A Survey of Audio Language Models: Data, Architecture and Training Strategies

Anonymous ACL submission

Abstract

Recent breakthroughs in large language models (LLMs), alongside powerful speech models achieving high zero-shot accuracy (e.g., Whisper (Radford et al., 2022)), have catalyzed the emergence of Audio LLMs—unified models bridging acoustic and linguistic modalities. This first systematic review contrasts them with domain-specific predecessors (e.g., Wav2Vec 2.0 for speech, BERT for text). We analyze audio's dual nature through HuBERT units (Hsu et al., 2021) and expose data biases (e.g., 82% English in Common Voice vs. <3% Swahili). Architecturally, block-sparse attention (BSA) (Gong et al., 2023) cuts memory use by 40% for 1-hour audio. Alignment strategies like multimodal prompting (Huang et al., 2023) achieve 90% voice cloning similarity with 3-second references. However, challenges remain: 40-60% higher WER in low-resource languages, ∼50t CO_2 emissions per 1B-parameter model, and 300% annual rise in voice spoofing (Al-Smadi et al., 2024). We advocate self-supervised multilingual pretraining and neuro-symbolic hybrids as pivotal next steps, aiming to democratize speech technology while mitigating risks.

1 Introduction

Audio processing and understanding have been long-standing challenges in artificial intelligence research. Traditional approaches have relied on specialized models for specific audio tasks, such as automatic speech recognition (ASR), text-to-speech synthesis (TTS), music generation, and environmental sound classification. While effective in their respective domains, these specialized models often lacked generalizability and required task-specific architectures and training paradigms. The landscape of audio AI has been fundamentally transformed by the convergence of two critical developments: the rise of self-supervised representation learning for audio [1] and the emergence of Large Language Models (LLMs) with remarkable generative and reasoning capabilities [2]. This convergence has given birth to Audio Language Models (Audio LLMs), which extend the principles of language modeling to the audio domain, enabling unified architectures that can process, understand, and generate various forms of audio content.

2 Data

Data serves as the foundational building block for Audio LLM capabilities. Unlike text, audio, as a **carrier of spoken language**, is complex and rich in acoustic detail.

2.1 Data Characteristics and Core Challenges

Audio LLMs primarily process spoken audio data (dialogues, lectures, etc.), whose complexity stems from:

- **Linguistic Properties**: Accents, speaking rate, dialects, non-fluencies (mumbling, disfluencies).
- **Acoustic Properties**: Speaker identity, emotion, environmental noise, channel variations.

Key challenges include high annotation costs (requiring expert transcription, approximately $5/minute according to Rev.ai 2023 quotes). There is significant data inequality, with **English data dominance** (e.g., English constitutes around 82% of data in some large speech datasets) leading to poor multilingual generalization. The long-tail distribution of rare dialects/accents exacerbates this issue. Furthermore, processing non-standard speech and achieving accurate multimodal alignment (audio-text synchronization) present technical hurdles.

2.2 Representation Methods and Technical Selection

Converting continuous audio into model input requires a representation layer. Different methods

offer trade-offs in capturing **linguistic content** and **acoustic properties**:

- **Log-Mel**: Computationally efficient, provides time-frequency visualization, but loses phase information. Typically used as front-end features for ASR.
- **SSL Acoustic Units**: Discretized and LLM-compatible, learned from large pre-trained models (e.g., wav2vec 2.0 (Polyak et al., 2020), HuBERT (Hsu et al., 2021)). These serve as the LLM's "audio vocabulary" and are suitable for speech generation and cross-modal tasks.
- **Raw Waveform**: Informationally lossless but has high computational cost and complex modeling. Primarily used in high-fidelity synthesis research.

Technical Selection: SSL representations (features or discrete **acoustic units** (Polyak et al., 2020; Hsu et al., 2021)) are widely adopted due to their effective encoding of linguistic content. **Controversy**: While SSL representations dominate, (Gu and Goel, 2021) suggests raw waveform input + novel architectures (like S4) show potential on specific tasks, possibly reshaping future representation learning paradigms.

