When Large Language Models Meet Speech: A Survey on Integration Approaches

Anonymous ACL submission

Abstract

Recent advancements in large language models (LLMs) have spurred interest in expanding their application beyond text-based tasks. A large number of studies have explored integrating other modalities with LLMs, notably speech modality, which is naturally related to text. This paper surveys the integration of speech with LLMs, categorizing the methodologies into three primary approaches: text-based, latent-representation-based, and audio-token-based integration. We also demonstrate how these methods are applied across various speech-related applications and highlight the challenges in this field to offer inspiration for future research.

1 Introduction

In recent years, the field of natural language processing (NLP) has been greatly reshaped by the development of large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023a; Google, 2024; Bai et al., 2023; DeepSeek-AI, 2024). These models have not only shown excellent ability in understanding and generating text but have also sparked interest in their potential applicability across other modalities, including speech. The integration of speech and LLMs offers a wide range of potential applications, including speech translation, conversational chatbots, and enhanced human-computer interaction in robotics.

Survey papers have reviewed speech language models (Peng et al., 2024; Cui et al., 2025), as well as audio language models (Latif et al., 2023) and multimodal language models (Ghosh et al., 2024a; Zhang et al., 2024b). However, there still lacks a survey specifically on the integration approaches of speech and LLMs, posing a challenge for researchers seeking to address this complex problem.

Distinct from other survey papers regarding speech and LLMs, this paper provides insight into the problem by specifically surveying the integration approaches of speech and LLMs. Studies on LLM tokenization (Chai et al., 2024; Tao et al., 2024) suggests that tokenization methods can affect the performance of LLMs. On the other hand, studies on tokenization methods for speech language modeling (Gat et al., 2023; Borsos et al., 2023) also show that speech tokenization methods can affect the performance of speech language models. In contrast to text processing, speech–LLM integration approaches are not limited to discrete tokenization, as presented in this paper. Therefore, studying the integration between speech and LLMs can be a key to innovations.

In this paper, we systematically categorize a substantial body of research on speech–LLM integration,[1] and provide a clear taxonomy of the integra-

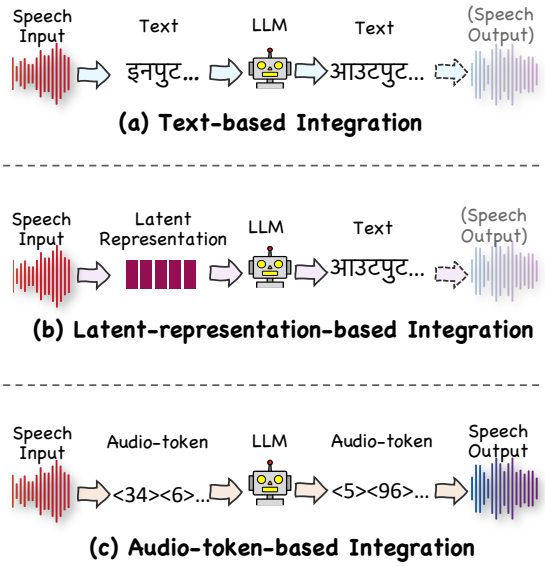

Figure 1: Overview of three fundamental approaches for speech-LLM integration.

[1] One challenge that affects the scope of studies is the lack of standard definition for LLMs. In this paper, we adopt the loose definition by Zhao et al. (2023), focusing on models with over 10 billion parameters, while also including notable

tion approaches. We broadly categorize the integration into the following three types: (a) **Text-based integration**: LLMs process textual data, integrated with speech-to-text and/or text-to-speech models; (b) **Latent-representation-based integration**: Latent vector representations that encode speech data are utilized, mainly as inputs to LLMs; (c) **Audio-token-based integration**: Speech tokens, such as semantic tokens and/or acoustic tokens, are used as the inputs/outputs for LLMs. The overview of these approaches is illustrated in Figure 1, and the detailed taxonomy with representative studies is presented in Figure 2.

2 Background

2.1 Language Modeling

Language modeling dates back to statistical models like N-gram models (Brown et al., 1992), which were central to early NLP and automatic speech recognition (ASR) systems (Bahl et al., 1983). Neural-network-based language modeling was introduced by Bengio et al. (2000), formulating the probability

$$P(w_{1:T}) = \prod_{t=1}^{T} P(w_t|w_{1:t-1}),$$

where w_t is the t-th token (text subwords or speech tokens) and $w_{i:j} = (w_i, w_{i+1}, \cdots, w_j)$ is the subsequence from i-th token to j-th token.

Advances in hardware enabled neural models to scale, leading to innovations such as sequence-to-sequence learning (Sutskever et al., 2014) and the attention mechanism (Bahdanau et al., 2015). The Transformer architecture (Vaswani et al., 2017) revolutionalizes language modeling using self-attention by efficiently modeling long-range dependencies utilizing parallelized computation. Decoder-only Transformer (Liu et al., 2018) gains prominence through the multi-task learning paradigm with generative pre-training and discrinative fine-tuning (Radford et al., 2018). Scaling these models (Radford et al., 2019; Brown et al., 2020) shows generalization to multiple tasks without explicit supervision and achieves comparative performance against task-specific models, which leads to the advent of LLM era. Techniques such as instruction tuning (Wei et al., 2022) and alignment to human preference via reinforcement learning (Ouyang et al., 2022) further advance LLM capabilities. For a comprehensive overview of LLMs, we refer the readers to LLM survey papers (Zhao et al., 2023; Minaee et al., 2024).

studies with smaller models.

2.2 Speech Representations

Speech is captured as waveform signals, sampled at a specific sampling rate and quantization value, and represented as a sequence of amplitude values. For deep learning applications, they are often converted to speech representations with shorter sequence lengths, such as log Mel filterbanks, which we refer to as **filterbanks**. Following the success of the unsupervised pre-training approach in other fields, self-supervised speech models (S3Ms) (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022) were introduced. These models predict masked frames to learn representations that capture phonetic information (Choi et al., 2024), along with a wide range of speech characteristics, such as speaker identity, paralinguistic information, and word-level information (Pasad et al., 2024). We refer to these contextualized frame-level features as **latent representations**.

To integrate speech (which is naturally continuous) with LLMs (which are designed to handle discrete tokens), **audio tokens** that are more discrete than previous representations have been studied recently. There are two mainstream approaches to modeling speech as audio tokens:[2] **semantic tokens** and **acoustic tokens**.[3] Semantic tokens can be obtained from S3Ms by discretizing latent representations using k-means. Acoustic tokens have been studied in the line of audio codecs (e.g., MP3 (ISO, 1993), Opus (Valin et al., 2012), etc.), which aim to compress audio signals. Neural audio codecs (Zeghidour et al., 2021; Défossez et al., 2023; Kumar et al., 2023) have recently gained traction as powerful tools to improve coding efficiency as well as perceptual quality. Because acoustic tokens exhibit better quality in producing output speech signals (Borsos et al., 2023) but fail to capture semantic information like semantic tokens do, recent studies combine both to obtain better representations for speech-language modeling (Wang et al., 2023d; Zhang et al., 2024c; Défossez et al., 2024).

