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],
Zhou Zhao[1], **Shinji Watanabe**[3]

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

{rongjiehuang, limingze, zhaozhou}@zju.edu.cn, {dongchao98}@stu.pku.edu.cn,
{jiatongs, xuankaic, dongsli}@cs.cmu.edu,
{yuningwu}@ruc.edu.cn, {shinjiw}@ieee.org

https://github.com/AIGC-Audio/AudioGPT

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 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.

1 Introduction

Nowadays, Large language models (LLMs) (Devlin et al., 2018; Raffel et al., 2020; Brown et al., 2020; Ouyang et al., 2022; Zhang et al., 2022a) 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), music (Huang et al., 2021; Liu et al., 2022a), sound (Yang et al., 2022; Huang et al., 2023a), 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.

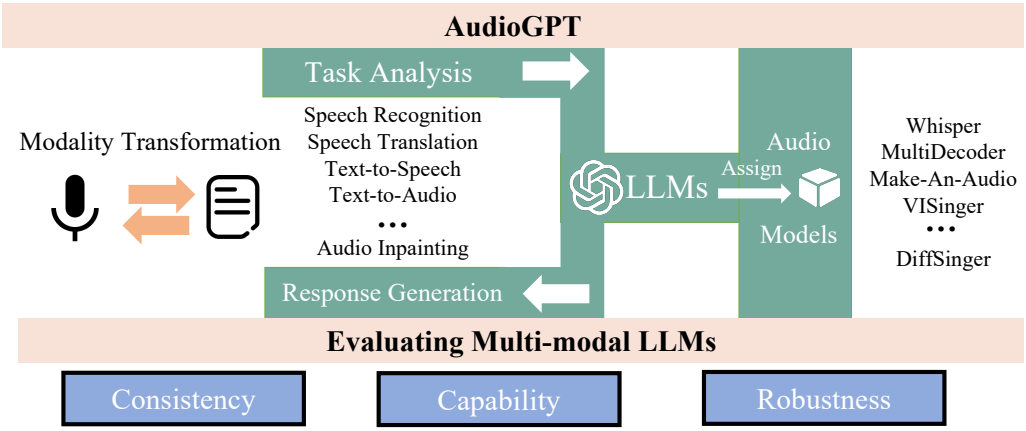
* Equal contributions

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 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 general-purpose interface (Wu et al., 2023; Shen et al., 2023) which empowers AudioGPT to solve numerous audio understanding and generation tasks; 2) Instead of training a spoken language model, we connect LLMs with input/output interface (ASR, TTS) for speech conversations; As illustrated in Figure 1, the whole process of 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 language 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 the audio foundation models for understanding and generation.
- Response Generation. Generating and returning a final response to users after the execution of audio foundation models.

Figure 1: A high-level overview of AudioGPT. AudioGPT can be divided into four stages, including modality transformation, task analysis, model assignment, and response generation. It equips ChatGPT with audio foundation models to handle complex audio tasks and is connected with a modality transformation interface to enable spoken dialogue. We design principles to evaluate multi-modal LLMs in terms of consistency, capability, and robustness.



As a blossoming research topics (Wu et al., 2023; Shen et al., 2023; Huang et al., 2023b), there is an increasing demand for evaluating the performance of multi-modal LLMs in understanding human intention and organizing the cooperation of multiple foundation models. In this work, we outline the design principles and process of evaluating AudioGPT in terms of consistency, capability, and robustness. Experimental results demonstrate the capabilities of AudioGPT for processing complex audio information in multi-round dialogue, covering a series of AI tasks including generating and understanding speech, music, sound, and talking head.

Key contributions of the paper include:

- We propose AudioGPT, which equips ChatGPT with audio foundation models to handle complex audio tasks. As a general-purpose interface, ChatGPT is connected with a modality transformation interface to enable spoken dialogue.
- We outline the design principles and process of evaluating multi-modal LLMs, and test AudioGPT in terms of consistency, capability, and robustness.
- Demonstrations present the efficiency of AudioGPT in audio understanding and generation with multiple rounds of dialogue, which empowers humans to create rich and diverse audio content with unprecedented ease.

