AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head

Rongjie Huang[1*], Mingze Li[1*], Dongchao Yang[2*], Jiatong Shi[3*], Xuankai Chang[3], Zhenhui Ye[1], Yuning Wu[4], Zhiqing Hong[1], Jiawei Huang[1], Jinglin Liu[1], Yi Ren[1], Yuexian Zou[2], Zhou Zhao[1], Shinji Watanabe[3]

[1]Zhejiang University
[2]Peking University
[3]Carnegie Mellon University
[4]Remin University of China

Abstract

Large language models (LLMs) have exhibited remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. Despite the recent success, current LLMs are not capable of processing complex audio information or conducting spoken conversations (like Siri or Alexa). In this work, we propose a multi-modal AI system named AudioGPT, which complements LLMs (i.e., ChatGPT) with 1) foundation models to process complex audio information and solve numerous understanding and generation tasks; and 2) the input/output interface (ASR, TTS) to support spoken dialogue. With an increasing demand to evaluate multi-modal LLMs of human intention understanding and cooperation with foundation models, we outline the principles and processes and test AudioGPT in terms of consistency, capability, and robustness. Experimental results demonstrate the capabilities of AudioGPT in solving **16** AI tasks with speech, music, sound, and talking head understanding and generation in multi-round dialogues, which empower humans to create rich and diverse audio content with unprecedented ease. Code can be found in https://github.com/AIGC-Audio/AudioGPT

Introduction

Nowadays, Large language models (LLMs) (Devlin et al. 2018; Raffel et al. 2020; Brown et al. 2020; Ouyang et al. 2022) are posing a significant impact on the AI community, and the advent of ChatGPT and GPT-4 leads to the advancement of natural language processing. Based on the massive corpora of web-text data and powerful architecture, LLMs are empowered to read, write, and communicate like humans.

Despite the successful applications in text processing and generation, replicating this success for audio modality (speech (Ren et al. 2020; Huang et al. 2022a; Hsu et al. 2021), music (Huang et al. 2021), sound (Yang et al. 2022), and talking head (Wu et al. 2021; Ye et al. 2023)) is limited, while it is highly beneficial since: 1) In real-world scenarios, humans communicate using spoken language across daily conversations, and utilize spoken assistant (e.g., Siri or Alexa) to boost life convenience; 2) As an inherent part of intelligence, processing audio modality information is a necessity to achieve artificial general intelligence. Understanding and generating speech, music, sound, and talking head could be the critical step for LLMs toward more advanced AI systems.

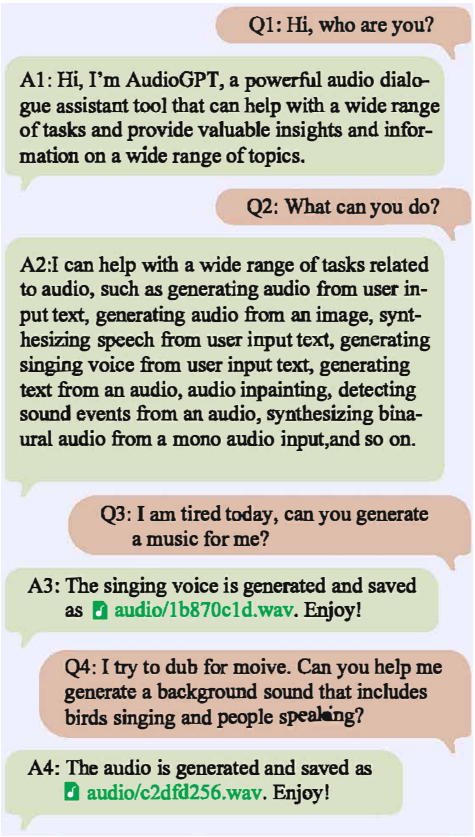

Figure 1: Qualitative analysis on multiple rounds of dialogue between humans and AudioGPT. Refer to https://aigc-audio.github.io/AudioGPT.github.io/ for details.

Despite the benefits of audio modality, training LLMs that support audio processing is still challenging due to the following issues: 1) Data: Obtaining human-labeled speech data is an expensive and time-consuming task, and there are only a few resources available that provide real-world spoken dialogues. Furthermore, the amount of data is limited compared

[*] Equal contributions

Task	Domain	Model
Speech Recognition	Speech	Whisper (Radford et al. 2022)
Speech Translation	Speech	MultiDecoder (Dalmia et al. 2021)
Style Transfer	Speech	GenerSpeech (Huang et al. 2022b)
Speech Enhancemen	Speech	ConvTasNet (Luo and Mesgarani 2019)
Speech Separation	Speech	TF-GridNet (Wang et al. 2022)
Mono-to-Binaural	Speech	NeuralWarp (Richard et al. 2021)
Audio Inpainting	Sound	Make-An-Audio (Huang et al. 2023)
Sound Extraction	Sound	LASSNet (Liu et al. 2022b)
Target Sound Detection	Sound	TSDNet (Yang et al. 2021)
Sound Detection	Sound	Pyramid Transformer (Xin et al. 2022)
Talking Head Synthesis	Talking Head	GeneFace (Ye et al. 2023)
Text-to-Speech	Speech	FastSpeech 2 (Ren et al. 2020)
Text-to-Audio	Sound	Make-An-Audio (Huang et al. 2023)
Audio-to-Text	Sound	MAAC (Ye et al. 2021)
Image-to-Audio	Sound	Make-An-Audio (Huang et al. 2023)
Singing Synthesis	Music	DiffSinger (Liu et al. 2022a) VISinger (Zhang et al. 2022)

Table 1: Supported Tasks in AudioGPT

to the vast corpora of web-text data, and multi-lingual conversational speech data is even scarcer; and 2) Computational resources: Training multi-modal LLMs from scratch is computationally intensive and time-consuming. Given that there are already existing audio foundation models that can understand and generate speech, music, sound, and talking head, it would be wasteful to start training from scratch.

