Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260384392

Integrating Speech into Large Language Models: Architectures, Training Strategies, and Emerging Challenges

Jiajun Li*

College of Artificial Intelligence, Tianjin University, Tianjin, China

**Corresponding author: Jiajun Li.*

Abstract

This paper provides a thorough examination of methods for incorporating speech into large language models (LLMs), with particular emphasis on architectural frameworks, training methodologies, and evaluation protocols. The analysis assesses three critical dimensions: the performance of speech encoders on tasks such as speech recognition, translation, dialogue, and affective computing, cross-modal alignment training techniques, and approaches for integrating speech encoders with LLMS. A total of 18 studies published between 2023 and 2024 were included in the survey. The unified decoder framework, the encoder-adapter LLM pipeline, and the multi-stream hierarchical model are the three architectural approaches identified. Each methodology demonstrates unique trade-offs between modularity and integration depth. Our findings indicate that dual-encoder architectures and hierarchical token representations significantly improve model robustness. Additionally, catastrophic forgetting is effectively mitigated in cross-modal training through curriculum learning and activation tuning. The computational efficiency, uniformity of evaluation, and scaling performance of spoken language models are persistently challenged, in contrast to text-based models. To further investigate real-time full-duplex communication, systematic scaling techniques for speech foundation models, and low-resource language documentation, additional research is needed.

Keywords

speech language models, multimodal large language models, cross-modal alignment, speech-text integration, foundation models