2 Related Works

2.1 Large Language Models

The research areas of AI are being revolutionized by the rapid progress of Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Zhang et al., 2022a), where they can serve as a general-purpose language task solver, and the research paradigm has been shifting towards the use of LLMs. They have long been considered a core problem in natural language processing and demonstrated remarkable abilities for tasks such as machine translation (Gulcehre et al., 2017; Baziotis et al., 2020), open-ended dialogue modeling (Hosseini-Asl et al., 2020; Thoppilan et al., 2022), and even code completion (Svyatkovskiy et al., 2019; Liu et al., 2020).

Among them, Kaplan et al. (2020) studied the impact of scaling on the performance of deep learning models, showing the existence of power laws between the model and dataset sizes and the performance of the system. Language models (LMs) at scale, such as GPT-3 (Brown et al., 2020) have demonstrated remarkable performance in few-shot learning. FLAN (Wei et al., 2021) is proposed to improve the zero-shot performance of large language models, which would expand their reach to a broader audience. LLaMA (Touvron et al., 2023) shows that it is possible to achieve state-of-the-art performance by training exclusively on publicly available data, without resorting to proprietary datasets. The advent of ChatGPT and GPT-4 leads to rethinking the possibilities of artificial general intelligence (AGI).

2.2 Spoken Generative Language Models

Self-supervised learning (SSL) has emerged as a popular solution to many speech processing problems with a massive amount of unlabeled speech data. HuBERT (Hsu et al., 2021) is trained with a masked prediction with masked continuous audio signals. Inspired by vector quantization (VQ) techniques, SoundStream (Zeghidour et al., 2021) and Encodec (Défossez et al., 2022) present the hierarchical architecture for high-level representations that carry semantic information.

Most of these models build discrete units in a compact and discrete space, which could be modeled with an autoregressive Transformer whose predictions are then mapped back to the original signal space. Hayashi & Watanabe (2020) leverage discrete VQ-VAE representations to build speech synthesis models via autoregressive machine translation. "textless NLP" (Kharitonov et al., 2022; Huang et al., 2022c) is proposed to model language directly without any transcription by training autoregressive generative models of low-bitrate audio tokens. AudioLM (Borsos et al., 2022) and MusicLM (Agostinelli et al., 2023) follow a similar way to address the trade-off between coherence and high-quality synthesis, where they cast audio synthesis as a language modeling task and leverage a hierarchy of coarse-to-fine audio discrete units in a discrete representation space.

Recently, Nguyen et al. (2023) leverage the success of discrete representation and introduce the first end-to-end generative spoken dialogue language model. However, due to the data and computational resource scarcity mentioned above, it would be challenging to train spoken generative language models from scratch that enables the processing of complex audio information. Differently, we regard LLMs (i.e., ChatGPT) as the general-purpose interface and leverage various audio foundational models to solve audio understanding and generation tasks, where AudioGPT is further connected with modality transformation to support speech conversations.

3 AudioGPT

3.1 System Formulation

As briefly discussed in Sec. 1, AudioGPT is a prompt-based system, defined as

$$\text{AudioGPT} = (\mathcal{T}, \mathcal{L}, \mathcal{M}, \mathcal{H}, \{\mathcal{P}_i\}_{i=1}^{P}), \tag{1}$$

where $\mathcal{T}$ is a modality transformer, $\mathcal{L}$ is a dialogue engine (i.e., large language model, LLM), $\mathcal{M}$ is a prompt manager, $\mathcal{H}$ is a task handler, and $\{\mathcal{P}_i\}_{i=1}^{P}$ is a set of P audio foundation models. Let a context with $(n-1)$-rounds interactions to be defined as $C = \{(q_1, r_1), (q_2, r_2), ..., (q_{n-1}, r_{n-1}))\}$, where q_i is the query of i^{th} round and r_i is the response of i^{th} round. Denoted a new query q_n, the execution of the AudioGPT is to generate the response r_n as formulated in:

$$r_n = \text{AudioGPT}(q_n, C) \tag{2}$$

During inference, AudioGPT can be decomposed into four major steps: 1) **Modality transformation**: transfer various input modalities within q_n into a query q'_n with a consistent modality; 2) **Task analysis**: utilize the dialogue engine $\mathcal{L}$ and the prompt manager $\mathcal{M}$ to parse (q'_n, C) into structure arguments a_n for the task handler $\mathcal{H}$; 3) **Model assignment**: the task handler $\mathcal{H}$ consumes structured arguments a_n and send the arguments to its corresponding audio task processor $\mathcal{P}_s$, where s is the selected task index, and 4) **response generation**: after execution of $\mathcal{P}_s(a_n)$, the final response r_n is generated through $\mathcal{L}$ by combining information from $(q'_n, C, \mathcal{P}_s(a_n))$.