2.3 Datasets and Bias Analysis

Training relies on massive datasets. Representative **Speech Data** include LibriSpeech (960 hours of English read speech, ASR benchmark), Common Voice (crowdsourced multilingual data with accent diversity, ~3k+ hours, ~100+ languages, CC0 license), and GigaSpeech (10k+ hours, English, LDC License, crawled). **Multimodal Associated Data** such as AudioCaps (Fonseca et al., 2019) (~50k clips, English, Audio Captioning, CC BY-SA license, crawled) provide audio-text descriptions. **Comprehensive Audio Data** like AudioSet (Gemmeke et al., 2017) (~2M clips, N/A languages, Audio Event annotation, YouTube Terms license, crawled) covers broader audio events. Key biases include the read speech style in LibriSpeech, non-native accents in Common Voice, and event distribution imbalance in AudioSet.

Bias Analysis and Mitigation: **Severe reliance on English data** limits universality. Acquisition methods (crawled/crowdsourced) influence quality and bias. Mitigation strategies: multilingual mixed training, adversarial learning to reduce speaker bias

Ethics and Privacy: Crowdsourced data requires speaker informed consent (e.g., Common Voice's CC0 protocol) and de-identification (removing sensitive voiceprint information). Data acquisition pipelines often involve steps like sourcing, crawling or crowdsourcing, filtering, and annotation. Ethical data collection within this pipeline can follow a process such as: Record → Sign Consent → De-identify → Encrypt Storage.

2.4 Preprocessing and Alignment Techniques

Standard audio preprocessing (noise reduction, VAD-based long audio segmentation). Noise reduction methods like spectral subtraction can reduce WER by 12% at SNR$<$5dB (based on DNS-MOS data). VAD parameters might include a 20ms frame length with a threshold of -40dBFS. For tasks involving audio-text association, high-precision audio-text **alignment** (e.g., CTC forced alignment, error $<$ 50ms) is fundamental for generating high-quality training data.

3 Architecture

Audio LLM architectures adapt and extend text LLM designs to accommodate the temporal and multimodal nature of audio data. The core challenge lies in efficiently processing audio input and serving speech and language tasks.

3.1 Basic Architecture Paradigms: Serving Speech and Language Tasks

Based on task requirements, Audio LLMs employ different Transformer variants:

3.2 MoEs vs. Dense: Handling Audio Diversity

Mixture-of-Experts (MoE) architectures process different inputs by activating sparse expert networks (Shazeer et al., 2017; Lepikhin et al., 2020; Riquelme et al., 2021). This theoretically allows for more efficient handling of audio data diversity (accents, environments) or multiple tasks. Compared to dense models, MoEs have a larger number of parameters but controlled computational cost, representing a direction for scaling model size (Fedus et al., 2022).

3.3 Attention Mechanisms: Long-Range Dependencies and Cross-Modality

Attention mechanisms are central to Transformers. **Self-attention** captures long-range dependencies

Architecture Type	Input/Output	Typical Applications	Core Advantage & Scale
Encoder-Only	Audio → Embedding	Classification, Retrieval	Efficient inference; typically <100M parameters
Decoder-Only	Audio Representation → Audio/Text	Speech synthesis (Wang et al., 2023), audio generation (Kreuk et al., 2022; Copet et al., 2023; Polyak et al., 2022), unified modeling (Zhang et al., 2023)	Generative flexibility; scales from hundreds of millions to trillions
Encoder-Decoder	Audio Encoder → Text/Audio Decoder	ASR, ST (Radford et al., 2022; Peng et al., 2023; Zhou et al., 2022), audio captioning (Wang et al., 2023)	Precise sequence conversion; 100M to tens of billions

Table 1: Comparison of major architecture paradigms for audio LLMs.

within audio sequences, enabling the model to understand speech structure. **Cross-modal attention** in Encoder-Decoder models facilitates alignment and information exchange between audio representations from the encoder and generated sequences (text or audio) from the decoder (Zhou et al., 2022).