[2] Different studies use different terms for these tokens (e.g., Lakhotia et al. (2021) used the term "units" for semantic tokens). We use the terms semantic tokens and acoustic tokens as in Borsos et al. (2023) and the term audio tokens to represent either or both of them similar to Défossez et al. (2024).

[3] We note that there are other efforts than semantic or acoustic tokens to discretize speech signals (Bai et al., 2024a).

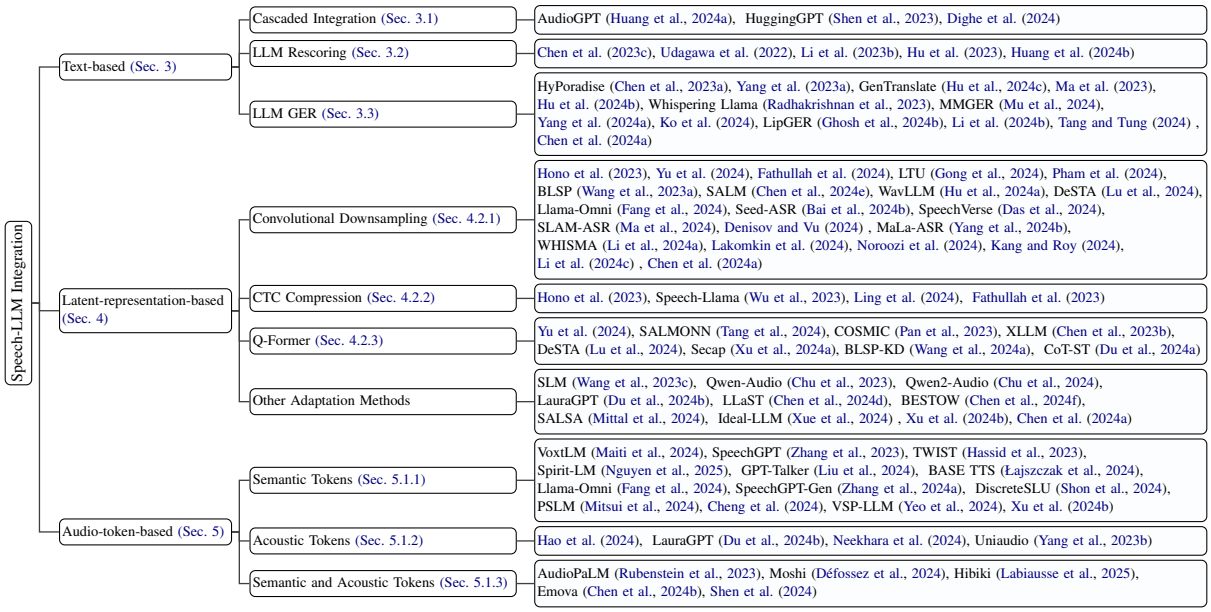

Figure 2: Taxonomy for speech-LLM integration. "Latent-representation-based" integration is categorized by different modality adaptation strategies. "Other Adaptation Methods" also includes studies that do not explicitly mention their modality adaptation methods.

3 Text-based Integration

This approach primarily utilizes text as both input and output for LLMs. One of the most direct implementations of this method is through **cascaded integration**, which is highly adaptable to various tasks. Within tasks like ASR, LLMs are not limited to working with the final recognition results but can also operate on intermediate hypotheses. This allows for sophisticated operations like **LLM rescoring**, and **LLM GER** (Generative Error Correction).

3.1 Cascaded Integration

Cascaded integration is the most simple and straightforward approach for integrating speech with LLMs. This involves the backbone LLM being supported by ASR and Text-to-Speech (TTS) interfaces for handling speech inputs and outputs, as well as the LLM invoking external models to solve speech-related tasks. This approach has been employed in models such as AudioGPT (Huang et al., 2024a) and HuggingGPT (Shen et al., 2023) to expand the applications of LLMs. The primary advantage of cascaded integration is its ease of implementation, allowing various tasks to be integrated into a single model with minimal effort. However, this method suffers from accuracy problems due to error propagation and efficiency problems due to the latency of processing through multiple steps.

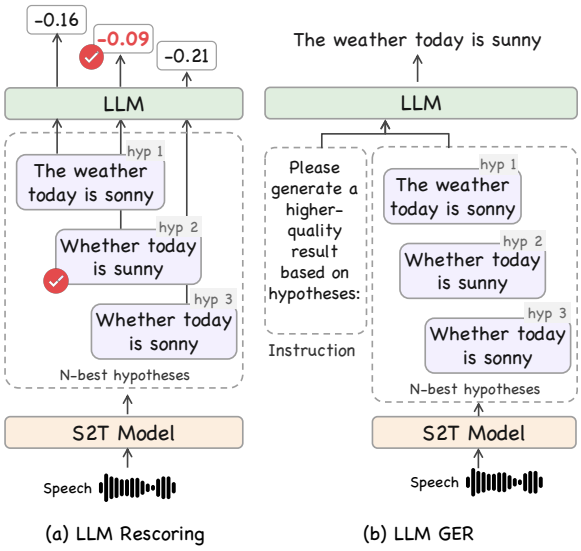

Figure 3: Two text-based approaches using hypotheses for improving speech-to-text tasks.

3.2 LLM Rescoring

Before research on LLMs explosively expanded, ASR models had already begun using rescoring mechanisms with external language models to improve transcription accuracy (McDermott et al., 2019; Shin et al., 2019; Salazar et al., 2020; Xu et al., 2022). Recently, there have been efforts towards utilizing LLMs for rescoring, and it has been proven to be effective (Chen et al., 2023c; Udagawa et al., 2022).

As shown in Figure 3 (a), this process involves generating a list of n-best hypotheses $\mathcal{H} = \{Y_1, Y_2, ..., Y_n\}$ from the speech input X through the initial ASR decoding, which is then re-evaluated using an LLM to find the hypothesis with higher linguistic coherence to improve accuracy. The rescoring process can be expressed as

$$Y^* = \text{argmax}_{Y_i \in \mathcal{H}} \left[(1-\lambda) \log p_{\text{AM}}(Y_i|X) + \lambda \log p_{\text{LLM}}(Y_i) \right]$$

where p_{AM} and p_{LLM} is the probability according to the ASR model (Acoustic Modle, AM) and the LLM, and λ is a weight balancing the contributions of two models.