3.2 Modality Transformation

As discussed in Sec. 3.1, the first stage aims to transform the query q_n into a new query q'_n in a consistent format. The user input query q_n includes two parts: a query description $q_n^{(d)}$ and a set of query-related resources of size k, $\{q_n^{(s_1)}, q_n^{(s_2)}, ..., q_n^{(s_k)}\}$. In AudioGPT, the query description $q_n^{(d)}$ can be either in textual or audio (i.e., speech) format. And the modality transformer $\mathcal{T}$ first checks the modality of query description $q_n^{(d)}$. If the query description $q_n^{(d)}$ is in audio, $\mathcal{T}$ is then responsible for converting $q_n^{(d)}$ in audio to textual modality as:

$$q'_n = (q_n'^{(d)}, \{q_n^{(s_1)}, ..., q_n^{(s_k)}\}) = \begin{cases} (q_n^{(d)}, \{q_n^{(s_1)}, ..., q_n^{(s_k)}\}) & \text{if } q_n^{(d)} \text{ is text,} \\ (\mathcal{T}(q_n^{(d)}), \{q_n^{(s_1)}, ..., q_n^{(s_k)}\}) & \text{if } q_n^{(d)} \text{ is audio.} \end{cases} \tag{3}$$

3.3 Task Analysis

As introduced in Sec. 3.1, the task analysis step focuses on extracting structured argument a_n from (q'_n, C). Specifically, the context C is fed into the dialogue engine $\mathcal{L}$ ahead of the argument extraction. Based on the types of query resources $\{q_n^{(s_1)}, q_n^{(s_2)}, ..., q_n^{(s_k)}\}$ from q'_n, the task handler $\mathcal{H}$ first classifies the query into different task families, which is classified through I/O modalities. Then, given the task family selected, the query description $q_n^{\prime(d)}$ is passed into the prompt manager $\mathcal{M}$ to generate argument a_n, including the selected audio foundation model $\mathcal{P}_p$ and its corresponding task-related arguments $h_{\mathcal{P}_p}$, where p is the index of the selected audio model from the audio model set $\{\mathcal{P}_i\}_{i=1}^{P}$.

$$(\mathcal{P}_p, h_{\mathcal{P}_p}) = \mathcal{L}(\mathcal{M}(\mathcal{H}(q'_n), q_n^{\prime(d)}), C), \tag{4}$$

where $\mathcal{H}(q'_n)$ is the task family selected by the task handler $\mathcal{H}$. Noted that, for an audio/image-input task family, $h_{\mathcal{P}_p}$ may also contain the necessary resources (e.g., audio or images) from the previous context C.

As aforementioned, the task family is determined through the task handler $\mathcal{H}$ by considering the I/O modality. To be specific, the families are:

- Audio-to-Text
  - Speech Recognition: Transcribe human speech

Table 1: Supported Tasks in AudioGPT

Task	Input	Output	Domain	Model
Speech Recognition	Audio	Text	Speech	Whisper (Radford et al., 2022)
Speech Translation	Audio	Text	Speech	MultiDecoder (Dalmia et al., 2021)
Style Transfer	Audio	Audio	Speech	GenerSpeech (Huang et al., 2022b)
Speech Enhancement	Audio	Audio	Speech	ConvTasNet (Luo & Mesgarani, 2019)
Speech Separation	Audio	Audio	Speech	TF-GridNet (Wang et al., 2022)
Mono-to-Binaural	Audio	Audio	Speech	NeuralWarp (Grabocka et al., 2018)
Audio Inpainting	Audio	Audio	Sound	Make-An-Audio (Huang et al., 2023a)
Sound Extraction	Audio	Audio	Sound	LASSNet (Liu et al., 2022b)
Sound Detection	Audio	Event	Sound	Pyramid Transformer (Xin et al., 2022)
Talking Head Synthesis	Audio	Video	Talking Head	GeneFace (Ye et al., 2023)
Text-to-Speech	Text	Audio	Speech	FastSpeech 2 (Ren et al., 2020)
Text-to-Audio	Text	Audio	Sound	Make-An-Audio (Huang et al., 2023a)
Audio-to-Text	Audio	Text	Sound	MAAC (Ye et al., 2021)
Image-to-Audio	Image	Audio	Sound	Make-An-Audio (Huang et al., 2023a)
Singing Synthesis	Musical Score	Audio	Music	DiffSinger (Liu et al., 2022a) VISinger (Zhang et al., 2022b)