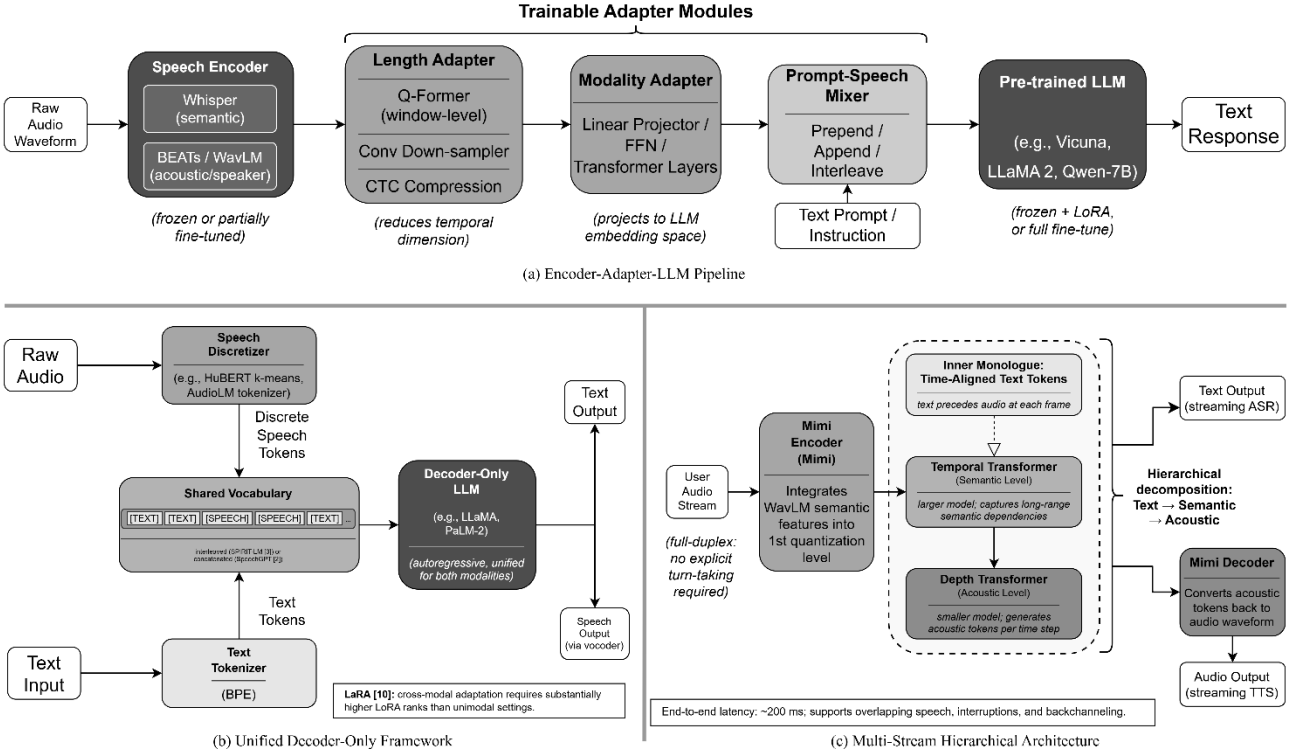
1. Introduction

Large language models (LLMs) have revolutionized natural language processing. The capabilities of models such as GPT-4 and LLaMA in text comprehension, generation, and contextual reasoning enable the development of foundational models that integrate multiple modalities into unified architectures [1]. Simultaneously, speech processing has made significant progress, as models such as HuBERT, wav2vec 2.0, and Whisper generate high-level representations directly from unprocessed audio [1, 2]. Speech is a multidimensional medium that encompasses paralinguistic elements, including sentiment, speaker identity, and prosodic characteristics, in addition to linguistic content.

Despite this advancement, technical obstacles remain. Speech signals are hierarchically structured across multiple levels of abstraction, are continuous, and temporally protracted, whereas LLMs are designed for textual input and operate on discrete token sequences [2]. To address this representational disparity, it is essential to implement architectural components, such as adapters, length compressors, and cross-modal projectors, as well as training methodologies that maintain the LLM's linguistic proficiency [3]. The absence of standardized evaluation methodologies further complicates cross-system comparisons and obscures optimal design decisions [1]. Recent scaling experiments have demonstrated that speech LMs transmit linguistic information at a significantly slower rate than text LLMs, suggesting that the bottleneck lies in integration rather than in fundamental design [4].

This study evaluates these concerns by conducting a narrative analysis of 18 peer-reviewed and preprint articles published between 2023 and 2024. The articles were procured from ACL, EMNLP, NeurIPS, ICLR, and ICASSP. Studies were included if they presented innovative designs or training methodologies for speech-LLM integration, conducted systematic evaluations of speech tasks, or provided analytical insights into scaling behavior or cross-modal alignment. The investigation is organized around three primary themes: downstream evaluation, architectural design, and training methodology. We investigate the divergent and complementary design characteristics of representative systems, such as the AudioPaLM [6] and the Qwen-Audio series [7]. Table 1 illustrates that the three paradigms offer distinct trade-offs suited to different deployment scenarios.

Figure 1: Three Architectural Models for the Integration of Voice and Llm: (A) Encoder-adapter-llm Pipeline, (B) Unified Decoder-only Framework, and (C) Multi-stream Hierarchical Design.



2. Architectural Paradigms for Speech-LLM Integration

The analyzed designs are characterized by three paradigms, each embodying a unique hypothesis about speech-text interaction within a unified model. The paradigm, encoder configuration, LLM backbone, and essential design innovation of representative systems are summarized in Table 1.

Table 1: Comparative Analysis of Representative Speech–llm Models

Model	Paradigm	Speech Encoder(s)	LLM Backbone	Key Innovation
SALMONN [3]	Encoder-Adapter-LLM	Whisper + BEATs	Vicuna-13B	Dual-encoder; window-level Q-Former; activation tuning