3.3 LLM Generative Error Correction

Based on the concept of rescoring, LLM GER represents a new approach that fully utilizes the generative capability of LLMs. As shown in Figure 3 (b), the method utilizes an instructional prompt together with the n-best hypotheses list to guide the LLM in generating transcription predictions. Chen et al. (2023a) described this task performed by the LLM as hypotheses-to-transcription (H2T), which they explore under both fine-tuning and few-shot settings. Further studies have explored this direction with PEFT (Hu et al., 2024b) or in-context learning (Yang et al., 2023a; Ma et al., 2023).

Instead of merely selecting the best hypothesis from a pre-existing set, this method allows the LLM to generate a new transcription based on the hypotheses. By doing so, LLMs can potentially produce a result with a quality better than all initial hypotheses. Such an approach has proven to be well-suited for other tasks such as speech-to-text translation (S2TT), which require more generative capabilities of the model (Hu et al., 2024c). Lin et al. (2024) adopted the mixture-of-experts architecture (Fedus et al., 2022; Jiang et al., 2024) to let different experts handle different types of generative errors produced by task-specific models including ASR and ST models.

4 Latent-representation-based Integration

In this integration approach, a **speech encoder** is used to process the speech input and generate latent representations that are directly fed into the LLM, bypassing the embedding layer. The speech encoder is usually a Transformer-based model, which can be pre-trained on large-scale speech data or trained from scratch for the speech-LLM integration.

A critical issue of this approach is the sequence length gap between the speech and text modalities. Speech features, often sampled at rates of 50 to 100 frames per second, result in longer sequences compared to text tokens. Consequently, some form of **modality adaptation** mechanism is necessary to bridge these modalities, ensuring that the latent representations align with the LLM's embedding space.

4.1 Speech Encoder

For the speech encoder, various pre-trained models can be utilized to provide representations learned from large-scale speech data. This includes S3Ms (e.g. HuBERT (Hsu et al., 2021)) as well as the encoder of pre-trained ASR models (e.g. Whisper (Radford et al., 2023)). To connect the pre-trained speech encoder and the LLM, an *adapter* (also referred to as a *bridge network* or a *module connector*) is usually adopted.

An alternative approach is to train a speech encoder from scratch, specifically for the LLM integration. The structure of the speech encoder is usually a multi-layer Transformer (Vaswani et al., 2017) encoder or Conformer (Gulati et al., 2020).

4.2 Modality Adaptation

The modality adaptation mechanism is adopted for the speech encoder to map original frame-wise representations to token-wise representations. This allows the LLM to process the speech sequence in a manner similar to how it processes the text sequence, without having to deal with overly long sequences caused by long-form speech inputs.

The adaptation is usually accomplished by the adapter between pre-trained models, which is introduced in 4.1, or any part of the self-designed trained-from-scratch speech encoder. To this end, various adaptation methods have been proposed. Apart from some rarely-used strategies such as random downsampling (Wang et al., 2023c), most of the methods can be categorized into three groups, as shown in Figure 4: convolutional downsampling, CTC compression, and Q-former. According to comparisons conducted by Hono et al. (2023) and Yu et al. (2024), the Q-former generally outperforms convolutional downsampling, which in turn outperforms CTC compression. Next, we will detail these strategies in the following sections.

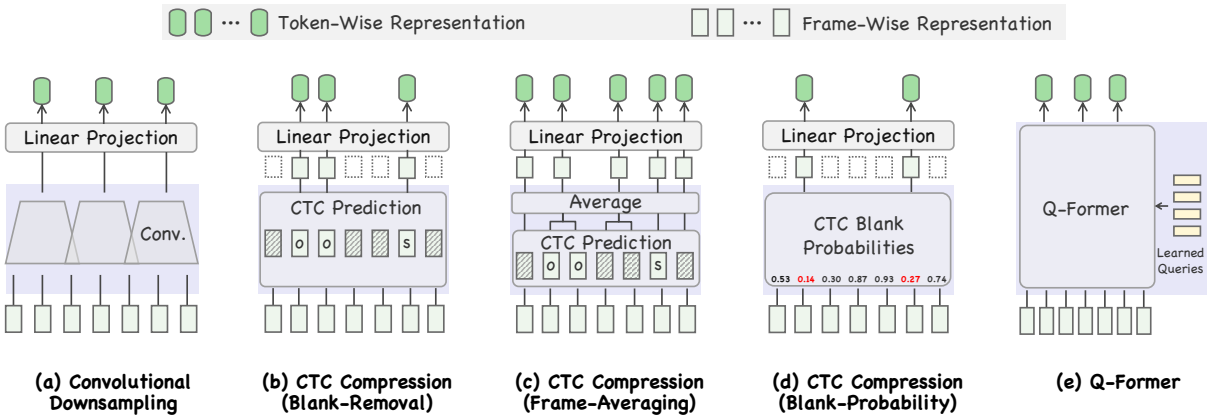

Figure 4: Different strategies for modality adaptation.

4.2.1 Convolutional Downsampling

A basic strategy where the speech representation sequence is downsampled from its original length using convolutional layers (Hono et al., 2023). When using the same number for kernel size and stride, it becomes equivalent to simply stacking multiple representation vectors together, which may only benefit parameter efficiency (Fathullah et al., 2024). For better alignment with the LLM's embeddings, some variants add a fully connected layer or a multi-head Transformer network after the convolutional layers (Yu et al., 2024).

4.2.2 CTC Compression

CTC compression involves two steps: (i) Training the speech encoder, or a part of it, on the ASR task based on Connectionist Temporal Classification (CTC) (Graves et al., 2013); (ii) Compressing the representation sequence based on the results of CTC predictions. Several specific strategies for the step (ii) include:

Blank-removal Discarding all the frames predicted as blanks (Hono et al., 2023; Wu et al., 2023).

Frame-averaging Averaging the latent representations of consecutive frames whose CTC predictions are the same label (Hono et al., 2023; Wu et al., 2023).

Blank-probability Discarding any frames with a CTC probability for blank exceeding a threshold (Ling et al., 2024).

4.2.3 Q-Former

Many studies (Yu et al., 2024; Tang et al., 2024; Pan et al., 2023) adopt Q-Former for the modality adaptation. Q-Former is a Transformer-based module converting variable-length input sequences into fixed-length output query representations. It is initially proposed for vision–text modality alignment (Li et al., 2023a).

4.3 Training Strategy

The optimal approach is to train the whole model consisting of the speech encoder and the LLM. However, fully fine-tuning the LLM is computationally expensive. Consequently, PEFT methods such as LoRA (Xu et al., 2024c), are often adopted to reduce resource consumption.