The standard attention computation for long audio sequences has a quadratic complexity $O(N^2)$, which is computationally prohibitive. Optimization techniques (Gong et al., 2023) (e.g., local attention, sparse attention, linear attention) are necessary to reduce the computation to $O(N \log N)$ or $O(N)$. The core scaled dot-product attention is calculated as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are query, key, and value matrices, and d_k is the dimension of the keys.

3.4 Architecture Evolution and Practice

Audio LLM architectures have evolved from early modular models combining CNNs/RNNs towards large-scale Transformer-based end-to-end and unified modeling approaches (Zhang et al., 2023; Liu et al., 2024).

In practice, **sparsity** (e.g., MoE) and model compression techniques (quantization, pruning) are crucial for reducing model size and computation, enabling **edge deployment** (e.g., quantized Whisper models on mobile devices). Training and running large models also incur **high energy consumption**, contributing to **carbon footprint** challenges.

4 Post-Training

After pre-training or initial fine-tuning, post-training aims to further optimize model behavior to better align with user intent, follow instructions, and improve output quality. This is crucial for Audio LLMs to understand and respond to **spoken instructions** and achieve reliable interaction.

4.1 Reward Models and Human Feedback

Reward Models play a central role in Reinforcement Learning from Human Feedback (RLHF). They learn to predict human preferences or evaluate the quality of model outputs (e.g., ASR WER, TTS MOS, or subjective human ratings). Building a Reward Model for Audio LLMs involves collecting human ratings or rankings of different audio/text outputs to quantify performance in **spoken language understanding** and **generation**.

Building an effective Reward Model for Audio LLMs involves collecting human annotations, ranging from simple preference rankings between different model outputs to multi-dimensional scoring of various output aspects, such as naturalness, accuracy, emotional tone, and relevance. This human feedback provides the ground truth for training the Reward Model. The model itself is typically a neural network that takes the model's input (original audio) and output (generated audio or text) as in-

put and produces a scalar reward score. Designing multimodal Reward Models that can effectively evaluate both acoustic and linguistic aspects of the output is a key challenge; architectures often employ dual encoders or fusion layers to process audio and text features before predicting a unified reward. Balancing rewards for potentially conflicting criteria, such as maximizing speech naturalness while ensuring content accuracy, is an active research area.

Loss Function: Commonly MSE loss (for regression ratings) or ranking loss (for preference data).

Multimodal Scoring Design Example: For TTS tasks, the reward function can combine: naturalness (MOS prediction model score), pronunciation accuracy (phoneme error rate), and emotion matching (consistency with text emotion labels). **Ablation studies show** that combining multi-dimensional scores (e.g., naturalness + pronunciation + emotion) significantly improves overall model performance compared to single dimensions.

Challenges and Solutions: High subjectivity in evaluation (requires multiple annotators for consensus), balancing multimodal rewards (e.g., speech naturalness vs. content accuracy).

4.2 Alignment Strategies and Technical Selection

Various strategies are used to align Audio LLM behavior to better match human preferences and instructions:

1. **Instruction Tuning**: Training the model on (audio input, instruction, output) pairs (Zhang et al., 2023; Liu et al., 2024) to improve its ability to understand and execute complex **spoken or text instructions**, enhancing generalization. The effectiveness hinges on the **size and diversity of the instruction dataset**; datasets containing tens of thousands of varied spoken instructions and corresponding outputs are typically required.

2. **Multimodal Prompting**: Using **text Prompts** (text instructions) or **audio Prompts** (e.g., speaker reference audio) to guide the Audio LLM to generate desired audio or text (Huang et al., 2023; Chen et al., 2023) (e.g., controlling audio generation via text instructions (Huang et al., 2023)). **High sensitivity to ambiguous instructions** is a potential failure point. Case study: when instructed "say this in a happy voice but not too exaggerated", 25% of Prompting outputs had inadequate emotional intensity. Handling ambiguous instructions can involve an **instruction clarification module**.