  - Speech Translation: Translate human speech into another language
  - Audio Caption: Describe audio in text
- Audio-to-Audio
  - Style Transfer: Generate human speech with styles derived from a reference
  - Speech Enhancement: Improve the speech quality by reducing background noise
  - Speech Separation: Separate mix-speech of different speakers
  - Mono-to-Binaural: Generate binaural audio given mono one
  - Audio Impainting: Inpaint audio given user input mask
- Audio-to-Event
  - Sound Extraction: Selectly extract a part of audio based on description
  - Sound Detection: Predict the event timelines in audio
- Audio-to-Video
  - Talking Head Synthesis: Generate a talking human portrait video given input audio
- Text-to-Audio
  - Text-to-Speech: Generate human speech given user input text
  - Text-to-Audio: Generate general audio given user description
- Image-to-Audio
  - Image-to-Audio: Generate audio from image
- Score-to-Audio
  - Singing Synthesis: Generate singing voice given input text, note and duration Sequence

3.4 Model Assignment

Given the selected model $\mathcal{P}_p$ and its corresponding arguments $h_{\mathcal{P}_p}$, this step assigns the related resources to the model and executes the model $\mathcal{P}_p$ to get the task output $o_{\mathcal{P}_p}$:

$$o_{\mathcal{P}_p} = \mathcal{P}_p(\{q_n^{(s_1)}, q_n^{(s_2)}, ..., q_n^{(s_k)}\}, h_{\mathcal{P}_p}). \tag{5}$$

To keep the efficiency of AudioGPT, we conduct the audio model initialization during either environmental setups or server initialization.

3.5 Response Generation

The response generation is highly related to the select task $\mathcal{P}_p$ and its output $o_{\mathcal{P}_p}$. Specifically, for audio generation tasks, AudioGPT shows both the waveform in an image and the corresponding audio file for downloading/playing; for tasks that generate text, the model directly returns the transcribed text; for the video generation task, the output video and some related image frames are shown; for classification tasks, a posteriorgram of categories is shown over the time span.

4 Evaluating Multi-Modal LLMs

4.1 Overview

The rapid development of multi-modal LLMs (Wu et al., 2023; Shen et al., 2023; Huang et al., 2023b) has significantly increased the research demand for evaluating its performance and behavior in understanding human intention, performing complex reasoning, and organizing the cooperation of multiple audio foundation models.

In this section, we outline the design principles and process of evaluating multi-modal LLMs (i.e., AudioGPT). Specifically, we evaluate the LLMs in the following three aspects: 1) Consistency, which measures whether the LLMs properly understand the intention of a user, and assigns the audio foundation models closely aligned with human cognition and problem-solving; 2) Capability, which measures the performance of audio foundation models in handling complex audio tasks, understanding and generating speech, music, sound, and talking head in a zero-shot fashion; and 3) Robustness, which measures the ability of LLMs deals with special cases.

4.2 Consistency

Figure 2: A high-level overview of consistency evaluation.

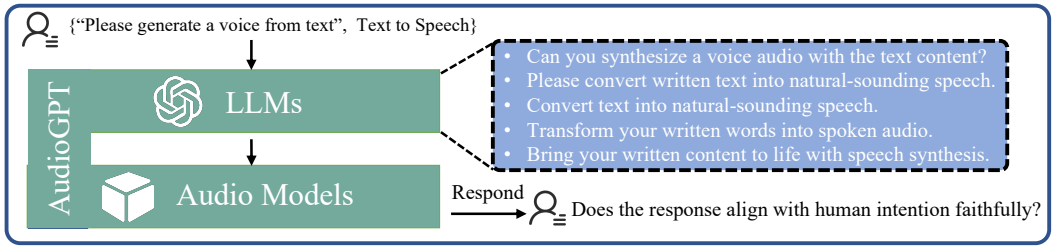


Table 2: Ratings that have been used in the evaluation of query-answer consistency.