In scenarios where a pre-trained speech encoder is utilized, the primary focus is training the adapter connecting two pre-trained models. Whether to fine-tune the speech encoder or the LLM with PEFT seems optional as the decision differs across studies. On this topic, several studies (Hono et al., 2023; Pham et al., 2024) have conducted systematic investigations by comparing the effects of fine-tuning versus freezing different modules.

As for training the speech encoder from scratch, it is commonly done jointly with fine-tuning LLM with PEFT. Wu et al. (2023) proposed employing a two-stage training process where PEFT is not initiated until the speech encoder has undergone some initial training, ensuring a more stable training progression.

5 Audio-token-based Integration

As explained in Section 2.2, there are two types of audio tokens: semantic tokens and acoustic tokens. Studies use either or both, treating them as independent tokens like text or converting them back to latent representations. When semantic tokens are used for the output, a separate model to con-

vert them into filterbanks and to waveform signals similar to TTS models is often used.

5.1 Speech Language Models without LLMs

5.1.1 Semantic Token

Lakhotia et al. (2021) introduced the concept of generative spoken language modeling, which uses semantic tokens as both the input and output of a language model. Polyak et al. (2021) and Kharitonov et al. (2022) extended the idea by incorporating prosodic information. These models only used audio, and although they could output temporarily coherent contents, semantic understanding ability, and audio quality remained as challenges.

5.1.2 Acoustic Token

The use of acoustic tokens for language modeling was popularized in the task of TTS (Wang et al., 2023b; Chen et al., 2024c), where a few seconds of sample speech is used to generate coherent speech as those samples given text. Wang et al. (2023d) extended the idea to cover multi-tasking of ASR, machine translation (MT), and TTS.

5.1.3 Integration of Semantic and Acoustic tokens

AudioLM (Borsos et al., 2023) was introduced to incorporate both semantic and acoustic tokens. They adopted a two-stage approach, where the model first predicts semantic tokens and then subsequently predicts acoustic tokens. Zhang et al. (2024c) employed distillation from semantic representation to the first quantizer to combine semantic and acoustic tokens. Their generation approach is similar to Wang et al. (2023b), where an autoregressive model produces the first-layer token and a non-autoregressive model produces the rest.

5.2 Integration of LLMs into Speech Language Models

The use of LLMs as an underlying model of speech language models has become more popular following the success of LLMs. Hassid et al. (2023) showed that training speech language models using underlying pretrained text language models could improve performance. VoxtLM (Maiti et al., 2024) uses OPT (Zhang et al., 2022) as an underlying LLM and conducts multi-task fintuning it on ASR, TTS, and language modeling on text and semantic tokens. AudioPaLM (Rubenstein et al., 2023) adopts a similar approach to Borsos et al. (2023) but also enables text input/output, leveraging PaLM (Chowdhery et al., 2023) as the underlying LLM. TWIST and SpeechGPT (Zhang et al., 2023) use Llama (Touvron et al., 2023a) as the underlying model and use semantic tokens as both input and output. Spirit-LM (Nguyen et al., 2025) uses Llama 2 (Touvron et al., 2023b) as the underlying model and uses a mixture of semantic and text tokens for both input and output.

Défossez et al. (2024) introduced a new architecture for speech language modeling. Similarly to previous studies, they trained a text LLM and a codec model similar to Zhang et al. (2024c). Their architecture enables duplex ability of listening and generating speech by using a hierarchical autoregressive model consisting of two Transformer modules, namely Temporal Transformer and Depth Transformer (Lee et al., 2022). This architecture is also shown to be effective for speech-to-speech translation (S2ST) (Labiausse et al., 2025).

6 Comparative Analysis

6.1 Advantages and Disadvantages

It is important to understand the advantages and disadvantages of the three integration approaches when employing them for different applications. We summarize their pros and cons as follows:

- **Degree of integration:** Latent-representation-based > Audio-token-based > Text-based.
- **Interpretability:** Text-based > Audio-token-based > Latent-representation-based.
- **Speech generation ability:** Text-based and audio-token-based approach can generate speech, whereas latent-representation-based approaches typically cannot.

Based on these characteristics, different approaches excel in different scenarios:

- Latent-representation-based and audio-token-based approaches are better than text-based approach, in scenarios where sufficient resources (data, computational power, and time) are available or when real-time processing is required. Deeper integration reduces error propagation and latency but demands more extensive resources.
- Text-based approach is better than latent-representation-based and audio-token-based approaches, in scenarios where resources are limited or when greater interpretability is required.
- Latent-representation-based approach is better than audio-token-based, in scenarios where

