocoders generate audio waveforms one sample at a time, with each sample conditioned on the previously generated samples (Oord et al., 2016). This approach allows for high-quality audio synthesis due to its sequential nature and the ability to capture intricate temporal dependencies within the audio signal. However, the sequential generation process can be computationally expensive and time-consuming, making autoregressive models less efficient compared to parallelized methods like GAN-based vocoders.

Flow-based Vocoder. Flow-based vocoder aims to establish a series of invertible transformations that map a simple distribution, such as a Gaussian, to the complex distribution of audio samples. This mechanism allows for efficient sampling and density evaluation, enabling the model to synthesize audio in parallel rather than sequentially, which sig-

Approach	Speech Tokenizer	Language Model	Vocoder
Minmo (Chen et al., 2025)	SenseVoice (SpeechTeam, 2024)	Qwen2.5 (Yang et al., 2024b)	CosyVoice 2 (Du et al., 2024c)
Lyra (Zhong et al., 2024)	Whisper (Radford et al., 2023)	Qwen2-VL (Wang et al., 2024b)	HuBERT + HiFi-GAN
Flow-Omni (Yuan et al., 2024)	Whisper Encoder + Linear Projector	Qwen2 (Yang et al., 2024a)	Flow Matching (Transformer + MLP) + HiFi-GAN
SLAM-Omni (Chen et al., 2024b)	Whisper Encoder + Linear Projector	Qwen2	-
OmniFlatten (Zhang et al., 2024c)	CosyVoice Encoder (Du et al., 2024b)	Qwen2	CosyVoice Decoder (Du et al., 2024b)
SyncLLM (Veluri et al., 2024)	HuBERT (Hsu et al., 2021)	LLaMA-3 (Dubey et al., 2024)	HiFi-GAN (Kong et al., 2020; Polyak et al., 2021)
EMOVA (Chen et al., 2024a)	SPIRAL (Huang et al., 2022)	LLaMA-3	VITS (Kim et al., 2021a)
Freeze-Omni (Wang et al., 2024d)	Transformer (Vaswani et al., 2017)	Qwen2	TiCodec (Ren et al., 2024)
IntrinsicVoice (Zhang et al., 2024d)	HuBERT	Qwen2	HiFi-GAN
Mini-Omni2 (Xie and Wu, 2024b)	Whisper	Qwen2	Mini-Omni (Xie and Wu, 2024a)
SALMONN-omni (Yu et al., 2024)	Mamba Streaming Encoder (Gu and Dao, 2024)	-	VoiceCraft (Peng et al., 2024) + Codec Decoder
Zeng et al. (Zeng et al., 2024b)	Whisper + VQ	GLM (GLM et al., 2024)	CosyVoice
Parrot (Meng et al., 2024)	VQ-VAE	LLaMA-3, Mistral, Gemma 2	HiFi-GAN
GPST (Zhu et al., 2024)	EnCodec (Défossez et al., 2023)	Transformer	Codec Decoder
GLM-4-Voice (Zeng et al., 2024a)	Whisper + VQ (Défossez et al., 2024)	GLM-4-9B-Base (GLM et al., 2024)	CosyVoice
Moshi (Défossez et al., 2024)	Mimi (Défossez et al., 2024)	Transformer*	Mimi
VITA (Fu et al., 2024)	CNN + Transformer + MLP (Fu et al., 2024)	Mixtral (Jiang et al., 2024)	Text-to-Speech Toolkit (Fu et al., 2024)
LSLM (Ma et al., 2024)	vq-wav2vec (Baevski et al., 2020a)	Decoder-Only Transformer	UniVATS (Du et al., 2024a)
SPIRIT-LM (Nguyen et al., 2024)	HuBERT, VQ-VAE (Van Den Oord et al., 2017), speechprop	LLaMA-2 (Touvron et al., 2023b)	HiFi-GAN
TWIST (Hassid et al., 2024)	HuBERT	OPT (Zhang et al., 2022b), LLaMA (Touvron et al., 2023a)	HiFi-GAN
PSLM (Mitsui et al., 2024)	HuBERT	NekoMata (Sawada et al., 2024)	HiFi-GAN
VOXTLM (Maiti et al., 2024)	HuBERT	OPT (Zhang et al., 2022b)	HiFi-GAN
Voicebox (Le et al., 2024)	EnCodec	Transformer* (Vaswani et al., 2017)	HiFi-GAN
Park et al. (Park et al., 2024)	AV-HuBERT (Shi et al., 2022)	OPT	HiFi-GAN
USDM (Kim et al., 2024)	XLS-R (Babu et al., 2021)	Mistral	Voicebox (Le et al., 2023)
VioLA (Wang et al., 2024c)	EnCodec	Transformer*	Codec Decoder (Défossez et al., 2023)
FunAudioLLM (SpeechTeam, 2024)	SAN-M (Gao et al., 2020)	Transformer*	HiFTNet (Li et al., 2023)
SpeechGPT-Gen (Zhang et al., 2024b)	SpeechTokenizer (Zhang et al., 2024e)	LLaMA-2	SpeechTokenizer decoder (Zhang et al., 2024e)
ICoT (**?**)	SpeechTokenizer	LLaMA-2	SoundStorm
AnyGPT (Zhan et al., 2024)	SpeechTokenizer	LLaMA-2	SoundStorm
LauraGPT (Du et al., 2023)	Conformer*	Qwen (Bai et al., 2023)	Transformer + Codec Decoder
Spectron (Nachmani et al., 2024)	Conformer*	PaLM 2* (Anil et al., 2023b)	WaveFit (Koizumi et al., 2023)
AudioLM (Borsos et al., 2023)	w2v-BERT (Chung et al., 2021)	Decoder-Only Transformer*	SoundStream* (Zeghidour et al., 2021)
UniAudio (Yang et al., 2023b)	EnCodec, Hifi-codec (Yang et al., 2023a), Improved RVQGAN (Kumar et al., 2023)	Transformer*	Codec Decoder
Llama-Omni (Fang et al., 2024)	Whisper	LLaMA-3.1	HiFi-GAN
Mini-Omni (Xie and Wu, 2024a)	Whisper + ASR Adapter (Xie and Wu, 2024a)	Qwen2	TTS Adapter (Xie and Wu, 2024a)
tGSLM (Algayres et al., 2023)	Segmentation + SSE (Algayres et al., 2022) + Lexical embedder	Transformer*	Tacotron-2 + Waveglow (Shen et al., 2018; Prenger et al., 2019)
SpeechGPT (Zhang et al., 2023a)	HuBERT	LLaMA	HiFi-GAN
dGSLM (Nguyen et al., 2023b)	HuBERT	Dialogue Transformer (Nguyen et al., 2023b)	HiFi-GAN
SUTLM (Chou et al., 2023)	HuBERT	Transformer*	-
pGSLM (Kharitonov et al., 2022)	HuBERT	MS-TLM (Kharitonov et al., 2022)	HiFi-GAN
GSLM (Lakhotia et al., 2021)	HuBERT, CPC (Oord et al., 2018), Wav2vec 2.0 (Baevski et al., 2020b)	Transformer*	Tacotron-2 + Waveglow