Model	Paradigm	Speech Encoder(s)	LLM Backbone	Key Innovation
WavLLM [5]	Encoder-Adapter-LLM	Whisper + WavLM	LLaMA 2-7B	Prompt-aware LoRA; curriculum learning
Gaido et al. [1]	Encoder-Adapter-LLM	Various SFMs	Various LLMs	Systematic taxonomy of five building blocks
Qwen-Audio [7]	Encoder-Adapter-LLM	Whisper-large-v2	Qwen-7B	Multi-task framework; 30+ audio tasks; hierarchical tags
SpeechGPT [2]	Unified Decoder-Only	HuBERT (discrete)	LLaMA-13B	Chain-of-modality instruction tuning
VoxtLM [8]	Unified Decoder-Only	HuBERT (discrete)	Pre-trained text LM	Unified Voxt vocabulary; four tasks
AudioPaLM [6]	Unified Decoder-Only	AudioLM tokens	PaLM-2 (8B)	Joint text–audio vocabulary; zero-shot S2ST
SPIRIT-LM [9]	Unified Decoder-Only	HuBERT (discrete)	LLaMA 2-7B	Interleaved speech–text pre-training; word-level alignment
LaRA [10]	Unified Decoder-Only	—	Various	High-rank adaptation needed for cross-modal transfer
Moshi [11]	Multi-Stream Hierarchical	Mimi neural codec (distilled from WavLM)	Helium-7B	Full-duplex dialogue; Inner Monologue; ~200ms latency
GPST [12]	Multi-Stream Hierarchical	Hierarchical codebook	—	Global–local Transformer; 24 kHz synthesis; approximately one-third the parameters of AudioLM

2.1 The Encoder-adapter-LLM Pipeline

Through one or more adapter modules, this framework establishes a connection between a pre-trained LLM and a partially or completely frozen speech encoder. SALMONN [3] and WavLLM [5] both implement dual-encoder architectures; however, their objectives are distinct. SALMONN integrates Whisper and BEATs to encompass both verbal content and non-verbal audio semantics, while WavLLM integrates Whisper and WavLM to distinguish semantic content from speaker characteristics. At the adapter stage, SALMONN implements a window-level Q-Former that partitions encoder output into fixed-size windows and executes cross-attention with learnable queries, thereby generating a concise token sequence for the LLM. Instead, WavLLM employs per-encoder linear projectors in conjunction with a prompt-aware LoRA adapter that automatically adjusts the adaptation strength based on the input task. This contrast emphasizes a fundamental trade-off: SALMONN prioritizes representational breadth through multi-source encoding, whereas WavLLM concentrates on adaptive task routing by conditioning on task identity.

Qwen-Audio [7] broadens the scope of this architecture by resolving label-granularity inconsistencies across diverse datasets using a hierarchical tag-based system by concurrently training on over 30 audio task types. Neither SALMONN nor WavLLM addressed this scalability issue. Gaido et al. [1] conduct a thorough examination of the paradigm in which they categorize pipeline configurations into five functional components: the speech foundation model, length adapter, modality adapter, prompt-speech mixer, and LLM backbone. Their analysis of 9 published systems indicates no agreement on the most appropriate components. Unique task performance profiles are produced by prompt-speech blending strategies, including prepending, appending, and interleaving. Multi-layer Transformers and feed-forward networks comprise modal converters. Segmented Q-Formers, convolutional down-samplers, and CTC compression units are all examples of length adapters.

2.2 Unified Decoder-only Frameworks

By integrating speech discretization into token sequences within a cohesive autoregressive decoder, this method enables a fluid generative process across both modalities. The cross-modal alignment methodology and prioritized capabilities of systems in this paradigm are markedly different.

SpeechGPT [2] implements a three-stage training methodology that includes modality-adaptation pre-training, cross-modal instruction fine-tuning, and chain-of-modality tuning. This approach initially establishes