Task	Dataset	Metrics	Model	Integration Method	Performance
ASR	**Librispeech** *dev-clean | dev-other | test-clean | test-other*	WER ↓	Whisper-large-v2 (Radford et al., 2023)	Non-LLM	— | — | 2.7 | 5.2
			HyPoradise (Chen et al., 2023a)	Text-based → LLM GER	— | — | 1.8 | 3.7
			BLSP (Wang et al., 2023a)	Latent-representation-based → Convolutional Downsampling	— | — | 10.4 | —
			SpeechVerse (Das et al., 2024)	Latent-representation-based → Convolutional Downsampling	— | — | 2.1 | 4.4
			Seed-ASR (Bai et al., 2024b)	Latent-representation-based → Convolutional Downsampling	— | — | **1.5** | **2.8**
			SALMONN (Tang et al., 2024)	Latent-representation-based → Q-Former	— | — | 2.1 | 4.9
			SLM (Wang et al., 2023c)	Latent-representation-based → Other Adaptation Methods	— | — | 2.6 | 5.0
			Qwen-Audio (Chu et al., 2023)	Latent-representation-based → Other Adaptation Methods	1.8 | 4.0 | 2.0 | 4.2
			Qwen2-Audio (Chu et al., 2024)	Latent-representation-based → Other Adaptation Methods	**1.3** | **3.4** | 1.6 | 3.6
			LauraGPT (Du et al., 2024b)	Latent-representation-based → Other Adaptation Methods	— | — | 4.4 | 7.7
			SpeechGPT-Gen (Zhang et al., 2024a)	Audio-token-based → Semantic Token	— | — | 2.4 | —
			SLAM-ASR (Ma et al., 2024)	Latent-representation-based → Convolutional Downsampling	— | — | 1.9 | 3.8
			BESTOW (Chen et al., 2024f)	Latent-representation-based → Other Adaptation Methods	— | — | — | 3.2
			Hao et al. (2024)	Audio-token-based → Acoustic Tokens	3.7 | 6.6 | 3.4 | 7.1
	Fleurs *zh | en*	WER ↓	Whisper-large-v2 (Radford et al., 2023)	Non-LLM	4.2 | 14.7
			Huang et al. (2024b) w/ PaLM2	Text-based → LLM Rescoring	13.1 | —
			Seed-ASR (Bai et al., 2024b)	Latent-representation-based → Convolutional Downsampling	**3.43** | —
			Qwen2-Audio (Chu et al., 2024)	Latent-representation-based → Other Adaptation Methods	— | **7.5**
			DiscreteSLU (Shon et al., 2024)	Audio-token-based → Semantic Token	12.6 | —
	AISHELL-2 *Mic | iOS | Android*	WER ↓	Seed-ASR (Bai et al., 2024b)	Latent-representation-based → Convolutional Downsampling	**2.2** | **2.2** | **2.2**
			Qwen-Audio (Chu et al., 2023)	Latent-representation-based → Other Adaptation Methods	3.3 | 3.1 | 3.3
			Qwen2-Audio (Chu et al., 2024)	Latent-representation-based → Other Adaptation Methods	3.0 | 3.0 | 2.9
			LauraGPT (Du et al., 2024b)	Latent-representation-based → Other Adaptation Methods	— | 3.2 | —
	VoxPopoli *All*	WER ↓	SpeechVerse (Das et al., 2024)	Latent-representation-based → Convolutional Downsampling	**6.5**
			AudioPaLM (Rubenstein et al., 2023)	Audio-token-based → Semantic and Acoustic Tokens	9.8
S2TT	**CoVoST2** *en-de | de-en | en-zh | zh-en | es-en | fr-en | it-en | ja-en*	BLEU ↑	Whisper-large-v2 (Radford et al., 2023)	Non-LLM	— | 36.3 | — | 18.0 | 40.1 | 36.4 | 30.9 | 26.1
			GenTranslate (Hu et al., 2024c)	Text-based → LLM GER	— | 39.2 | — | 21.6 | 42.0 | 41.7 | — | 22.9
			GenTranslate-V2 (Hu et al., 2024c)	Text-based → LLM GER	— | 40.6 | — | 23.3 | 43.6 | 42.7 | — | 25.4
			BLSP (Wang et al., 2023a)	Latent-representation-based → Convolutional Downsampling	24.4 | — | 41.3 | — | — | — | — | —
			Speech-Llama (Wu et al., 2023)	Latent-representation-based → CTC Compression	— | 27.1 | — | 12.3 | 27.9 | 25.2 | 25.9 | 19.9
			SALMONN (Tang et al., 2024)	Latent-representation-based → Q-Former	18.6 | — | 33.1 | — | — | — | — | —
			Qwen-Audio (Chu et al., 2023)	Latent-representation-based → Other Adaptation Methods	25.1 | 33.9 | 41.5 | 15.7 | 39.7 | 38.5 | 36.0 | —
			Qwen2-Audio (Chu et al., 2024)	Latent-representation-based → Other Adaptation Methods	**29.9** | 35.2 | **45.2** | 24.4 | 40.0 | 38.5 | 36.3 | —
			LauraGPT (Du et al., 2024b)	Latent-representation-based → Other Adaptation Methods	— | — | 38.5 | — | — | — | — | —
			LLaST (Chen et al., 2024d)	Latent-representation-based → Other Adaptation Methods	— | 41.2 | — | 24.8 | **46.1** | **45.1** | **43.0** | **28.8**
			AudioPaLM (Rubenstein et al., 2023)	Audio-token-based → Semantic and Acoustic Tokens	— | **43.4** | — | **25.5** | 44.2 | 44.8 | — | 25.9
			Ideal-LLM (Xue et al., 2024)	Latent-representation-based → Other Adaptation Methods	25.9 | 38.5 | — | — | 41.5 | 40.0 | 38.0 | —
S2ST	**CVSS S2ST** *de-en | zh-en | es-en | fr-en | it-en | ja-en*	ASR-BLEU ↑	Translatotron 2 + pretraining + TTS aug (Jia et al., 2022)	Non-LLM	33.6 | 13.1 | 38.5 | 36.5 | 35.7 | 8.5
			AudioPaLM (Rubenstein et al., 2023)	Audio-token-based → Semantic and Acoustic Tokens	**37.2** | **20.0** | **40.4** | **38.3** | **39.4** | **20.9**
TTS	**AISHELL-1**	CER ↓ | SECS ↑ | MOSNet ↑	VALL-E Phone (Wang et al., 2023b)	Non-LLM	**4.75** | **0.91** | **3.22**
			VALL-E Token (Wang et al., 2023b)	Non-LLM	6.52 | **0.91** | 3.19
			LauraGPT (Du et al., 2024b)	Audio-token-based → Acoustic Tokens	6.91 | 0.90 | 3.14
	LibriTTS	WER ↓ | SECS ↑ | MOSNet ↑	VALL-E Phone (Wang et al., 2023b)	Non-LLM	**4.30** | 0.92 | 3.28
			VALL-E Token (Wang et al., 2023b)	Non-LLM	6.57 | **0.93** | 3.28
			SpeechGPT-Gen (zero-shot) (Zhang et al., 2024a)	Audio-token-based → Semantic Token	3.10 | 0.63 | 3.63
			VoxtLM (Maiti et al., 2024)	Audio-token-based → Semantic Token	— | — | **4.36**
			LauraGPT (Du et al., 2024b)	Audio-token-based → Acoustic Tokens	8.62 | 0.91 | 3.26
	Topic-StoryCloze	TSC ↑	Spirit-LM (Nguyen et al., 2025)	Audio-token-based → Semantic Token	82.9
			TWIST-1.3B (Hassid et al., 2023)	Audio-token-based → Semantic Token	70.6
			TWIST-7B (Hassid et al., 2023)	Audio-token-based → Semantic Token	74.1
			TWIST-13B (Hassid et al., 2023)	Audio-token-based → Semantic Token	76.4
			Moshi (Défossez et al., 2024)	Audio-token-based → Semantic and Acoustic Tokens	**83.6**
	StoryCloze	SSC ↑	Spirit-LM (Nguyen et al., 2025)	Audio-token-based → Semantic Token	61.0
			TWIST-1.3B (Hassid et al., 2023)	Audio-token-based → Semantic Token	52.4
			TWIST-7B (Hassid et al., 2023)	Audio-token-based → Semantic Token	55.3
			TWIST-13B (Hassid et al., 2023)	Audio-token-based → Semantic Token	55.4
			Moshi (Défossez et al., 2024)	Audio-token-based → Semantic and Acoustic Tokens	**62.7**

Table 1: Quantitative comparison based on applications. The **Bold** denotes the best result.

speech is only considered as input. Latent-representation-based methods provide the deepest integration when generating textual or other downstream outputs.

- Audio-token-based is better than latent-representation-based, in scenarios where speech is also considered as output. Generating speech from latent representations remains challenging, making audio-token-based approaches more suitable in these cases.