Table 1: Summarization of the architectural choice of speech tokenizer, language model, and vocoder in popular SpeechLMs. "-" represents non-existence or not indicated, * means the architecture is mainly based on the written one, "A, B" means the authors experimented with both "A" and "B" as the component, and "A + B" means "A" and "B" are combined to serve as the component.

nificantly enhances both speed and quality (Prenger et al., 2019). Compared to GAN-based vocoders, Flow-based vocoders typically need more parameters and memory to train the model, which hinders them from being effectively utilized (Kumar et al., 2019) in SpeechLMs.

VAE-based Vocoders. Variational Autoencoders (VAEs) are powerful generative models that learn to encode input data into a compressed latent space while allowing for the reconstruction of the original data (Van Den Oord et al., 2017; Huang et al., 2019). However, VAE is seldom explored as the underlying architecture of vocoders.

Diffusion-based Vocoder. Diffusion models have emerged in recent years as a powerful class of generative models that can be used for high-fidelity speech synthesis. They work by gradually adding noise to the input data (e.g. audio waveforms) to create a sequence of increasingly noisy representations, then learning to reverse this process to generate new samples (Kong et al., 2021; Chen et al., 2021b; Lee et al., 2022). For instance, DiffWave (Kong et al., 2021) uses Denoising Diffusion Probabilistic Models (DDPM) to synthesize audio. Additionally, CosyVoice (Du et al., 2024b) introduces a Conditional Flow-Matching (CFM) model that serves as a vocoder in TTS systems.

C Discussion on Different Discrete Features

The choice of discrete feature types (see Section 4.1.1) used for training significantly affects the quality of speech generated by SpeechLMs, often resulting in trade-offs (Borsos et al., 2023). For example, while semantic tokens align well with text and excel in producing semantically coherent speech, the generated speech often lacks acoustic details, such as high-frequency information. Recovering and enhancing these details typically requires post-processing, like a diffusion model, which significantly increases the model's latency. Conversely, acoustic tokens can facilitate the generation of high-fidelity audio but often struggle with inaccuracies in content generation (Zhang et al., 2024e). Researchers have tried two ways to balance these trade-offs. The first involves combining semantic and acoustic tokens into a single sequence. AudioLM (Borsos et al., 2023) proposes a hierarchical modeling scheme that first models semantic tokens from w2v-bert (Chung et al., 2021) and then uses these tokens to predict acoustic tokens from SoundStream (Zeghidour et al., 2021), which ultimately generates speech. However, this kind of approach increases sequence length, which increases modeling complexity and latency. The second strategy leverages *mixed tokens* (Section 3.1.3) to jointly model semantic and acoustic information, showing promising results in Moshi (Défossez et al., 2024) and SpeechGPT-Gen (Zhang et al., 2024b).

D Discussion on Speech-Text Representation Alignment

The primary goal of aligning text and speech representations is to leverage the strengths of text-based models to enhance speech-based models. Researchers have found that training a SpeechLM is significantly more challenging than training a TextLM. This difficulty arises because text serves as a concentrated form of knowledge, while speech requires models to independently learn the rules of spoken language. Aligning text and speech representations has demonstrated effectiveness, but it involves various trade-offs. First, text primarily conveys semantic information, which can improve a SpeechLM's semantic modeling capabilities but may compromise its ability to capture paralinguistic features, such as tone and emotion, during alignment. Second, there are two main inference approaches for the aligned models: text-present and text-independent. Text-present inference decodes text and speech simultaneously, which may increase latency but enhances the SpeechLM's reasoning abilities (Xie and Wu, 2024a) and reduces possible hallucinations (Défossez et al., 2024). Conversely, text-independent inference is more efficient but may lack stability. Furthermore, the question of whether to incorporate text modality to enhance SpeechLM performance remains an open question, especially considering that humans typically acquire spoken language skills before mastering written language.