Rating	Consistency	Definition
20	Completely inconsistent	Very annoying and objectionable inconsistency.
40	Mostly inconsistent	Annoying but not objectionable inconsistency.
60	somewhat consistent	Perceptible and slightly annoying inconsistency
80	Mostly consistent	Just perceptible but not annoying inconsistency.
100	Completely consistent	Imperceptible inconsistency

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.

More specifically, as shown in Figure 2, the consistency evaluation is carried out in three steps for each task in the benchmark. In the first step, we request human annotators to provide prompts for each task in a format of {prompts, task_name}. This allows us to evaluate the model's ability to comprehend complex tasks and identify the essential prompts needed for successful task assignments. In the second step, we leverage the outperformed language generation capacity of LLMs to produce descriptions with the same semantic meanings while having different expressions, enabling a comprehensive

evaluation of whether LLMs understands the intention of a broader amount of user. Finally, we use crowd-sourced human evaluation via Amazon Mechanical Turk, where AudioGPT is prompted with these natural language descriptions corresponding to a variety of tasks and intentions. Human raters are shown the response of multi-modal LLMs and a prompt input and asked "Does the response closely align with human cognition and intention faithfully?". They must respond with "completely", "mostly", or "somewhat" on a 20-100 Likert scale, which is documented with 95% confidence intervals (CI).

4.3 Capability

As the task executors for processing complex audio information, audio foundation models have a significant impact on handling complex downstream tasks. Taking AudioGPT as an example, we report evaluation matrics and downstream datasets for understanding and generating speech, music, sound, and talking head in Table 3.

Table 3: Evaluating audio foundation models in AudioGPT.

Task	Audio Model	Dataset	Metrics
Speech Recognition	Whisper	LibriTTS	WER, CER
Speech Translation	MultiDecoder	MUSTC	BLEU
Style Transfer	GenerSpeech	ESD	MCD, FFE, MOS
Speech Enhancement	ConvTasNet	CHiME4	SNR, PESQ, STOI
Speech Separation	TF-GridNet	WSJ0-2mix	SNR, PESQ, STOI
Mono-to-Binaural	NeuralWarp	BinauralDataset	L2 Error, PESQ, MRSTFT
Audio Inpainting	Make-An-Audio	AudioCaption	MOS
Sound Extraction	LASSNet	AudioCaption	SNR, PESQ
Sound Detection	Pyramid Transformer	AudioSet	mAP
Target Sound Detection	TSDNet	URBAN-SED	F-score
Talking Head Synthesis	GeneFace	LRS3-TED	FID, LMD
Text-to-Speech	Fastspeech2	LJSpeech	MCD, FFE, MOS
Text-to-Audio	Make-an-Audio	AudioCaption	FID, KL, CLAP, MOS
Audio-to-Text	MAAC	Clotho-v2	CIDEr-D
Image-to-Audio	Make-An-Audio	AudioCaption	MOS
Singing Synthesis	DiffSinger, VISinger	OpencPOP	MCD, FFE, MOS

4.4 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:

- Long chains of evaluation: Multi-modal LLMs are expected to handle long chains of evaluation while considering short and long context dependencies in multi-modal generation and reuse. A chain of tasks can be presented either as a query that requires sequential application of candidate audio models, as consecutive queries that ask for different tasks, or as a mixture of the two types.
- Unsupported tasks: Multi-modal LLMs should be able to provide reasonable feedback to queries that require unsupported tasks not covered by the foundation models.
- Error handling of multi-modal models: Multi-modal foundation models can fail due to different reasons, such as unsupported arguments or unsupported input formats. In such scenarios, multi-modal LLMs need to provide reasonable feedback to queries that explain the encountered issue and suggest potential solutions.
- Breaks in context: Multi-modal LLMs are expected to process queries that are not in a logical sequence. For instance, the user may submit random queries in a query sequence but continue to proceed with previous queries that have more tasks.

To evaluate the robustness, we conduct a three-step subjective user rating process, similar to the steps discussed in Sec.4.2. In the first step, human annotators provide prompts based on the above four

categories. In the second step, the prompts are fed into the LLM to formulate a complete interaction session. Finally, a different set of subjects recruited from multi-modal LLMs rate the interaction on the same 20-100 scale as described in Sec.4.2.