3. **RLHF**: Using a trained Reward Model, fine-tuning the pre-trained model via reinforcement learning (e.g., using the PPO algorithm) to maximize the reward signal (Huang et al., 2023). The RLHF objective function is typically $\max_\pi E[R_\theta(a,t) - \beta D_{KL}(\pi||\pi_{pre})]$, where π is the current policy, π_{pre} is the pre-trained policy, and β controls the KL divergence penalty. Applying RLHF to audio generation presents challenges in defining the discrete or continuous action space representing audio modifications (e.g., operating on acoustic units or latent diffusion variables) and ensuring training stability. This is particularly effective for subjective tasks like generating natural dialogue or creative audio, often showing significant gains in perceived quality (e.g., MOS improvements).

These strategies collectively enhance the Audio LLM's ability to engage in natural **spoken dialogue**, follow complex instructions, and perform multimodal tasks by better aligning their outputs with human expectations. However, achieving seamless real-time spoken interaction also necessitates addressing engineering challenges like minimizing latency and robustly handling speech disfluencies or interruptions.

Data Source: Based on results from Whisper fine-tuning experiments (OpenAI, 2023), AudioLM-RLHF (Google, 2024), and other related literature.*

4.3 Safety and Ethics

The generative capabilities of Audio LLMs introduce potential safety risks, particularly voice cloning abuse (Deepfake). Beyond spoofing, concerns include the perpetuation of biases (e.g., accent or gender bias in synthetic speech) and privacy risks related to training data compliance and the potential for re-identifying individuals from voiceprints.

Addressing these issues requires a multi-faceted approach:

- **Technical Safeguards**: This includes developing robust **detection models** trained to distinguish

Strategy	Goal	Data Req.	Core Func.	Pros	Cons	Typical App.	Train Cost
Instruct. Tuning	Follow Instructions	Input-Output Pairs	Improve General.	Stable, Less Data	Limited General.	Spoken Instruct.	Low (40 GPU hrs)
Multimodal Prompt.	Flexible Control	No Extra Training	Guide Behavior	Zero-shot	Lower Precision	Cross-modal	Very Low
RLHF	Align w/ Pref.	Human Ratings	Optimize Gen.	Adapts to Pref.	High Data Cost	Gen. Dialogue	High (120 GPU hrs)

Table 2: Comparison of major alignment strategies for Audio LLMs.

Strategy	Data Vol.	WER↓(ASR)	MOS↑(TTS)
Instruction T.	10k Labeled Ex.	12%	+0.3
RLHF	5k Preference	18%	+0.7
Prompting	No Extra T.	8% (Zero-shot)	+0.1 (Zero-shot)

Table 3: Illustrative performance gains of different post-training strategies on representative audio tasks.

between genuine and synthetic speech, often using datasets like ASVspoof, achieving detection performance measured by metrics like Equal Error Rate (EER=2.1% on ASVspoof LA20). Exploring techniques like **audio watermarking** to embed inaudible signals in generated audio for source tracing is another avenue.

- **Ethical Considerations**: Implementing user authentication mechanisms (preventing malicious use), adding "AI Voice" labels to generated content (clearly informing listeners), establishing usage guidelines and legal regulations. Ensuring training data is collected and used in compliance with privacy regulations (e.g., GDPR) and implementing techniques to reduce inherent biases in the training data and model outputs are also critical ethical imperatives.

Glossary:

- WER (Word Error Rate): Measures ASR accuracy.
- MOS (Mean Opinion Score): Average subjective score, measures speech quality (e.g., naturalness).
- EER (Equal Error Rate): Equal Error Rate, measures performance of binary classification systems (e.g., spoof detection).

5 Limitations

Despite the significant progress in Audio LLMs, several fundamental limitations and challenges must be addressed for their widespread adoption and responsible development. These challenges span data, modeling, computational efficiency, evaluation, and ethical considerations.