E More on Speech Interaction Paradigm

In addition to real-time interaction (see Section 4.3), another advanced interaction ability of SpeechLMs is Interactive Period Recognition (IPR), which refers to the ability to recognize whether the users are interacting with it or not. SpeechLMs should provide response during the interactive period and remain silent during the non-interactive period. IPR is essential for creating a

Dataset	Type	Phase	Hours	Year
LibriSpeech (Panayotov et al., 2015)	ASR	Pre-Training	1k	2015
Multilingual LibriSpeech (Pratap et al., 2020)	ASR	Pre-Training	50.5k	2020
LibriLight (Kahn et al., 2020)	ASR	Pre-Training	60k	2019
People dataset (Galvez et al., 2021)	ASR	Pre-Training	30k	2021
VoxPopuli (Wang et al., 2021a)	ASR	Pre-Training	1.6k	2021
Gigaspeech (Chen et al., 2021a)	ASR	Pre-Training	40k	2021
Common Voice (Ardila et al., 2020)	ASR	Pre-Training	2.5k	2019
VCTK (Veaux et al., 2013)	ASR	Pre-Training	0.3k	2017
WenetSpeech (Zhang et al., 2022a)	ASR	Pre-Training	22k	2022
LibriTTS (Zen et al., 2019)	TTS	Pre-Training	0.6k	2019
CoVoST2 (Wang et al., 2021b)	S2TT	Pre-Training	2.8k	2020
CVSS (Jia et al., 2022)	S2ST	Pre-Training	1.9k	2022
VoxCeleb (Nagrani et al., 2017)	Speaker Identification	Pre-Training	0.4k	2017
VoxCeleb2 (Chung et al., 2018)	Speaker Identification	Pre-Training	2.4k	2018
Spotify Podcasts (Clifton et al., 2020)	Podcast	Pre-Training	47k	2020
Fisher (Cieri et al., 2004)	Telephone conversation	Pre-Training	2k	2004
SpeechInstruct* (Zhang et al., 2023a)	Instruction-following	Instruction-Tuning	-	2023
InstructS2S-200K* (Fang et al., 2024)	Instruction-following	Instruction-Tuning	-	2024
VoiceAssistant-400K* (Xie and Wu, 2024a)	Instruction-following	Instruction-Tuning	-	2024

Table 2: A summary of popular datasets used in the pre-training and instruction-tuning phase of SpeechLMs. * means it is the speech version of the text dataset synthesized using TTS. S2ST and S2TT represent speech-to-speech translation and speech-to=text translation, respectively.

Modeling Method	Example	Explanation
Speech-only	[SPEECH] S12 S34 S33 ... S11 S59	Only the speech sequence is provided.
Text-only	[TEXT] A quick brown fox jumps over a lazy dog.	Only the text sequence is provided.
Concatenated speech-text	[SPEECH] S12 S34 S33 ... S11 S59 [TEXT] A quick brown fox jumps over a lazy dog.	The speech sequence and text sequence are concatenated together.
Alternating speech-text	[SPEECH] S12 S34 S33 [TEXT] brown fox jumps over a lazy [SPEECH] S11 S59	The sequence is interleaved with speech and text tokens.

Table 3: Four different methods of jointly modeling speech and text tokens.

natural conversational flow, allowing the model to avoid unnecessary interruptions. It is crucial for situations where a small group of users is having a discussion, as the SpeechLM needs to discern when to join in and when to stay silent. Additionally, it is important for the model to learn when to disregard instructions when users are not speaking at it. One approach to achieving IPR is through a Voice Activity Detection (VAD) module. MiniCPM-o 2.6 (OpenBMB, 2024) integrates a VAD module to ensure the model responds only when the input audio surpasses a predefined VAD threshold. Inputs below this threshold are considered noise and ignored. VITA (Fu et al., 2024) takes a different approach by training the SpeechLM to distinguish between query speech and non-query audio. The model learns to output an end-of-sequence token to terminate its response when non-query audio is detected.

F Empirical Analysis of SpeechLM Performance

This section provides an overview of the empirical performance of SpeechLMs on widely used benchmarks. We begin by examining the overall performance of SpeechLM and then delve into the performance of the tokenizer and vocoder components of SpeechLMs.

F.1 Overall Performance

The overall performance of SpeechLM is typically assessed based on its capability to model linguistic information (as shown in Table 4) and its success in achieving high performance in spoken QA tasks (Table 5). The results illustrate a clear improvement in SpeechLM's performance across various model architectures and the influence of training data size. GSLM, which is trained entirely from scratch, serves as the baseline for comparison. TWIST demonstrates notable enhancements by uti-

Models	sWUGGY	sBLIMP	sTSC	sSC
GSLM	64.8	54.2	66.6	53.3
AudioLM	71.5	64.7	-	-
TWIST	72.2	56.5	-	-
SPIRIT-LM	69.0	58.3	82.9	61.0
Moshi	74.3	58.9	83.0	60.8
GLM-4-Voice	-	-	82.9	62.4

Table 4: Evaluation accuracy (%) of different SpeechLMs on common linguistic evaluation benchmarks. sTSC and sSC represent sTopic-StoryCloze and sStoryCloze, respectively. The results are adapted from the corresponding SpeechLM papers.