5 Experiments

5.1 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 (Chase, 2022). 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. The current manuscript mainly covers the system description, where the experiments are designed more for demonstration.

5.2 Case Study on Multiple Rounds Dialogue

Figure 3 shows a 12-rounds dialogue case of AudioGPT, which demonstrates the capabilities of AudioGPT 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.

5.3 Case Study on Simple Tasks

AudioGPT equips ChatGPT with audio foundation models, where ChatGPT is regarded as the general-purpose interface to solve numerous audio understanding and generation tasks. We test AudioGPT on a wide range of audio tasks in generating and understanding speech, music, sound, and talking head, where some cases are illustrated in Figure 4 and 5.

6 Limitation

Although AudioGPT excels at solving complex audio-related AI tasks, limitations could be observed in this system as follows: 1) **Prompt Engineering**: AudioGPT uses ChatGPT to connect a large number of foundation models, and thus it requires prompt engineering to describe audio foundation models in natural language, which could be time-consuming and expertise-required; 2) **Length Limitation**: the maximum token length in ChatGPT may limit the multi-turn dialogue, which also influences the user's context instruction, and 3) **Capabliity Limitation** AudioGPT relies heavily on audio foundation models to process audio information, which is heavily influenced by the accuracy and effectiveness of these models.

7 Conclusion

In this work, we presented AudioGPT, which connected ChatGPT with 1) audio foundation models to handle challenging audio tasks, and 2) a modality transformation interface to enable spoken dialogue. By combining the advantages of ChatGPT and audio-modality solvers, AudioGPT presented strong capacities in processing audio information in the following four stages: modality transformation, task analysis, model assignment, and response generation. To assess the ability of multi-modal LLMs in human intention understanding and cooperation with foundation models, we outlined the design principles and processes, and evaluated AudioGPT in terms of consistency, capability, and robustness. Experimental results demonstrated the outperformed abilities of AudioGPT in solving AI tasks with speech, music, sound, and talking head understanding and generation in multi-round dialogues, empowering humans to create rich and diverse audio content with unprecedented ease. The current manuscript mainly covers the system description, where the experiments are designed more for demonstration.

Figure 3: Qualitative analysis on multiple rounds of dialogue between humans and AudioGPT.