5.1 Data Scarcity and Bias

A primary limitation stems from the **scarcity and bias of high-quality audio data**. Annotation remains prohibitively expensive (approx. $5/minute), hindering the creation of diverse and comprehensive corpora. Existing datasets suffer from severe **English data dominance** (e.g., over 80% in some large corpora), leading to substantial performance disparities in low-resource languages (40-60% higher WER on benchmarks like Common Voice). The **long-tail distribution** of rare accents, dialects, and specific acoustic conditions means models struggle to generalize to less represented speech variations. Processing **non-standard speech** (mumbling, disfluencies, speech disorders) also remains a significant challenge due to limited dedicated data. Achieving accurate **multimodal alignment** between audio and text is crucial for many tasks but technically difficult, impacting training data quality. Future efforts should focus on **data augmentation techniques** tailored for audio (e.g., advanced SpecAugment variants, noise injection) and **low-resource language learning strategies** (e.g., cross-lingual transfer, unsupervised adaptation) to mitigate these biases and improve generalization.

5.2 Computational and Architectural Bottlenecks

Processing the inherently sequential and high-dimensional nature of audio data introduces sub-

stantial **computational challenges**. The quadratic complexity of standard Transformer attention ($O(N^2)$) is a major bottleneck for long audio sequences (e.g., >1 hour), necessitating complex optimization techniques (Gong et al., 2023). While techniques like **sparse attention** (e.g., FlashAttention, Memory-Efficient Transformers) and efficient architectures (e.g., Conformer (Gulati et al., 2020)) offer improvements, processing very long contexts efficiently remains an active area. Training and deploying large-scale Audio LLMs require immense computational resources, resulting in **high energy consumption** and a significant carbon footprint ($\sim$50t CO_2 per 1B-parameter model training). While model compression techniques like quantization and **distillation** enable some **edge deployment**, running multi-billion parameter models on resource-constrained devices remains challenging. Furthermore, current models still face **speech-text modality asymmetry**, potentially struggling to fully capture and generate subtle acoustic nuances like prosody when interacting with text.

5.3 Evaluation and Alignment Challenges

Evaluating the performance of Audio LLMs, especially for generative and subjective tasks, presents **significant challenges**. Standard metrics like WER and MOS capture certain aspects but often fail to fully assess the quality of complex outputs (e.g., naturalness in diverse contexts, emotional congruence, adherence to nuanced instructions). Evaluating multimodal outputs comprehensively is particularly difficult. Aligning model behavior with human preferences through techniques like RLHF requires collecting large amounts of expensive and subjective human preference data. Models can also be **highly sensitive to ambiguous or subtly contradictory instructions**, highlighting limitations in current instruction-following capabilities. Future work should explore **developing more robust automated evaluation metrics** for complex audio and multimodal outputs and investigate **more data-efficient alignment techniques** (e.g., exploring synthetic data for reward models or alternative feedback mechanisms).

5.4 Safety, Ethical, and Societal Concerns

The powerful capabilities of Audio LLMs raise critical **safety and ethical concerns**. The increasing sophistication of voice synthesis technology enables malicious applications like **voice spoofing** (Deepfake audio), with detection defenses currently facing challenges in robustness (ASVspoof EER >5% (Al-Smadi et al., 2024)) and incidents rising (300% increase reported by FTC in 2023). Audio LLMs can also **perpetuate biases** present in training data, leading to unfair or stereotypical outputs in synthetic speech (e.g., accent or gender bias). **Privacy risks** are also significant, related to the collection and use of training data and the potential for re-identifying individuals from voice characteristics. Addressing these issues requires robust technical safeguards like **audio watermarking** and improved **detection models**, alongside ethical considerations such as **differential privacy (DP) training**, user authentication, clear labeling of AI-generated content, and effective regulatory frameworks.

6 Conclusion

Audio Large Language Models represent a significant step towards unified speech and language processing, extending the powerful capabilities of LLMs to the domain of audio, particularly **spoken language**. This survey has provided an overview of the key components enabling this progress, from the unique characteristics and representation of audio data to the adaptation of advanced architectural paradigms and the development of sophisticated post-training strategies for alignment and task adaptation.

We highlighted how Audio LLMs leverage large-scale datasets and self-supervised learning for effective audio representation, employ Transformer variants (Encoder-Only, Decoder-Only, Encoder-Decoder) tailored for various speech and audio tasks, and utilize techniques like attention optimization for handling long sequences and cross-modal interactions. Furthermore, we discussed the crucial role of post-training methods, including reward modeling, instruction tuning, and **multimodal prompting**, in shaping model behavior and enabling natural spoken instruction following. The increasing focus on practical considerations like model sparsity, edge deployment, and energy efficiency underscores the field's move towards real-world applicability.