6.2 Quantitative Comparison

We provide a comprehensive performance comparison in Table 1, comparing studies with different integration approaches across multiple tasks: ASR, S2TT, S2ST, TTS. The table highlights key metrics such as word error rate (WER), BLEU score, and others, alongside each model's integration method and backbone components.

While Table 1 summarizes many state-of-the-art results, direct comparisons can be challenging due to variations in backbone LLMs, acoustic front-ends, and training protocols. For instance, although BLEU scores on CoVoST 2 (Wang et al., 2021) indicate that deeper integration methods (e.g., latent-representation-based) often yield improved performance, model size and training resources also play a significant role. Future work should focus on more uniform experimental conditions to isolate the impact of each approach.

Nonetheless, the metrics in Table 1 serve as a starting point for evaluating how well each method handles different tasks. The results suggest that deeper integration (e.g., latent-representation-based) can be beneficial when sufficient computational resources and training data are available, whereas text-based methods often offer greater interpretability and simpler pipelines.

7 Challenges

7.1 Text-based Integration

Text-based integration preserves the original LLM input modality, which is text. Because the model is already optimized for textual inputs, this approach typically requires the least adaptation from the LLM. However, transforming speech into text before feeding it into the LLM inevitably introduces a layer of abstraction and potential information loss such as prosody and emotion, which limits the downstream performance.

Therefore, the main challenge is how to convey the rich information of source speech through text. In current practice, the intermediate text is often generated based on the highest probability (Chen et al., 2023a; Yang et al., 2023a; Radhakrishnan et al., 2023). Although this strategy maximizes intermediate accuracy, it may not be optimal for the LLM and its final outputs. One possible direction is to inject controlled randomness during decoding to ensure additional diversity. However, degrading accuracy could also have a negative effect, creating a trade-off. How to balance diversity and accuracy remains an open research question that requires future investigation.

7.2 Latent-representation-based Integration

In contrast to text-based integration, latent-representation-based integration employs an intermediate representation that is closer to the source speech but more distant from the LLM's natural input space. Consequently, the primary challenge lies in aligning these acoustic representations with the textual embedding space of the LLM. Section 4.2 introduces multiple modality adaptation mechanisms to address this and enable LLMs to handle speech more effectively.

While the primary focus of most research introduced in this paper has been on adapting LLMs to better handle speech data, there is comparatively less attention on how these adaptations affect performance on text-based tasks. The alignment of text and speech data, along with their latent representations, remains underexplored, with only a few studies focusing on this issue (Wang et al., 2023a; Lu et al., 2024).

A promising solution for improving speech-text alignment is the generation of high-quality synthetic speech data to supplement real datasets, thus closing the gap between the speech modality and text-based LLMs. More thorough investigations into how to align speech and text representations in complex languages (beyond English) could also strengthen the multimodal capabilities of LLMs.

7.3 Audio-token-based Integration

As stated in Sections 2.2 and 5, S3Ms (and their derived semantic tokens) mainly capture phonetic information (Choi et al., 2024), while acoustic tokens offer higher fidelity in generating speech signals (Borsos et al., 2023).

Integrating both representations, as presented in Section 5.1.3, is one approach that requires further investigation. While this approach achieves strong performance across various downstream tasks (Table 1), there are still limited studies on this approach, where the models are English-centric models that rely on millions of hours of speech data. Minimizing computational requirements is essential to extend this approach to languages lacking the vast resources available for English.

Some studies suggest that latent-representation-based integration can outperform audio-token-based integration (Wang et al., 2024b). However, it is possible that this is due to the suboptimal representation of audio tokens (Gat et al., 2023), requiring further improvement of tokenization methods and comparison with other integration approaches.

7.4 Fair Comparison Across Integration Approaches

Beyond these integration-approach-specific challenges, there is a notable gap in comparing the different integration approaches under a unified setting. Most existing works focus on one approach, making it difficult to assess their relative merits consistently. A fair comparison among these integration methods could clarify how different factors affect performance. Developing standardized benchmarks, protocols, and reporting practices for speech-LLM research would help future work isolate the core differences between these approaches.

8 Conclusion

In this survey, we systematically categorize the integration of LLMs with speech modality, outlining three primary approaches: text-based, latent-representation-based, and audio-token-based integration. Each method demonstrates unique strengths in performing different speech-related tasks. However, numerous challenges remain in this evolving field, presenting significant opportunities for future research.

Limitations

In our survey of the integration of speech with LLMs, we have covered the significant developments within this field. Despite our comprehensive approach, we acknowledge that the rapid pace of development in speech processing and LLM means that some recent advances or discussions might have escaped our attention and are not fully addressed within this paper.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Thirty-fourth Conference on Neural Information Processing Systems*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.

He Bai, Tatiana Likhomanenko, Ruixiang Zhang, Zijin Gu, Zakaria Aldeneh, and Navdeep Jaitly. 2024a. dmel: Speech tokenization made simple. *Preprint*, arXiv:2407.15835.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, et al. 2024b. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2523–2533.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yekun Chai, Yewei Fang, Qiwei Peng, and Xuhong Li. 2024. Tokenization falling short: On subword robustness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1582–1599, Miami, Florida, USA. Association for Computational Linguistics.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Chng Eng Siong. 2023a. Hyporadise: An open baseline for generative speech recognition with large language models. 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*.

Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, Ensiong Chng, and Chao-Han Huck Yang. 2024a. It's never too late: Fusing acoustic information into large language models for automatic speech recognition. *arXiv preprint arXiv:2402.05457*.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023b. X-LLM: bootstrapping advanced large language models by treating multi-modalities as foreign languages. *CoRR*, abs/2305.04160.

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

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu

Wei. 2024c. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *Preprint*, arXiv:2406.05370.

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

Tongzhou Chen, Cyril Allauzen, Yinghui Huang, Daniel S. Park, David Rybach, W. Ronny Huang, Rodrigo Cabrera, Kartik Audhkhasi, Bhuvana Ramabhadran, Pedro J. Moreno, and Michael Riley. 2023c. Large-scale language model rescoring on long-form data. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Xi Chen, Songyang Zhang, Qibing Bai, Kai Chen, and Satoshi Nakamura. 2024d. Llast: Improved end-to-end speech translation system leveraged by large language models. *arXiv preprint arXiv:2407.15415*.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C. Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024e. SALM: speech-augmented language model with in-context learning for speech recognition and translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 13521–13525. IEEE.

Zhehuai Chen, He Huang, Oleksii Hrinchuk, Krishna C Puvvada, Nithin Rao Koluguri, Piotr Żelasko, Jagadeesh Balam, and Boris Ginsburg. 2024f. Bestow: Efficient and streamable speech language model with the best of two worlds in gpt and t5. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 147–154. IEEE.