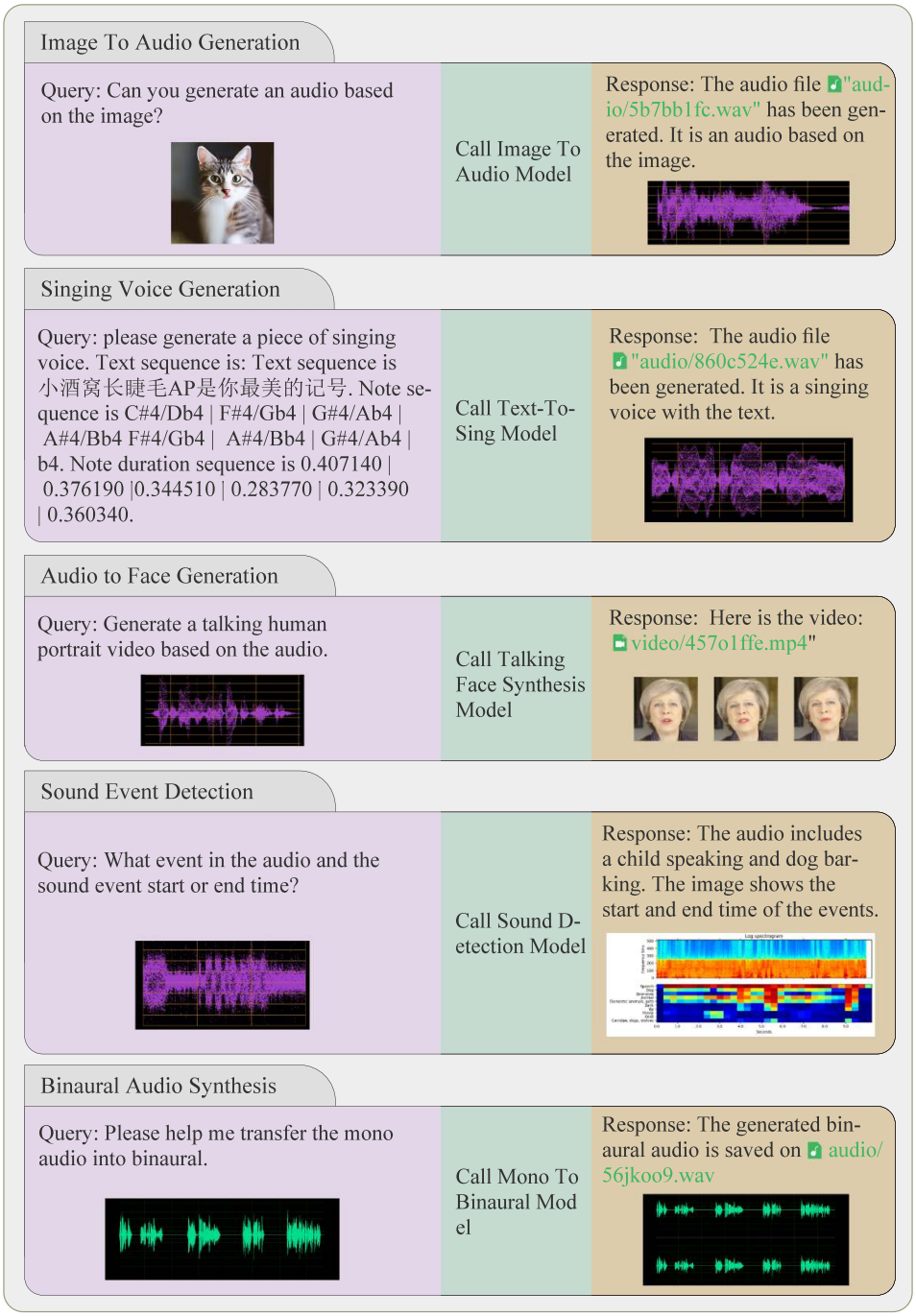


Figure 4: Qualitative analysis on simple tasks.