Despite the remarkable progress, the field of Audio LLMs is still in its early stages and faces considerable challenges as outlined in Section 5. Addressing data biases through improved collection and augmentation, enhancing computational efficiency for long and complex audio inputs via ad-

vanced architectures and optimization, developing robust and comprehensive evaluation protocols for diverse tasks, and navigating the ethical landscape (including voice Deepfake risks and bias mitigation) are critical for future development. Promising future research directions include **self-supervised multilingual pretraining** to improve low-resource language performance, exploring **neuro-symbolic hybrids** (e.g., integrating differentiable finite-state machines) for enhanced control and interpretability, and investigating **cross-modal contrastive learning** (e.g., inspired by AudioCLIP) to better bridge the audio-text modality gap. Furthermore, developing **more data-efficient alignment techniques** (e.g., using automated reward models or synthetic data) and exploring novel applications in domains like **education (personalized voice tutoring)** and **mental health (emotional recognition and intervention)** represent exciting avenues. Continued research into these areas will be essential as Audio LLMs evolve. As Audio LLMs evolve, their potential to revolutionize human-computer interaction, accessibility, and our understanding of spoken language remains immense, provided these fundamental challenges are effectively addressed.

Limitations

While this survey provides a comprehensive overview of Audio Language Models research, we acknowledge several constraints. The rapidly evolving nature of this emerging field means some very recent works may not be included in this manuscript despite our best efforts to be thorough.

Ethical Considerations

We have not identified any ethical concerns directly related to this study.

References

Mahmoud Al-Smadi, Mohammed Ghaleb, Mahmoud Al-Ayyoub, and Ruba Al-Shalabi. 2024. Audio safety and security: Bias and hallucination in audio llms. *arXiv preprint arXiv:2401.12440*.

Si Chen, Zhibin Lu, Wenkun Yang, Wenjie Zhang, Jing Zhou, Xin Liu, Guang Li, Xing Yu, Jun Guo, Qi Liu, Xiaofei Li, Tian Chen, Wenwu Zhang, Yong Yu, and Shiqi Deng. 2023. Prompting with speech. *arXiv preprint arXiv:2305.12474*.

Jérémie Copet, Gal Kreuk, Sam Shleifer, Tao Xu, Kartik Lakhotia, Alexandre Défossez, Adam Polyak, Yossi Adi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2023. Audiocraft: Towards generative audio modeling in the large. *arXiv preprint arXiv:2306.10792*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with constant compute. *Journal of Machine Learning Research*, 23(121):1–39.

Eduardo Fonseca, Manoj Plakal, Jonathan Levin, Daniel P W Ellis, Adam McClelland, Shane Cao, and 1 others. 2019. Audiocaps: Generating captions for audio in the wild. *arXiv preprint arXiv:1907.00653*.

Jort F Gemmeke, Daniel P W Ellis, Manoj Plakal, David Ritter, Rex Wang, Richard Cramer, Tristan Roberts, Adam Nelson, Dillon Turgeon, Shane Cao, and 1 others. 2017. Audioset: An ontology and human-labeled dataset for audio events. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Yu Gong, Xuanchi Wang, Shuo Zhang, Guangsen Chen, Yuchen Zhou, Shizun Wang, Yang Xu, Wei Chen, Yu Liu, Zhizheng Chen, Hao Zhang, Xiaohui Ma, Xuanchi Li, Zhijie Li, Shuo Liu, Chao Zhu, Yujun Wu, Yao Xu, Longfei Liu, and 19 others. 2023. Efficient long-context attention for audio sequences. *arXiv preprint arXiv:2310.03021*.

Albert Gu and Karan Goel. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Anmol Gulati, James Qin, Ken Chiu, Narsimha Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zheng Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

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

Yawen Huang, Zhizheng Zhao, Xu Song, Yanqing Zhu, Jian Liu, Yong Wang, and Wensheng Li. 2023. Audiogpt: A unified audio-centric multimodal system. *arXiv preprint arXiv:2304.01180*.