	Web-Questions		Llama-Questions		Trivia QA	
Models	$S \rightarrow T$	$S \rightarrow S$	$S \rightarrow T$	$S \rightarrow S$	$S \rightarrow T$	$S \rightarrow S$
GSLM	-	1.5	-	4.0	-	-
AudioLM	-	2.3	-	7.0	-	-
TWIST	-	2.2	-	0.5	-	-
SpeechGPT	6.5	-	21.6	-	14.8	-
Spectron	6.1	-	22.9	-	-	-
Moshi	26.6	9.2	62.3	21.0	22.8	7.3
GLM-4-Voice	32.2	15.9	64.7	50.7	39.1	26.5
Freeze-Omni	72.0	-	53.9	-	44.7	-
MinMo	55.0	39.9	78.9	64.1	48.3	37.5

Table 5: Evaluation accuracy (%) of different SpeechLMs on common spoken QA benchmarks in speech-to-text and speech-to-speech modes. The results are adapted from the corresponding SpeechLM papers.

lizing pre-trained TextLM checkpoints, showcasing the advantages of transfer learning from text-based models. SPIRIT-LM builds upon these improvements with its innovative interleaved speech-text alignment training strategy, enabling more effective cross-modal understanding. While SPIRIT-LM delivers impressive results through its alignment-based training, Moshi surpasses it by leveraging a much larger dataset comprising 7 million hours of speech data. Finally, GLM-4-Voice sets a new performance standard by pre-training on approximately 13 million hours of synthetic interleaved speech-text data.

F.2 Tokenizer Performance

Table 6 presents a comparison among three types of tokenizers: a semantic tokenizer (HuBERT), an acoustic tokenizer (EnCodec), and a mixed tokenizer (SpeechTokenizer). The first two metrics assess semantic performance, while the last two evaluate acoustic performance. As discussed in Section 3.1, the semantic tokenizer demonstrates superior semantic modeling, the acoustic tokenizer excels in acoustic modeling, and the mixed tokenizer strikes a balance between the two.

F.3 Vocoder Performance

The comparative analysis presented in Tables 7 and 8 demonstrates the trade-offs between different vocoder architectures across multiple evaluation dimensions. Table 7 reveals that autoregressive models (WaveNet and WaveRNN) achieve superior perceptual quality than the GAN-based vocoder (MelGAN) as measured by Mean Opinion Score (MOS). However, Table 8 illustrates the computational efficiency advantages of GAN-based vocoders. MelGAN demonstrates exceptional efficiency with only 3.05M parameters and 3.01 GFLOPS, achieving real-time factors of 0.001 on GPU and 0.029 on CPU—orders of magnitude faster than diffusion-based approaches. Parallel WaveGAN offers an optimal balance, maintaining competitive audio quality while requiring only 1.34M parameters and achieving RTF values of 0.002 (GPU) and 0.576 (CPU).

The results indicate that while autoregressive vocoders (and diffusion-based vocoders) excel in audio fidelity, GAN-based architectures provide the computational efficiency essential for real-time speech synthesis. This efficiency-quality trade-off explains the prevalent adoption of GAN-based vocoders in SpeechLMs, where low-latency inference is critical for interactive applications while maintaining acceptable perceptual quality.

G Technical Specifications of Common SpeechLM Components

This section expands on Section 3 by providing the technical notations and specifications for some of the commonly adopted components in SpeechLMs, including speech tokenizers and vocoders.

G.1 Speech Tokenizer

This subsection presents the notation for three representative speech tokenizers, each exemplifying a distinct category. We examine HuBERT as a semantic objective tokenizer, Encodec as an acoustic objective tokenizer, and SpeechTokenizer as a mixed objective tokenizer.

HuBERT. As a representative semantic objective tokenizer, HuBERT (Hsu et al., 2021) employs a feature encoder f_E to transform raw audio waveforms $\mathbf{a}$ into continuous embeddings $\mathbf{v}$, i.e., $f_E(\mathbf{a}; \theta_{f_E}) = \mathbf{v}$. These embeddings are then quantized into discrete speech tokens $\mathbf{s}$ via k-means clustering of MFCC features, denoted as $d(\text{MFCC}(\mathbf{a}); \theta_d) = \mathbf{s}$. The model is trained with

Tokenizer	Teacher	MI↑	WER↓	WER*↓	SIM↑
GroundTruth	-	-	-	4.58	1.0
HuBERT (semantic)	KM500	31.2	9.88	16.26	0.77
EnCodec (acoustic)	RVQ-1	16.5	61.52	38.34	0.92
EnCodec (acoustic)	RVQ-1:8	23.6	30.91	5.11	0.98
SpeechTokenizer (mixed)	RVQ-1 HuBERT avg	30.9	15.58	9.57	0.74
SpeechTokenizer (mixed)	RVQ-1:8 HuBERT avg	29.7	16.03	5.04	0.97

Table 6: Empirical Comparison results of different speech tokenizers (adapted from (Zhang et al., 2024e)). KM represents K-means. MI and WER represent Mutual Information and Word Error Rate. WER* and SIM represent word error rate and speaker similarity of resynthesized speech.