In this work, we introduce "AudioGPT", a system designed to excel in understanding and generating audio modality in spoken dialogues. Specifically, 1) Instead of training multi-modal LLMs from scratch, we leverage a variety of audio foundation models to process complex audio information, where LLMs (i.e., ChatGPT) are regarded as the interface (Wu et al. 2023; Shen et al. 2023) which empowers AudioGPT to solve **16** audio understanding and generation tasks; 2) Instead of training a spoken language model, we connect LLMs with interface (ASR, TTS) for speech conversations; AudioGPT can be divided into four stages:

- Modality Transformation. Using input/output interface for modality transformation between speech and text, bridging the gap between the spoken LLMs and ChatGPT.
- Task Analysis. Utilizing the dialogue engine and prompt manager to help ChatGPT understands the intention of a user to process audio information.
- Model Assignment. Receiving the structured arguments for prosody, timbre, and language control, ChatGPT assigns audio foundation models for understanding and generation.
- Response Generation. Generating and returning a response to users after execution of foundation models.

Evaluating Multi-Modal LLMs

Consistency

In the consistency evaluation for the zero-shot setting, models are directly evaluated on the questions without being provided any prior examples of the specific tasks, which evaluate whether multi-modal LLMS could reason and solve problems without explicit training.

Capability

As the task executors for processing complex audio information, audio foundation models have a significant impact on handling complex downstream tasks.

Robustness

We evaluate the robustness of multi-modal LLMs by assessing their ability to handle special cases. These cases can be classified into the following categories: 1) Long chains of evaluation; 2) Unsupported tasks; 3) Error handling of multi-modal models; and 4) Breaks in context.

Experiments

Experimental Setup

In our experiments, we employ the gpt-3.5-turbo of the GPT models as the large language models and guide the LLM with LangChain. The deployment of the audio foundation models requires only a flexible NVIDIA T4 GPU on hugging face space. We use a temperature of zero to generate output using greedy search and set the maximum number of tokens for generation to 2048.

Case Study on Multiple Rounds Dialogue

Shown in Figure 1, AudioGPT demonstrates the capabilities for processing audio modality, covering a series of AI tasks in generating and understanding speech, music, sound, and talking head. The dialogue involves multiple requests to process audio information and shows that AudioGPT maintains the context of the current conversation, handles follow-up questions, and interacts with users actively.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Dalmia, S.; Yan, B.; Raunak, V.; Metze, F.; and Watanabe, S. 2021. Searchable Hidden Intermediates for End-to-End Models of Decomposable Sequence Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1882–1896.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 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.

Huang, R.; Chen, F.; Ren, Y.; Liu, J.; Cui, C.; and Zhao, Z. 2021. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3945–3954.

Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*.

Huang, R.; Lam, M. W.; Wang, J.; Su, D.; Yu, D.; Ren, Y.; and Zhao, Z. 2022a. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*.

Huang, R.; Ren, Y.; Liu, J.; Cui, C.; and Zhao, Z. 2022b. GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech Synthesis. *arXiv preprint arXiv:2205.07211*.

Liu, J.; Li, C.; Ren, Y.; Chen, F.; and Zhao, Z. 2022a. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, X.; Liu, H.; Kong, Q.; Mei, X.; Zhao, J.; Huang, Q.; Plumbley, M. D.; and Wang, W. 2022b. Separate what you describe: language-queried audio source separation. *arXiv preprint arXiv:2203.15147*.

Luo, Y.; and Mesgarani, N. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8): 1256–1266.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J.; et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140): 1–67.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Richard, A.; Markovic, D.; Gebru, I. D.; Krenn, S.; Butler, G.; de la Torre, F.; and Sheikh, Y. 2021. Neural Synthesis of Binaural Speech from Mono Audio. In *International Conference on Learning Representations*.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. *arXiv preprint arXiv:2303.17580*.

Wang, Z.-Q.; Cornell, S.; Choi, S.; Lee, Y.; Kim, B.-Y.; and Watanabe, S. 2022. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. *arXiv preprint arXiv:2209.03952*.

Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; and Duan, N. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Wu, H.; Jia, J.; Wang, H.; Dou, Y.; Duan, C.; and Deng, Q. 2021. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1478–1486.

Xin, Y.; Yang, D.; Zou, Y.; Yang, D.; and Zou, Y. 2022. Audio Pyramid Transformer with Domain Adaption for Weakly Supervised Sound Event Detection and Audio Classification. *Proc. Interspeech 2022*, 1546–1550.

Yang, D.; Wang, H.; Zou, Y.; Cui, F.; and Wang, Y. 2021. Detect what you want: Target sound detection. *arXiv preprint arXiv:2112.10153*.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2022. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2023. GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis. *arXiv preprint arXiv:2301.13430*.

Ye, Z.; Wang, H.; Yang, D.; and Zou, Y. 2021. Improving the performance of automated audio captioning via integrating the acoustic and semantic information. *arXiv preprint arXiv:2110.06100*.

Zhang, Y.; Cong, J.; Xue, H.; Xie, L.; Zhu, P.; and Bi, M. 2022. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7237–7241. IEEE.