a fundamental understanding of speech tokens, subsequently enhances instruction-following, and ultimately enhances speech generation quality by conditioning it on an antecedent textual response. The curriculum's real-time deployment is impeded by sequential delays, despite its structure and effectiveness. VoxtLM [8] employs a multitask methodology that integrates speech recognition, synthesis, text generation, and speech continuation into a single shared decoder. Concurrent training improves performance relative to single-task baselines (TTS intelligibility CER decreased from 28.9 to 5.6), but it requires a more strenuous vocabulary acquisition process. Cross-modal transfer is prioritized over task diversity in AudioPaLM [6] and SPIRIT-LM [9]. AudioPaLM exhibits zero-shot cross-lingual speech-to-speech translation by integrating audio identifiers into its lexicon and constructing upon a robust text LLM (PaLM-2). Instead, SPIRIT-LM achieves cross-modal few-shot transfer by employing word-level interleaving during pre-training, thereby eliminating the need for task-specific fine-tuning. This distinction emphasizes the design decision between transfer via initialization and transfer via training-data architecture. LaRA [10] discovers that conventional low-rank adaptation is insufficient for cross-modal learning across all systems in this paradigm. This is because effective speech-text alignment requires substantially higher adaptation ranks than those required for unimodal fine-tuning. Consequently, parameter-efficient techniques that are designed for text-only scenarios do not transfer directly.

2.3 Multi-stream and Hierarchical Architectures

This framework meets the requirements for real-time verbal communication and high-quality voice synthesis by employing simultaneous multi-stream generation and hierarchical token modeling. Moshi [11] and GPST [12] both employ a hierarchical decomposition of speech synthesis into semantic and acoustic phases, each with its own set of objectives.

Moshi [11] endeavors to enable full-duplex spoken dialogue by representing user and system speech as two simultaneous audio token streams. This eliminates the necessity for explicit turn-taking and enables overlapping discourse and backchanneling. The Inner Monologue technique, which maintains streaming operation at approximately 200 ms latency, enhances linguistic coherence and factual precision by integrating synchronized text fragments within each frame. This results in a hierarchical structure of text, semantics, and acoustics. GPST [12] places a higher value on high-fidelity single-speaker speech than on interactive discourse. It concurrently processes both token types within a unified architecture, rather than sequentially, by integrating a global Transformer for long-range semantic structure and a local Transformer for fine-grained acoustic structure. This results in speaker likeness and voice quality comparable to those of multi-stage systems, while using approximately 33% fewer parameters. Additionally, it allows for the generation of spoken language and 24 kHz synthesis, which are not available in Moshi. Collectively, they substantiate the efficacy of hierarchical semantic-to-acoustic decomposition as a structural principle that applies to both high-fidelity offline generation (GPST) and real-time dialogue (Moshi).

2.4 Cross-paradigm Synthesis

The three paradigms consistently demonstrate a trade-off between modularity and integration depth. Encoder-adapter pipelines improve component reusability and reduce training costs; however, they generate information bottlenecks at the adapter interface and restrict the extent of cross-modal interaction. Decoder-only architectures overcome this constraint by integrating speech and text within a comprehensive representational framework, which facilitates stronger cross-modal transfer and few-shot generalization. However, this approach comes at the expense of more complex joint vocabulary development and increased training requirements. The most engineering effort is required to implement multi-stream architectures, which are the least mature in terms of deployment scope. However, they facilitate real-time streaming and concurrent modality streams, thereby maximizing integration depth. Consequently, the selection of a paradigm is primarily influenced by the application's requirements rather than by the paradigm’s intrinsic architectural quality.

3. Training Strategies and Cross-modal Alignment

Training methodologies that maintain the LLM's pre-existing linguistic expertise and align cross-modal representations are essential for the successful integration of speech and language abilities.

3.1 Multi-stage Training and Activation Tuning

Many encoder-adapter-LLM systems employ a two-phase protocol: an initial alignment phase that trains adapter modules on large ASR or audio captioning datasets with the LLM frozen, followed by instruction tuning on task-specific instruction-response pairs [1, 3]. SALMONN identifies this failure mode as task over-fitting [3], which impedes open-ended cross-modal reasoning, as the model's output distribution converges toward predictable transcription-style outputs due to the emphasis on short, deterministic responses in instruction tunings. SALMONN resolves this issue by employing a third-stage activation-tuning procedure that reduces the LoRA scaling factor, thereby yielding a diverse array of self-supervised responses. These responses are subsequently used as training data for a final fine-tuning phase, which restores performance on open-ended tasks such as audio narrative generation and spoken question answering with minimal supplementary annotation.