Metric	Corpus	WaveNet	WaveRNN	MelGAN	Parallel WaveGAN	WaveGrad	DiffWave	Griffin-Lim	Ground Truth
SSIM↑	LJ Speech	0.66	0.62	0.89	0.84	0.76	0.82	0.90	-
	LibriTTS	0.056	0.53	0.91	0.86	0.71	0.74	0.89	-
	VCTK	0.46	0.43	0.88	0.79	0.59	0.64	0.86	-
LS-MSE↓	LJ Speech	0.006	0.010	0.001	0.002	0.006	0.006	0.001	-
	LibriTTS	0.008	0.008	0.001	0.001	0.005	0.006	0.001	-
	VCTK	0.009	0.010	0.001	0.002	0.007	0.007	0.001	-
PSNR↑	LJ Speech	23.20	20.36	28.53	26.70	22.57	22.51	28.77	-
	LibriTTS	21.54	21.17	29.98	28.62	22.94	22.18	29.03	-
	VCTK	21.36	20.40	30.40	28.17	21.54	21.22	28.77	-
FAD↓	LJ Speech	1.05	3.43	1.51	0.92	3.12	3.62	2.69	0.31
	LibriTTS	1.55	2.60	2.95	1.41	3.10	3.74	4.27	1.23
	VCTK	0.99	3.59	1.76	1.22	4.10	5.59	3.92	0.61
MOS↑	LJ Speech	3.68±0.037	3.96±0.089	3.73±0.075	3.99±0.059	3.85±0.068	4.07±0.060	3.68±0.082	4.10±0.059
	LibriTTS	3.75±0.107	3.74±0.099	3.50±0.086	3.82±0.069	3.48±0.083	3.80±0.073	3.36±0.092	4.03±0.065
	VCTK	3.95±0.032	3.94±0.089	3.75±0.074	3.87±0.068	3.77±0.074	3.86±0.069	3.66±0.079	3.98±0.064

Table 7: Audio quality evaluation of vocoder models across three datasets (adapted from (AlBadawy et al., 2022)). SSIM, LS-MSE, PSNR, FAD, and MOS represent Structural Similarity Index Measure, Log-mel Spectrogram Mean Squared Error, Peak Signal-to-Noise Ratio, Fréchet Audio Distance, and Mean Opinion Score, respectively.

Model	#Param (M)	GFLOPS	RTF (GPU)	RTF (CPU)
WaveNet	3.79	89.65	-	-
WaveRNN	4.35	94.98	-	-
MelGAN	3.05	3.01	0.001	0.029
Parallel WaveGAN	1.34	31.26	0.002	0.576
WaveGrad	15.81	33.75	0.381	9.858
DiffWave	2.62	31.70	0.070	4.452

Table 8: Computational efficiency comparison of vocoder models (adapted from (AlBadawy et al., 2022)). It compares model complexity (parameters), computational requirements by Floating Point Operations per Second (GFLOPS), and inference speed measured by Real-Time Factor (RTF) on GPU and CPU platforms.

a masked prediction objective, which seeks to maximize the likelihood of the correct token at masked positions:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{a}\sim\mathcal{D}}\left[\sum_{i\in\mathcal{M}} -\log p(s_i \mid \mathbf{v}_{\setminus\mathcal{M}};\theta)\right], \quad (5)$$

where $\mathcal{M}$ denotes the masked indices. HuBERT further refines its speech tokens iteratively, updating the encoder and discretizer parameters at each step as

$$\mathbf{s}^{(n+1)} = d(f_E(\mathbf{a};\theta_{f_E}^{(n)});\theta_d^{(n)}). \quad (6)$$

This iterative process enables the learning of increasingly meaningful speech representations.

Encodec. As a representative acoustic objective tokenizer, EnCodec (Défossez et al., 2023) employs a convolutional encoder-decoder architecture with residual vector quantization (RVQ). The encoder f_E maps the raw audio waveform $\mathbf{a}$ to continuous embeddings $\mathbf{v}$, i.e., $\mathbf{v} = f_E(\mathbf{a};\theta_{f_E})$. These embeddings are then discretized using a multi-stage RVQ, where each stage r quantizes the residual from the previous stage:

$$\mathbf{s} = d(\mathbf{v};\theta_d) = \big(d_1(\mathbf{v};\theta_{d_1}),\ d_2(\mathbf{v}-\hat{\mathbf{v}}_1;\theta_{d_2}),\ \ldots,\ d_R(\mathbf{v}-\sum_{r=1}^{R-1}\hat{\mathbf{v}}_r;\theta_{d_R})\big), \quad (7)$$

with $\hat{\mathbf{v}}_r$ denoting the quantized embedding at stage r. The decoder f_D reconstructs the audio waveform from the quantized tokens, $\hat{\mathbf{a}} = f_D(\mathbf{s};\theta_{f_D})$.

This design enables EnCodec to produce discrete acoustic tokens that retain high-fidelity audio information suitable for downstream modeling.

SpeechTokenizer. As a representative mixed objective tokenizer, SpeechTokenizer (Zhang et al., 2024e) combines semantic and acoustic objectives by leveraging both HuBERT and residual vector quantization (RVQ) mechanisms. The encoder f_E first transforms the input audio waveform $\mathbf{a}$ into continuous embeddings $\mathbf{v}$, i.e., $\mathbf{v} = f_E(\mathbf{a}; \theta_{f_E})$. Discretization is performed via a multi-stage RVQ. The discretization process uses a multi-stage RVQ, which operates similarly to Encodec, except that the first RVQ stage distills tokens derived from HuBERT, while the subsequent stages quantize the residuals. This hybrid approach enables SpeechTokenizer to capture both high-level semantic and low-level acoustic information for robust speech representation learning.

G.2 Vocoder

We present the notation of HiFi-GAN (Kong et al., 2020) as it is the most used vocoder in SpeechLMs. HiFi-GAN synthesizes high-fidelity audio waveforms from mel-spectrograms or speech tokens using a generator-discriminator framework. The generator $G(\mathbf{s}; \theta_G)$ maps a sequence of speech tokens $\mathbf{s}$ to an output audio waveform $\mathbf{a}$, i.e.,

$$\mathbf{a} = Vo(\mathbf{s}; \theta_{Vo}) = G(\mathbf{s}; \theta_G), \tag{8}$$

where Vo denotes the vocoder function and $\theta_{Vo} = \theta_G$ are its parameters. HiFi-GAN employs multi-period and multi-scale discriminators, $D_{MPD}(\mathbf{a}; \theta_{MPD})$ and $D_{MSD}(\mathbf{a}; \theta_{MSD})$, to distinguish real from generated audio during adversarial training. At inference, only the generator G is used to efficiently reconstruct speech waveforms.