Yao-Fei Cheng, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, Wen Shen Teo, Siddhant Arora, and Shinji Watanabe. 2024. Task arithmetic for language expansion in speech translation. *arXiv preprint arXiv:2409.11274*.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-supervised speech representations are more phonetic than semantic. In *Interspeech 2024*, pages 4578–4582.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

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

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

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2025. Recent advances in speech language models: A survey. *Preprint*, arXiv:2410.03751.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. 2024. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

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

Pavel Denisov and Thang Vu. 2024. Teaching a multilingual large language model to understand multilingual speech via multi-instructional training. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 814–834. Association for Computational Linguistics.

Pranay Dighe, Yi Su, Shangshang Zheng, Yunshu Liu, Vineet Garg, Xiaochuan Niu, and Ahmed H. Tewfik. 2024. Leveraging large language models for exploiting ASR uncertainty. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12231–12235. IEEE.

Yexing Du, Ziyang Ma, Yifan Yang, Keqi Deng, Xie Chen, Bo Yang, Yang Xiang, Ming Liu, and Bing Qin. 2024a. Cot-st: Enhancing llm-based speech translation with multimodal chain-of-thought. *arXiv preprint arXiv:2409.19510*.

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. 2024b. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *Preprint*, arXiv:2310.04673.

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. *Preprint*, arXiv:2410.00037.

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, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2024. Prompting large language models with speech recognition abilities. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 13351–13355. IEEE.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2023. Towards general-purpose speech abilities for large language models using unpaired data. *arXiv preprint arXiv:2311.06753*.

William Fedus, Jeff Dean, and Barret Zoph. 2022. A review of sparse expert models in deep learning. *Preprint*, arXiv:2209.01667.

Itai Gat, Felix Kreuk, Tu Anh Nguyen, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, and Yossi Adi. 2023. Augmentation invariant discrete representation for generative spoken language modeling. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 465–477, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024a. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *CoRR*, abs/2404.07214.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Purva Chiniya, Utkarsh Tyagi, Ramani Duraiswami, and Dinesh Manocha. 2024b. Lipger: Visually-conditioned generative error correction for robust automatic speech recognition. *arXiv preprint arXiv:2406.04432*.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Gemini Team Google. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 6645–6649. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 5036–5040. ISCA.

Hongkun Hao, Long Zhou, Shujie Liu, Jinyu Li, Shujie Hu, Rui Wang, and Furu Wei. 2024. Boosting large language model for speech synthesis: An empirical study. *CoRR*, abs/2401.00246.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pretrained speech language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yukiya Hono, Koh Mitsuda, Tianyu Zhao, Kentaro Mitsui, Toshiaki Wakatsuki, and Kei Sawada. 2023. An integration of pre-trained speech and language models for end-to-end speech recognition. *CoRR*, abs/2312.03668.

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 Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Ke Hu, Tara N. Sainath, Bo Li, Nan Du, Yanping Huang, Andrew M. Dai, Yu Zhang, Rodrigo Cabrera, Zhifeng Chen, and Trevor Strohman. 2023. Massively multilingual shallow fusion with large language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024a. Wavllm: Towards robust and adaptive speech large language model. *CoRR*, abs/2404.00656.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and Engsiong Chng. 2024b. Large language models are efficient learners of noise-robust speech recognition. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and EngSiong Chng. 2024c. Gentranslate: Large language models are generative multilingual speech and machine translators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 74–90. Association for Computational Linguistics.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024a. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 23802–23804. AAAI Press.

W. Ronny Huang, Cyril Allauzen, Tongzhou Chen, Kilol Gupta, Ke Hu, James Qin, Yu Zhang, Yongqiang Wang, Shuo-Yiin Chang, and Tara N. Sainath. 2024b. Multilingual and fully non-autoregressive ASR with large language model fusion: A comprehensive study. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 13306–13310. IEEE.

ISO. 1993. ISO/IEC 11172-3:1993 information technology — coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s — part 3: Audio.

Ye Jia, Yifan Ding, Ankur Bapna, Colin Cherry, Yu Zhang, Alexis Conneau, and Nobu Morioka. 2022. Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation. In *Interspeech 2022*, pages 1721–1725.

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.

Wonjune Kang and Deb Roy. 2024. Prompting large language models with audio for general-purpose speech summarization. *arXiv preprint arXiv:2406.05968*.

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.

Yuka Ko, Sheng Li, Chao-Han Huck Yang, and Tatsuya Kawahara. 2024. Benchmarking japanese speech recognition on asr-llm setups with multi-pass augmented generative error correction. *arXiv preprint arXiv:2408.16180*.

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*.

Tom Labiausse, Laurent Mazaré, Edouard Grave, Patrick Pérez, Alexandre Défossez, and Neil Zeghidour. 2025. High-fidelity simultaneous speech-to-speech translation. *Preprint*, arXiv:2502.03382.

Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, et al. 2024. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*.

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

Egor Lakomkin, Chunyang Wu, Yassir Fathullah, Ozlem Kalinli, Michael L Seltzer, and Christian Fuegen. 2024. End-to-end speech recognition contextualization with large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12406–12410. IEEE.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W. Schuller. 2023. Sparks of large audio models: A survey and outlook. *CoRR*, abs/2308.12792.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11513–11522.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Mohan Li, Cong-Thanh Do, Simon Keizer, Youmna Farag, Svetlana Stoyanchev, and Rama Doddipatla. 2024a. Whisma: A speech-llm to perform zero-shot spoken language understanding. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1115–1122. IEEE.

Sheng Li, Chen Chen, Chin Yuen Kwok, Chenhui Chu, Eng Siong Chng, and Hisashi Kawai. 2024b. Investigating asr error correction with large language model and multilingual 1-best hypotheses. In *Proc. Interspeech*, pages 1315–1319.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023b. Prompting large language models for zero-shot domain adaptation in speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Yuang Li, Jiawei Yu, Min Zhang, Mengxin Ren, Yanqing Zhao, Xiaofeng Zhao, Shimin Tao, Jinsong Su, and Hao Yang. 2024c. Using large language model for end-to-end chinese asr and ner. *arXiv preprint arXiv:2401.11382*.

Yen-Ting Lin, Chao-Han Huck Yang, Zhehuai Chen, Piotr Zelasko, Xuesong Yang, Zih-Ching Chen, Krishna C Puvvada, Szu-Wei Fu, Ke Hu, Jun Wei Chiu, Jagadeesh Balam, Boris Ginsburg, and Yu-Chiang Frank Wang. 2024. Neko: Toward post recognition generative correction large language models with task-oriented experts. *Preprint*, arXiv:2411.05945.