WavLLM implements curriculum learning [5], which progresses from single-task training on fundamental inputs (ASR, speaker verification, emotion recognition) to multi-task training on complex composite audio segments. The development of structured analytical responses and chain-of-thought reasoning regarding speech content, which were previously unattainable for speech LLMs, is facilitated here.

3.2 Interleaved Speech-text Pre-training

In contrast, an alternative method directly integrates speech and text tokens during pre-training, thereby eliminating both explicit alignment objectives and task-labeled data. SPIRIT-LM [9] trains textual LLMS on sequences that alternate between speech and text tokens within the same context window by utilizing self-selected word-level alignments. Standard language modeling objectives implicitly establish cross-modal correspondences. The method's applicability to morphologically complex or low-resource languages may be limited by the primary determinant of feasibility: the availability of reliable word-level alignments. Nevertheless, this limitation can be overcome by implementing multi-stage instruction tuning. A similar yet distinct methodology is employed by AudioPaLM [6], which trains a pre-trained text LLM on a combination of speech and text data rather than exclusively interleaved sequences and enhances its vocabulary with audio tokens.

3.3 Scaling Behavior and Computational Efficiency

Cuervo and Marxer [4] conduct a thorough analysis of the scaling properties of spoken-language models. They train over 50 models with varying parameter counts and data volumes. The findings of their investigation suggest that the loss of speech LM test exhibits a power-law scaling pattern that is comparable to that of text LLMs. Nevertheless, the scaling efficiency of downstream linguistic performance—as assessed by syntactic (BLIMP) and semantic benchmarks (Topic Cloze, Story Cloze)—is approximately three orders of magnitude lower than that of text-based models in comparison. This provides a quantitative justification for developing speech-capable systems using pre-trained text LLMs rather than training them from scratch. The significance of data quality and curation for SLMs, regardless of computational scale, is underscored by their supplementary discovery that semantically enriched pre-training data—as demonstrated by an auditory adaptation of the TinyStories dataset—significantly enhances downstream semantic efficiency.

4. Downstream Applications and Evaluation

The models that were evaluated have been subjected to a variety of tasks. Illustrative application domains, benchmarks, and key findings are delineated in Table 2.

In addition to the individual results previously summarized, several cross-study observations warrant further discussion. Gaido et al. [1] observe that the need for standardized evaluation criteria is underscored by substantial variation in training datasets, language pairs, and LLM sizes, which impedes direct comparisons across studies in their comprehensive assessment of SFM+LLM systems for speech translation. The conventional WER metric inadequately penalizes semantic significance for errors in domain-specific terminology, a methodological issue that transcends any single system, as noted by Wang et al. [14]. The severity-aware WER (SWER) metric was developed to address this motivation, assigning weights to errors based on content type. Voas et al. [15] introduce the VG-SPICE task, which involves generating and modifying

visual scene graphs from spoken dialogue. The AViD-SP model they developed demonstrates that raw audio inputs provide additional information beyond what ASR transcriptions can capture. This principle has implications for a diverse array of downstream speech-LLM applications.