H Downstream Applications

SpeechLMs, unlike traditional ASR and TTS systems that focus on specific tasks, are generative foundation models capable of handling a variety of speech-only, text-only, and multi-modal tasks. This section explores their primary downstream applications, which include both traditional speech tasks and unique SpeechLM tasks. Unlike TextLMs which generate only semantic information, SpeechLMs can also model paralinguistic features like pitch and timbre, enhancing their capabilities. The applications of SpeechLMs are categorized into three main classes: semantic-related, speaker-related, and paralinguistic applications. We give an example of all the downstream applications in Table 9.

H.1 Semantic-Related Applications

Semantic-related applications involve key tasks that facilitate meaningful interactions between humans and machines. These applications require SpeechLMs to grasp the semantic meaning of input and generate responses that are contextually relevant and logically coherent. The primary Semantic-related applications of SpeechLMs are as follows.

Spoken Dialogue. Spoken dialogue is the most natural application of SpeechLMs. Spoken dialogue systems are designed to facilitate natural conversations between humans and machines in spoken format. They can engage users in interactive exchanges, understanding and generating responses based on the context of the conversation. Unlike TextLMs, SpeechLMs are able to perform conversations with humans directly in speech, which is a more natural way of communication. Note that SpeechLMs can not only perform speech-only dialogues but also perform cross-modal dialogues, such as taking texts as input and responding in speech format.

Speech Translation. Speech translation (ST) is the process of converting spoken language from one language to another. Similar to Spoken dialogue, SpeechLMs can perform ST in both single-modal and cross-modal settings. Specifically, the input and output of the ST task can be either in text or speech format.

Automated Speech Recognition. Automatic speech recognition (ASR) enables systems to convert spoken language into text. The input of ASR is a speech waveform, and the system outputs the transcription in textual form. For SpeechLMs, the input would be a combination of the speech waveform and the instruction to tell the model to perform ASR on the given speech.

Keyword Spotting. Keyword spotting can be considered a special type of ASR, where its primary objective is to identify specific words or phrases within continuous speech. While traditional ASR systems aim to transcribe entire spoken utterances into text, keyword spotting focuses specifically on identifying and extracting predefined keywords or phrases within continuous speech. The primary application of keyword spotting is to build voice-activated assistants in smart home devices. Those

Task	Input	Output
Semantic-related Applications		
Spoken Dialogue	Speech(I am so angry right now!)	Speech(Sorry to hear that. What happened?)
Speech Translation	Speech(Good morning my friend!) Please translate the speech into French.	Speech(Bonjour mon ami !)
Automatic Speech Recognition	Speech(A river is on your left side.) Please perform automatic speech recognition to the speech	The transcribed speech is: A river is on your left side.
Keyword Spotting	Speech(How's your day? (Speaking to somebody else) Hey, Siri.) Please detect if the given speech contains any pre-defined keywords.	Keyword detected: "Hey, Siri."
Text-to-Speech Synthesis	Please synthesize the following text into speech: "A river is on your left side."	Sure. Here is the synthesized speech: Speech(A river is on your left side.)
Intent Classification	Speech(Can you help me book a flight?) Please detect the intent of the given speech.	Intent detected: "Book a flight".
Slot Filling	Speech(I want to fly from New York to San Francisco on June 5th.) Please fill the following slots based on the given speech: departure city, destination city, and date	Sure. Here are the filled slots. Departure city: New York. Destination city: San Francisco. Date: June 5th.
Query by Example Spoken Term Detection	Speech(Can you retrieve the audio documents that contain the word "happy"?) [A batch of audio documents]	Sure. I have found documents 2 and 4 contain the word happy.
Speaker-related Applications		
Speaker Identification	The following speech comes from Lucy: "[Lucy's speech]". The following speech comes from Bob: "Bob's speech". Question: Who is the speaker of the following speech [Unknown speech]	The speaker of the following speech: "[Unknown speech]" is Bob.
Speaker Verification	Here are two speeches. Speech 1: [Speech 1]. Speech 2: [Speech 2]. Do they come from the same speaker?	No, they do not come from the same speaker.
Speaker Diarization	[An audio file of the meeting] Please provide the speaker diarization output in the following format: [Start Time] - [End Time]: Speaker [Speaker].	Sure. Here is the speaker diarization result: 00:00 - 00:15: Alice, 00:16 - 00:40: Bob, 00:41 - 01:00: Carol, 01:01 - 01:20: Alice, 01:21 - 01:45: Bob and Carol (overlapping speech), 01:46 - 02:00: Carol.
Voice-Conditioned Speech Generation	Can you use a robotic voice to say "Hello world!" to me?	Sure. Speech([In robotic voice] Hello world!)
Paralinguistic Applications		
Emotion Recognition	Speech(What are you doing) What emotion does the speech convey?	It seems that the speaker is pretty shocked and angry about something.
Speech Separation	Speech([Two overlapping speeches]) Can you separate the two speeches?	Sure. Here is the speech separation result. Speech 1: Speech(Should I go for this one?). Speech 2: Speech(That one seems good.)
Paralinguistics-Enhanced Generation	Speech(Please generate a speech that sounds happy.)	Speech([In happy mood] My friend just gave me a candy!!)