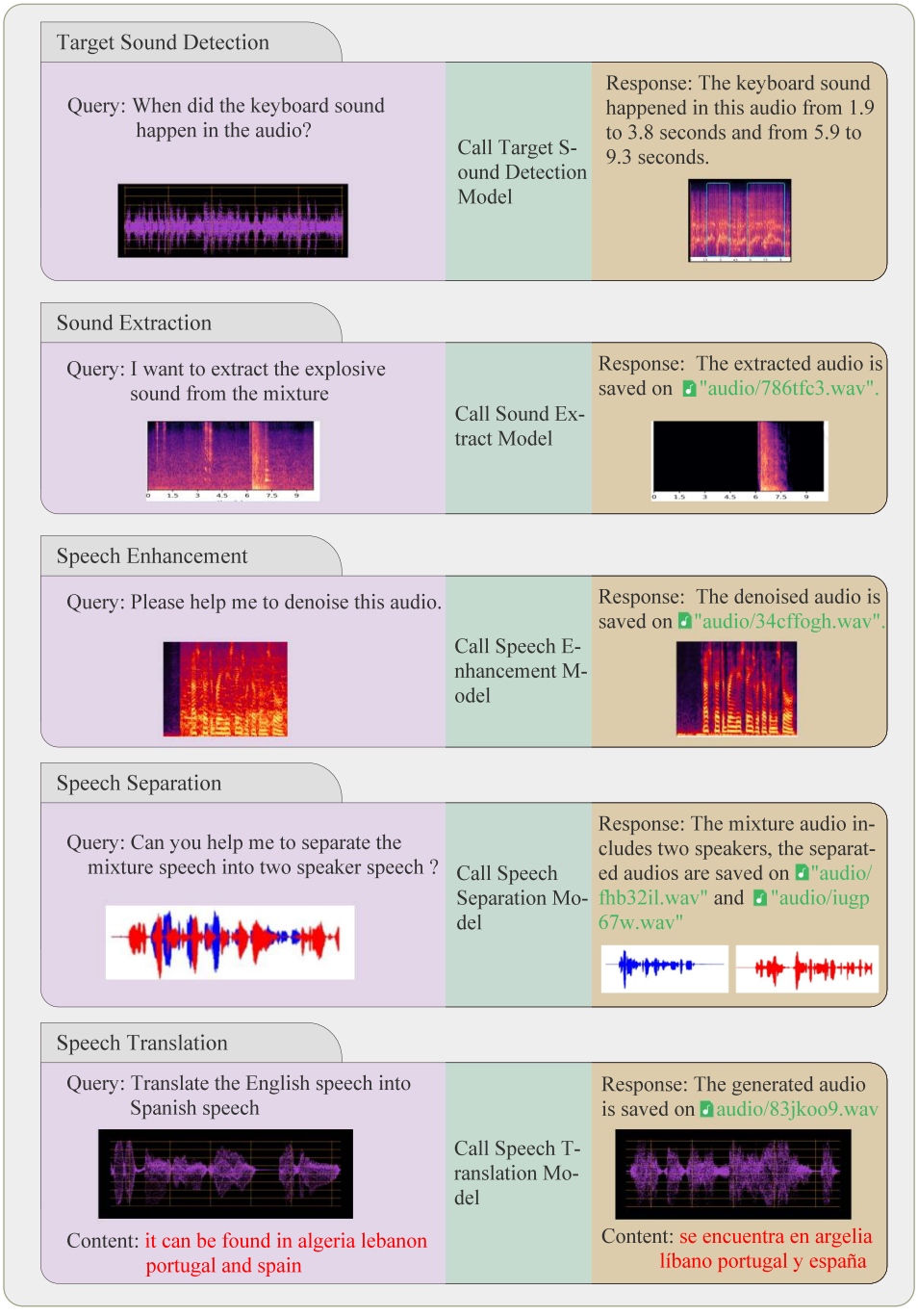


Figure 5: Qualitative analysis on simple tasks.

References

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Baziotis, C., Haddow, B., and Birch, A. Language model prior for low-resource neural machine translation. *arXiv preprint arXiv:2004.14928*, 2020.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

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

Chase, H. LangChain, 10 2022. URL `https://github.com/hwchase17/langchain`.

Dalmia, S., Yan, B., Raunak, V., Metze, F., and Watanabe, S. 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*, pp. 1882–1896, 2021.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

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

Grabocka, J., Schmidt Thieme, L., and etc. Neuralwarp: Time-series similarity with warping networks. *arXiv preprint arXiv:1812.08306*, 2018.

Gulcehre, C., Firat, O., Xu, K., Cho, K., and Bengio, Y. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148, 2017.

Hayashi, T. and Watanabe, S. Discretalk: Text-to-speech as a machine translation problem. *arXiv preprint arXiv:2005.05525*, 2020.

Hosseini-Asl, E., McCann, B., Wu, C.-S., Yavuz, S., and Socher, R. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191, 2020.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

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

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

Huang, R., Ren, Y., Liu, J., Cui, C., and Zhao, Z. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis. *arXiv preprint arXiv:2205.07211*, 2022b.

Huang, R., Zhao, Z., Liu, J., Liu, H., Ren, Y., Zhang, L., and He, J. Transpeech: Speech-to-speech translation with bilateral perturbation. *arXiv preprint arXiv:2205.12523*, 2022c.

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

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Liu, Q., et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023b.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kharitonov, E., Copet, J., Lakhotia, K., Nguyen, T. A., Tomasello, P., Lee, A., Elkahky, A., Hsu, W.-N., Mohamed, A., Dupoux, E., et al. textless-lib: A library for textless spoken language processing. *arXiv preprint arXiv:2202.07359*, 2022.

Liu, F., Li, G., Zhao, Y., and Jin, Z. Multi-task learning based pre-trained language model for code completion. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, pp. 473–485, 2020.

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

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

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

Nguyen, T. A., Kharitonov, E., Copet, J., Adi, Y., Hsu, W.-N., Elkahky, A., Tomasello, P., Algayres, R., Sagot, B., Mohamed, A., et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266, 2023.

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

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

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

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

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

Svyatkovskiy, A., Zhao, Y., Fu, S., and Sundaresan, N. Pythia: Ai-assisted code completion system. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2727–2735, 2019.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

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

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

Xin, Y., Yang, D., and Zou, Y. Audio pyramid transformer with domain adaption for weakly supervised sound event detection and audio classification. *Proc. Interspeech 2022*, pp. 1546–1550, 2022.

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

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

Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., and Zhao, Z. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507, 2021.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.

Zhang, Y., Cong, J., Xue, H., Xie, L., Zhu, P., and Bi, M. 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)*, pp. 7237–7241. IEEE, 2022b.