Table 2: Representative Speech-llm Applications Across Domains

Application Domain	Representative Work	Key Benchmarks	Notable Result
ASR & Speech Translation	Gaido et al. [1]; GenTranslate [13]	CoVoST-2, MuST-C, FLEURS, WMT	GenTranslate surpasses the previous state-of-the-art in both speech and machine translation
Multimodal Scientific ASR	Wang et al. [14]	MS-ASR (severity-aware WER)	SWER enhancement of 45% with visual context from presentations
Spoken Dialogue	Moshi [11]	Spoken QA, audio quality metrics	State-of-the-art performance among speech-to-text models in spoken query answering; full-duplex communication with an estimated 200ms latency
Semantic Parsing	Voas et al. [15]	VG-SPICE	Raw audio provides supplementary information that surpasses the scope of ASR transcriptions
Depression Detection	Zhang et al. [16]	DAIC-WOZ	State-of-the-art performance on audio-text benchmarks that employ lightweight acoustic landmarks
Language Documentation	He et al. [17]	FIELDWORK (37 languages)	In the glossing and underlying representation subtasks, cascaded systems outperform end-to-end models.
Pronunciation Assessment	Yan et al. [18]	speechocean762	All evaluation dimensions are improved by hierarchical modeling that employs a correlation-aware regularizer

5. Discussion

5.1 The Modularity-integration Trade-off and Its Implications

The three paradigms advance along a continuum of modularity and integration. Encoder-adapter-LLM pipelines are a prime example of modularity, as they are constructed using a component-based architecture. This architecture promotes incremental development, reduces training costs, and promotes the reuse of pre-trained encoders and LLMs. However, it also restricts cross-modal interaction and generates information bottlenecks at the adapter interface. This issue is resolved by unified decoder-only systems, which enhance cross-modal interaction and few-shot transfer by integrating speech and text within a single representational framework. However, this approach is accompanied by increased architectural complexity and more stringent training requirements. Multi-stream hierarchical systems are more sophisticated, as they provide concurrent modality streams and hierarchical token generation for real-time full-duplex communication. However, they are limited in scope of deployment and require significant engineering effort.

The results of LaRA [10] and Cuervo & Marxer [4] collectively indicate that speech–text integration is not amenable to unimodal fine-tuning and is more advantageous when facilitated by architectural and data design than when facilitated by scaling alone. The interplay of architectural innovation, data design, and training methodology will determine future progress.

5.2 Training Failure Modes and Data Efficiency

The task overfitting issue that SALMONN observes during instruction tuning [3] suggests a broader methodological risk: any speech-LLM fine-tuned on short, deterministic instruction data is at risk of falling into a similar distributional collapse. A systematic investigation is warranted to determine the efficacy of activation tuning and related regularization strategies across a variety of architectures and task distributions. The curriculum-learning experiments in WavLLM [5] further illustrate that training-order decisions significantly influence compositional reasoning capability. This factor has received relatively little attention compared to architectural design decisions.

5.3 Evaluation Standardization

The absence of well-defined evaluation criteria is a substantial structural constraint. The establishment of reliable performance baselines and the identification of optimal design choices in controlled settings are impeded by the variability in modeling sizes, language pairings, evaluation metrics, and training datasets across studies. Wang et al. [14] (severity-aware WER) and He et al. [17] (FIELDWORK corpus) provide valuable domain-specific contributions; however, additional standardization is required. The retention of paralinguistic information, robustness to acoustic noise and domain shifts, computational efficiency, latency, and task accuracy should be the focus of future evaluation frameworks. Controlled ablation studies that isolate specific architectural components—identified as critical by Gaido et al. [1]—will be required to ascertain which design choices enhance performance and under what conditions. However, such studies are exceedingly uncommon in the current literature.

6. Conclusion

18 studies on speech–LLM integration were surveyed across three architectural paradigms—encoder-adapter-LLM pipelines, unified decoder-only frameworks, and multi-stream hierarchical designs—each occupying a distinct position on the modularity–integration continuum. No single paradigm dominated all evaluation criteria. The architecture is not the sole determining factor; the training strategy and data curation are equally important, especially given the significant scaling inefficiency of spoken-language models compared to text LLMs.

Numerous directions require immediate attention. To conduct reliable cross-system comparisons, it is imperative to employ standardized multitask evaluation criteria. Cross-modal parameter-efficient adaptation techniques that surpass the rank limits of unimodal fine-tuning are scarce. In the absence of a comprehensive evaluation of hybrid training methodologies that integrate interleaved pre-training with curriculum-scheduled fine-tuning within a cohesive architecture and dataset, a substantial methodological gap remains. Further development is necessary for full-duplex streaming communication techniques to be widely adopted. Lastly, the pragmatic necessity and conceptually underexplored area of speech-LLM integration for low-resource and endangered languages are both evident. The potential of speech-to-text foundation models to match the generality and effectiveness of their text-only counterparts will depend on progress in these domains.