Table 9: Examples of the various capabilities of SpeechLMs.

Name	Eval Type	# Tasks	Audio Type	I/O
ABX (Versteegh et al., 2016; Dunbar et al., 2019; Nguyen et al., 2020)	Representation	1	Speech	$A \rightarrow -$
sWUGGY (Nguyen et al., 2020)	Linguistic	1	Speech	$A \rightarrow -$
sBLIMP (Nguyen et al., 2020)	Linguistic	1	Speech	$A \rightarrow -$
sStoryCloze (Hassid et al., 2024)	Linguistic	1	Speech	$A/T \rightarrow -$
STSP (Nguyen et al., 2024)	Paralinguistic	1	Speech	$A/T \rightarrow A/T$
MMAU (Sakshi et al., 2024)	Downstream	27	Speech, Sound, Music	$A \rightarrow T$
Audiobench (Wang et al., 2024a)	Downstream	8	Speech, Sound	$A \rightarrow T$
AIR-Bench (Yang et al., 2024c)	Downstream	20	Speech, Sound, Music	$A \rightarrow T$
SD-Eval (Ao et al., 2024)	Downstream	4	Speech	$A \rightarrow T$
SUPERB (Huang et al., 2024a)	Downstream	10	Speech	$A \rightarrow T$
Dynamic-SUPERB (Huang et al., 2024a)	Downstream	180	Speech, Sound, Music	$A \rightarrow T$
SALMON (Maimon et al., 2024)	Downstream	8	Speech	$A \rightarrow -$
VoiceBench (Chen et al., 2024c)	Downstream	8	Speech	$A \rightarrow T$
VoxEval (Cui et al., 2025)	Downstream	56	Speech	$A \rightarrow A$

Table 10: A summary of popular benchmarks for the evaluation of SpeechLMs. I/O, A, and T represent input/output format, audio, and text, respectively.

devices are activated when the specific keywords are triggered. Therefore, although SpeechLMs are capable of spotting and understanding more than just a couple of words, keyword spotting can be used to efficiently trigger SpeechLMs to respond to user inputs.

Text-to-Speech Synthesis. Text-to-speech synthesis (TTS) enables systems to synthesize written text into spoken language. In contrast to ASR, TTS takes text as input and outputs the converted speech waveform. Similarly, the input of the SpeechLMs would be a combination of the text to synthesize and the instruction, and the output is the synthesized speech.

Intent Classification. Intent classification is a critical task that identifies the underlying intention behind a user's input speech. The AI system can then perform certain actions based on the identified user intent (e.g., book a flight). Intent classification is particularly important in applications such as virtual assistants, customer service bots, and interactive voice response systems. To perform Intent Classification, it is more natural for SpeechLMs to take speech inputs and classify the results in text since it is easier to parse and classify the intent classification result in text than speech.

Slot Filling. Slot filling is an important task in spoken language understanding that involves identifying and extracting specific pieces of information from user inputs into predefined classes, such as intents, entities, and parameters that are essential for completing a task. For example, slot filling extracts the phrase "I want to fly from New York to San Francisco on June 5th." into distinct slots like "departure city" (New York), "destination city" (San Francisco), and "date" (June 5th). Similar to Intent Classification, it is more natural for SpeechLMs to take speech inputs and extract the pieces in texts.

Query by Example Spoken Term Detection. Another spoken term detection task is query by example spoken term detection (QbE-STD), which allows users to identify specific spoken terms or phrases within a larger audio stream by providing an example of the desired term. Unlike traditional keyword spotting methods that rely on predefined lists of keywords, QbE-STD leverages the flexibility of example-based querying, enabling users to specify their search terms through audio samples.

H.2 Speaker-Related Applications

Speaker-related applications refer to the tasks that involve the processing of information related to speaker identity. It could involve classification tasks such as identifying, verifying, and distinguishing individual speakers based on their unique vocal characteristics, as well as generation tasks such as maintaining or modifying the timbre of a given speech. While we acknowledge that voice characteristics can be considered paralinguistic information, we believe that speaker-related applications are unique because they enable SpeechLMs to function in complex scenarios such as participating in multi-speaker conversations. In this section, we survey common speaker-related applications of SpeechLMs.

Speaker Identification. Speaker identification is the process of recognizing a person's identity based on their voice characteristics. It is a multi-class classification of a given speech as input. SpeechLMs can perform this task by taking an input speech and outputting the classification result in text or speech format. Moreover, SpeechLMs can also identify different speakers implicitly. Specifically, it can chat with multiple speakers at the same time, distinguishing the words from different speakers and responding to each speaker appropriately.

Speaker Verification. Speaker verification involves determining whether the speakers of a pair of speeches match with each other. Unlike speaker identification, which is a multi-class classification process, speaker verification is a binary classification process.

Speaker Diarization. Speaker diarization is the process of partitioning an audio stream into segments according to the identity of the speakers. It predicts "who is speaking when" for each timestamp (Yang et al., 2021). A natural way to integrate speaker diarization into SpeechLMs is to have the model generate the transcript of each audio segment along with the identification of the speaker.

Voice-Conditioned Speech Generation. Voice-conditioned speech generation involves synthesizing speech based on the vocal characteristics of a specific speaker. This could involve voice cloning and voice conversion. Voice cloning utilizes a sample of the speaker's voice as a reference, enabling the model to reproduce the speaker's timbre when generating speech from input text. Voice conversion, on the other hand, modifies an existing speech signal to sound like it was produced by a different speaker while retaining the original content. Additionally, instead of giving the target vocal characteristics, SpeechLMs should also be able to adapt their output timbre based on various speech or text

instructions.

