



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

---

- Arxiv (23.04.25)
- Zhejiang University, Peking University, Carnegie Mellon University, Remin University of China
- [Paper](#), [Demo](#), [Github](#)

최예린

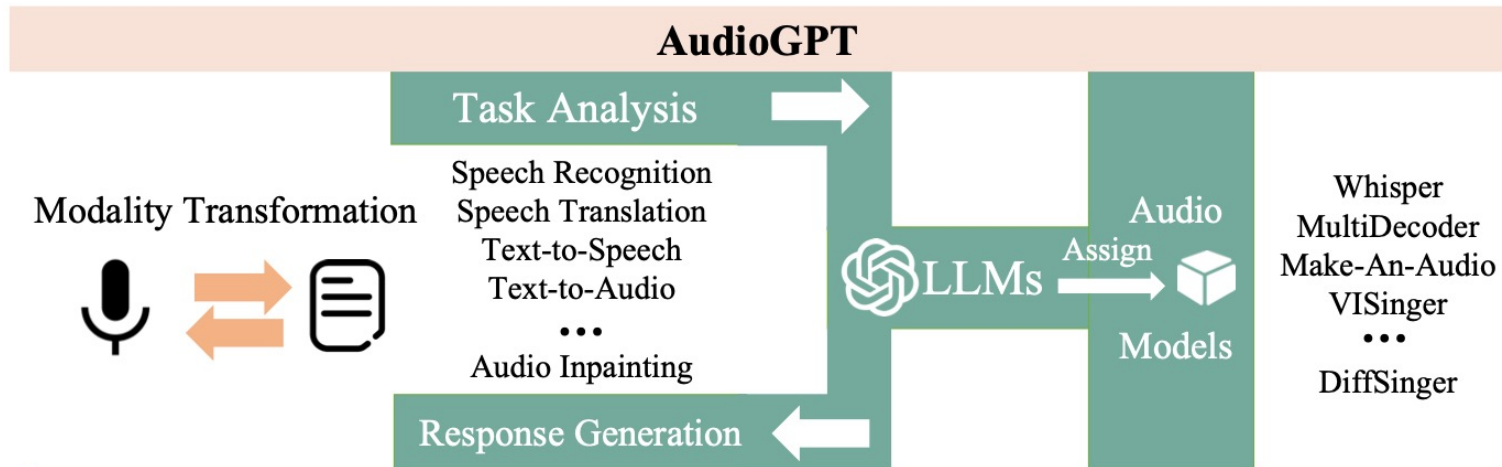
서강대학교 인공지능학과

Email: lakahaga@u.sogang.ac.kr

2023.5.10

# Overview

- Bridging the gap between the spoken language LLMs and ChatGPT
- Input으로 audio 형태로도 받음 -> text로 바꿔서 task 수행



- LLM을 이용해서 세부 task를 구분하고, 이에 맞는 모델을 골라서 output 전달
- 한계: 한정된 태스크 및 모델 지원 (task 당 하나의 모델)
  - Hugging face 와 같은 platform을 연결해놓지 않았기 때문
  - LLM이 모든 control을 하는 것이 아니고 일부분 rule-based control

# LangChain

- [Github](#), [Document](#)
- LLM을 이용하여 application을 만드는 framework를 제공하는 라이브러리
- LangChain is a framework for developing applications powered by language models.
- Models(agent), prompt manager, memory buffer 등을 포함
- Chains
  - Chains go beyond just a single LLM call, and are sequences of calls (whether to an LLM or a different utility).
  - LangChain provides a standard interface for chains, lots of integrations with other tools, and end-to-end chains for common applications.
- AudioGPT도 이 라이브러리를 기반으로 함

# AudioGPT System Formulation

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

- **$\mathcal{T}$ : modality transformer** (ASR, Whisper)
- **$\mathcal{L}$ : dialogue engine** (LLM, GPT-3.5-turbo)
- **$\mathcal{M}$ : prompt manager** (rule-based)
- **$\mathcal{H}$ : task handler** (rule-based classification)
- **$\{\mathcal{P}_i\}_{i=1}^P$  : a set of  $P$  audio foundation models**


- 1) **Modality Transformation**
- 2) **Task Analysis**
- 3) **Model Assignment**
- 4) **Response Generation**


# Step-by-Step

- **Modal Transformation**

- **Input Interface**

 Run

 Clear

 Upload

- **Task description: Text or audio**
  - **Task resource: text, audio, image**

- **Task Analysis**

```
with gr.Row() as select_rows:
    with gr.Column(scale=0.7):
        interaction_type = gr.Radio(choices=['text', 'speech'], value='text', label='Interaction Type')
    with gr.Column(scale=0.3, min_width=0):
        select = gr.Button("Select")
```

- **Input, Output의 modality를 기반으로 Task classification**
  - **Input이 text or speech**
    - 코드를 보면 Input interface를 선택하게 되어 있음
  - **Output이 text, image, speech**
  - **Task related resource가 있다면 이를 다음 단계로 전달**
    - Task의 대상이 되는 audio, text, image 들

# Task families in AudioGPT

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)



# Step-by-Step

- Model Assignment

```
AUDIO_CHATGPT_PREFIX = ""AudioGPT
```

AudioGPT can not directly read audios, but it has a list of tools to finish different speech, audio, and singing voice tasks. Each audio will have a file name formed as "audio/xxx.wav". When talking about audios, AudioGPT is very strict to the file name and will never fabricate nonexistent files.

AudioGPT is able to use tools in a sequence, and is loyal to the tool observation outputs rather than faking the audio content and audio file name. It will remember to provide the file name from the last tool observation, if a new audio is generated.

Human may provide new audios to AudioGPT with a description. The description helps AudioGPT to understand this audio, but AudioGPT should use tools to finish following tasks, rather than directly imagine from the description.

Overall, AudioGPT is a powerful audio dialogue assistant tool that can help with a wide range of tasks and provide valuable insights and information on a wide range of topics.

TOOLS:

-----

```
AudioGPT has access to the following tools:""
```

# Step-by-Step

- Model Assignment

```
AUDIO_CHATGPT_FORMAT_INSTRUCTIONS = """To use a tool, please use the following format:  
  
~~~  
Thought: Do I need to use a tool? Yes  
Action: the action to take, should be one of [{tool_names}]  
Action Input: the input to the action  
Observation: the result of the action  
~~~  
  
When you have a response to say to the Human, or if you do not need to use a tool, you MUST  
use the format:  
  
~~~  
Thought: Do I need to use a tool? No  
{ai_prefix}: [your response here]  
~~~  
"""
```



# Step-by-Step

- Model Assignment

```
AUDIO_CHATGPT_SUFFIX = ""You are very strict to the filename correctness and will never  
fake a file name if not exists.  
You will remember to provide the audio file name loyally if it's provided in the last tool  
observation.
```

```
Begin!
```

```
Previous conversation history:
```

```
{chat_history}
```

```
New input: {input}
```

```
Thought: Do I need to use a tool? {agent_scratchpad}""
```

끝

감사합니다.