References

[1] Gaido, M., Papi, S., Negri, M., & Bentivogli, L. (2024). Speech translation with speech foundation models and large language models: What is there and what is missing? In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 14760–14778). Association for Computational Linguistics.

[2] Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., & Qiu, X. (2023). SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 15757–15773). Association for Computational Linguistics.

[3] Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., & Zhang, C. (2024). SALMONN: Towards generic hearing abilities for large language models. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024).

[4] Cuervo, S., & Marxer, R. (2024). Scaling properties of speech language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 351–361). Association for Computational Linguistics.

[5] Hu, S., Zhou, L., Liu, S., Chen, S., Meng, L., Hao, H., Pan, J., Liu, X., Li, J., Sivasankaran, S., Liu, L., & Wei, F. (2024). WavLLM: Towards robust and adaptive speech large language model. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 4552–4572). Association for Computational Linguistics.

[6] Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borber, Z., Riesa, J., Tanaka, K., Lamania, T., Chen, J., Ghaffarizadeh, A., Mengibar, R., & others. (2023). AudioPaLM: A large language model that can speak and listen. arXiv preprint arXiv:2306.12925.

[7] Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., & Zhou, J. (2023). Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv preprint arXiv:2311.07919.

[8] Maiti, S., Peng, Y., Choi, S., Jung, J.-W., Chang, X., & Watanabe, S. (2024). VoxtLM: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE.

[9] Nguyen, T. A., Muller, B., Yu, B., Costa-jussà, M. R., Elbayad, M., Popuri, S., Ropers, C., Duquenne, P.-A., Algayres, R., Mavlyutov, R., Gat, I., Williamson, M., Synnaeve, G., Pino, J., Sagot, B., & Dupoux, E. (2024). SPIRIT-LM: Interleaved spoken and written language model. arXiv preprint arXiv:2402.05755.

[10] Shaik, Z. H., Hegde, P., Bannulmath, P., & T, D. K. (2024). LaRA: Large rank adaptation for speech and text cross-modal learning in large language models. In Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics.

[11] Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., & Zeghidour, N. (2024). Moshi: A speech-text foundation model for real-time dialogue. arXiv preprint arXiv:2410.00037.

[12] Zhu, Y., Su, D., He, L., Xu, L., & Yu, D. (2024). Generative pre-trained speech language model with efficient hierarchical transformer. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

[13] Hu, Y., Chen, C., Yang, C.-H. H., Li, R., Zhang, D., Chen, Z., & Chng, E. S. (2024). GenTranslate: Large language models are generative multilingual speech and machine translators. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

[14] Wang, M., Wang, Y., Vu, T.-T., Shareghi, E., & Haffari, R. (2024). Exploring the potential of multimodal LLM with knowledge-intensive multimodal ASR. In Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics.

[15] Voas, J., Mooney, R., & Harwath, D. (2024). Multimodal contextualized semantic parsing from speech. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

[16] Zhang, X., Liu, H., Xu, K., Zhang, Q., Liu, D., Ahmed, B., & Epps, J. (2024). When LLMs meet acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In Proceedings of Interspeech 2024. ISCA.

[17] He, T., Choi, K., Tjuatja, L., Levin, L., Neubig, G., & Mortensen, D. R. (2024). WAV2GLOSS: Generating interlinear glossed text from speech. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

[18] Yan, B. C., Li, J. T., Wang, Y. C., Wang, H. W., Lo, T. H., Hsu, Y. C., Chao, W. C., & Chen, B. (2024). An effective pronunciation assessment approach leveraging hierarchical transformers and pre-training strategies. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.