H.3 Paralinguistic Applications

Paralinguistics refers to the non-verbal elements of communication that accompany spoken language, including vocal attributes that convey meaning beyond the words themselves. These elements, such as pitch, timbre, rate of speech, and pauses, significantly influence the message interpretation. Additionally, variations in these elements can evoke different emotions, so we include emotion-related tasks as part of paralinguistic applications.

Emotion Recognition. Emotion recognition task involves identifying and classifying the emotion carried by a given speech into predefined classes. Similar to speaker identification, SpeechLMs are capable of not only directly performing this task but also implicitly recognizing users' emotions through their speech queries and responding accordingly.

Speech Separation. Speech separation refers to the process of isolating individual speech signals from a mixture of sounds, such as when multiple speakers are talking simultaneously. When separating the input speech, SpeechLMs can not only output the contents of each person in speech but also in text format (i.e., transcriptions).

Paralinguistics-Enhanced Generation. Paralinguistics-enhanced generation refers to the process of instructing SpeechLMs to produce speech that exhibits specific paralinguistic characteristics. Users can define these characteristics in their prompts, allowing the model to generate speech that aligns with their specifications. Examples of paralinguistics-enhanced generation include synthesizing speech with a specific style, speaking at a fast pace, and even singing. This capability distinguishes SpeechLMs from TextLMs and facilitates a more engaging and interactive form of communication with the AI models.

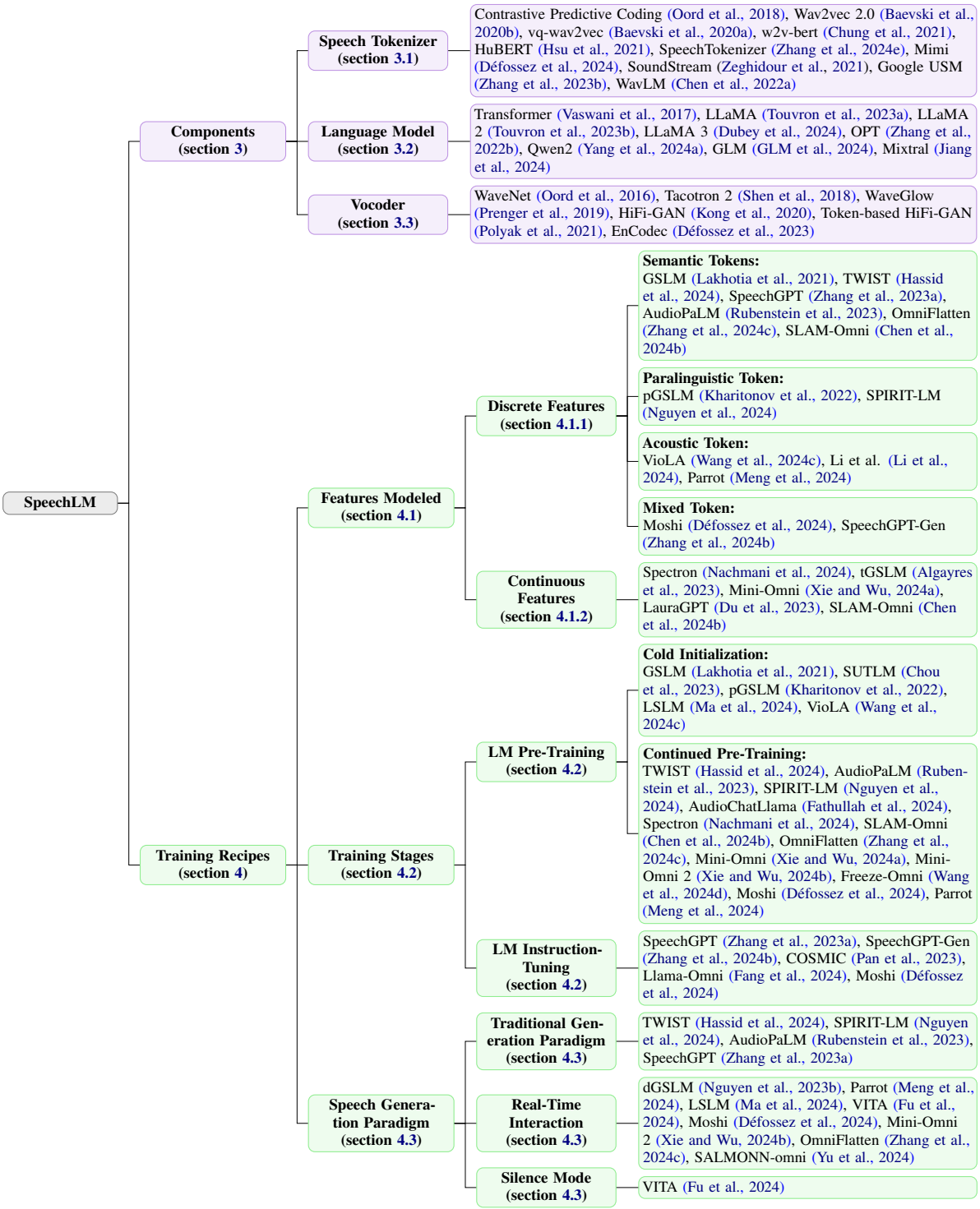


Figure 2: Taxonomy of Speech Language Models.