Shaoshi Ling, Yuxuan Hu, Shuangbei Qian, Guoli Ye, Yao Qian, Yifan Gong, Ed Lin, and Michael Zeng. 2024. Adapting large language model with speech for fully formatted end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11046–11050. IEEE.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. 2024. Generative expressive conversational speech synthesis. *CoRR*, abs/2407.21491.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024. Desta: Enhancing speech language models through descriptive speech-text alignment. *CoRR*, abs/2406.18871.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark J. F. Gales, and Kate M. Knill. 2023. Can generative large language models perform ASR error correction? *CoRR*, abs/2307.04172.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.

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.

Erik McDermott, Hasim Sak, and Ehsan Variani. 2019. A density ratio approach to language model fusion in end-to-end automatic speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 434–441. IEEE.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *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. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 2692–2700. Association for Computational Linguistics.

Ashish Mittal, Darshan Prabhu, Sunita Sarawagi, and Preethi Jyothi. 2024. Salsa: Speedy asr-llm synchronous aggregation. *arXiv preprint arXiv:2408.16542*.

Bingshen Mu, Yangze Li, Qijie Shao, Kun Wei, Xucheng Wan, Naijun Zheng, Huan Zhou, and Lei Xie. 2024. Mmger: Multi-modal and multi-granularity generative error correction with llm for joint accent and speech recognition. *arXiv preprint arXiv:2405.03152*.

Paarth Neekhara, Shehzeen Hussain, Subhankar Ghosh, Jason Li, Rafael Valle, Rohan Badlani, and Boris Ginsburg. 2024. Improving robustness of llm-based speech synthesis by learning monotonic alignment. *arXiv preprint arXiv:2406.17957*.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, et al. 2025. Spirit-lm: Interleaved spoken and written language model.

Transactions of the Association for Computational Linguistics, 13:30–52.

Vahid Noroozi, Zhehuai Chen, Somshubra Majumdar, Steve Huang, Jagadeesh Balam, and Boris Ginsburg. 2024. Instruction data generation and unsupervised adaptation for speech language models. *arXiv preprint arXiv:2406.12946*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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. *CoRR*, abs/2311.02248.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What Do Self-Supervised Speech Models Know About Words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Jing Peng, Yucheng Wang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2024. A survey on speech large language models. *Preprint*, arXiv:2410.18908.

Van Tung Pham, Yist Y. Lin, Tao Han, Wei Li, Jun Zhang, Lu Lu, and Yuxuan Wang. 2024. A comprehensive solution to connect speech encoder and large language model for ASR. *CoRR*, abs/2406.17272.

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 *Interspeech 2021*, pages 3615–3619.

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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegner. 2023. Whispering llama: A cross-modal generative error correction framework for speech recognition. *arXiv preprint arXiv:2310.06434*.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2699–2712. Association for Computational Linguistics.

Maohao Shen, Shun Zhang, Jilong Wu, Zhiping Xiu, Ehab AlBadawy, Yiting Lu, Mike Seltzer, and Qing He. 2024. Get large language models ready to speak: A late-fusion approach for speech generation. *arXiv preprint arXiv:2410.20336*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580.

Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using BERT for speech recognition. In *Proceedings of The 11th Asian Conference on Machine Learning, ACML 2019, 17-19 November 2019, Nagoya, Japan*, volume 101 of *Proceedings of Machine Learning Research*, pages 1081–1093. PMLR.

Suwon Shon, Kwangyoun Kim, Yi-Te Hsu, Prashant Sridhar, Shinji Watanabe, and Karen Livescu. 2024. Discreteslu: A large language model with self-supervised discrete speech units for spoken language understanding. *arXiv preprint arXiv:2406.09345*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yixuan Tang and Anthony KH Tung. 2024. Contextualized speech recognition: rethinking second-pass rescoring with generative large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6478–6485.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Takuma Udagawa, Masayuki Suzuki, Gakuto Kurata, Nobuyasu Itoh, and George Saon. 2022. Effect and analysis of large-scale language model rescoring on competitive ASR systems. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 3919–3923. ISCA.

Jean-Marc Valin, Koen Vos, and Timothy B. Terriberry. 2012. Definition of the Opus Audio Codec. RFC 6716. Internet Engineering Task Force (IETF).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023a. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, and Jiajun Zhang. 2024a. Blsp-kd: Bootstrapping language-speech pre-training via knowledge distillation. *arXiv preprint arXiv:2405.19041*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023b. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Dingdong Wang, Mingyu Cui, Dongchao Yang, Xueyuan Chen, and Helen Meng. 2024b. A comparative study of discrete speech tokens for semantic-related tasks with large language models. *Preprint*, arXiv:2411.08742.

Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K. Rubenstein, Lukas Zilka, Dian Yu, Golan Pundak, Nikhil Siddhartha, Johan Schalkwyk, and Yonghui Wu. 2023c. SLM: bridge the thin gap between speech and text foundation models. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023d. Viola: Unified codec language models for speech recognition, synthesis, and translation. *Preprint*, arXiv:2305.16107.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and Yu Wu. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. 2022. Rescorebert: Discriminative speech recognition rescoring with bert. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6117–6121. IEEE.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024a. Secap: speech emotion captioning with large language model. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Yaoxun Xu, Shi-Xiong Zhang, Jianwei Yu, Zhiyong Wu, and Dong Yu. 2024b. Comparing discrete and continuous space llms for speech recognition. *arXiv preprint arXiv:2409.00800*.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. 2024c. Qa-lora: Quantization-aware low-rank adaptation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hongfei Xue, Wei Ren, Xuelong Geng, Kun Wei, Longhao Li, Qijie Shao, Linju Yang, Kai Diao, and Lei Xie. 2024. Ideal-llm: Integrating dual encoders and language-adapted llm for multilingual speech-to-text. *arXiv preprint arXiv:2409.11214*.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023a. Generative speech recognition error correction with large language models and task-activating prompting. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Chao-Han Huck Yang, Taejin Park, Yuan Gong, Yuanchao Li, Zhehuai Chen, Yen-Ting Lin, Chen Chen, Yuchen Hu, Kunal Dhawan, Piotr Żelasko, et al. 2024a. Large language model based generative error correction: A challenge and baselines forspeech recognition, speaker tagging, and emotion recognition. *arXiv preprint arXiv:2409.09785*.

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

Guanrou Yang, Ziyang Ma, Fan Yu, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024b. Mala-asr: Multimedia-assisted llm-based asr. *arXiv preprint arXiv:2406.05839*.

Jeong Hun Yeo, Seunghee Han, Minsu Kim, and Yong Man Ro. 2024. Where visual speech meets language: VSP-LLM framework for efficient and context-aware visual speech processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 11391–11406. Association for Computational Linguistics.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting speech encoder and large language model for ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12637–12641. IEEE.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *Preprint*, arXiv:2107.03312.

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

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

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024b. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 12401–12430. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024c. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.