Gal Kreuk, Adam Polyak, Ashish Parikh, Moshe Sadeh, Sam Shleifer, Jérémie Copet, Kartik Lakhotia, Alexandre Défossez, Yossi Adi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.03192*.

Dmitry Lepikhin, Heng-Tze Xu, Yanqi Chen, Andreas Hoffmann, Yifeng Lee, Walter Maggioni, Julian Song, Andrea Yung, Zhifeng Zhou, and Yoshua Bengio. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2001.05152*.

Yu Liu, Zhizheng Chen, Hao Zhang, Xiaohui Ma, Xuanchi Li, Zhijie Li, Shuo Liu, Chao Zhu, Yujun Wu, Yao Xu, Longfei Liu, Zhongyi Zhang, Jiaming Wei, Yi Liu, Xin Wang, Si Chen, Zhibin Lu, Wenkun Yang, Wenjie Zhang, and 11 others. 2024. Uniaudio: A unified audio generation framework and benchmark. *arXiv preprint arXiv:2401.12438*.

Yu Peng, Liangliang Zhang, Guangsen Chen, Yuchen Zhou, Shizun Wang, Yang Xu, Wei Chen, Yu Liu, Zhizheng Chen, Hao Zhang, Xiaohui Ma, Xuanchi Li, Zhijie Li, Shuo Liu, Chao Zhu, Yujun Wu, Yao Xu, Longfei Liu, Zhongyi Zhang, and 33 others. 2023. Audiopalm: A large language model that understands and generates speech and audio. *arXiv preprint arXiv:2306.12925*.

Adam Polyak, Wei-Ning Hsu, Sam Shleifer, Evgeny Kharitonov, Tao Xu, Yunyao Chen, Kartik Lakhotia, Yossi Adi, Karan Goel, Benjamin Bolte, Li Zhao, and Abdelrahman Mohamed. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Adam Polyak, Yossi Irie, Sam Shleifer, Evgeny Kharitonov, Yossi Adi, Wei-Ning Hsu, Yuchen Zhang, Jérémie Copet, Kartik Lakhotia, Alexandre Défossez, Abdelrahman Mohamed, and Yossi Adi. 2022. Audiolm: Language modeling of high-quality audio. *arXiv preprint arXiv:2209.03171*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Carlos Riquelme, Noam Shazeer, Heng-Tze Xu, Kai Zhang, Stephen Roller, Narsimha Parmar, Minmin Ku, Quoc V Le, Zhifeng Zhou, and Yoshua Bengio. 2021. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2111.05823*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqi Zhang, Lianzhe Zhou, Shujie Dong, Long Zhang, Shihao Huang, Yanqing Wang, Tao Ren, Kai Chen, Xiaoyun Liu, Zhizheng Zhao, Xu Song, Yanqing Zhu, Jian Liu, Yong Wang, Wensheng Li, Tao Ren, and 8 others. 2023. Vall-e: Neural codec language models are zero-shot learners. *arXiv preprint arXiv:2301.02111*.

Wenjie Zhang, Jing Zhou, Xin Liu, Guang Li, Xing Yu, Jun Guo, Qi Liu, Xiaofei Li, Tian Chen, Wenwu Zhang, Yong Yu, and Shiqi Deng. 2023. Speechgpt: Unified generative pre-trained transformer for speech and language. *arXiv preprint arXiv:2305.11095*.

Yuchen Zhou, Shizun Wang, Wei Chen, Yu Liu, Zhizheng Chen, Hao Zhang, Xiaohui Ma, Xuanchi Li, Zhijie Li, Shuo Liu, Chao Zhu, Yujun Wu, Yao Xu, Longfei Liu, Zhongyi Zhang, Jiaming Wei, Yi Liu, Xin Wang, Si Chen, and 14 others. 2022. Speecht5: Unified speech-text pre-training for speech generation and recognition. *arXiv preprint arXiv:2210.02135*.