SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities

- Arxiv (23.05.19)
- Fudan University
- Paper, Github

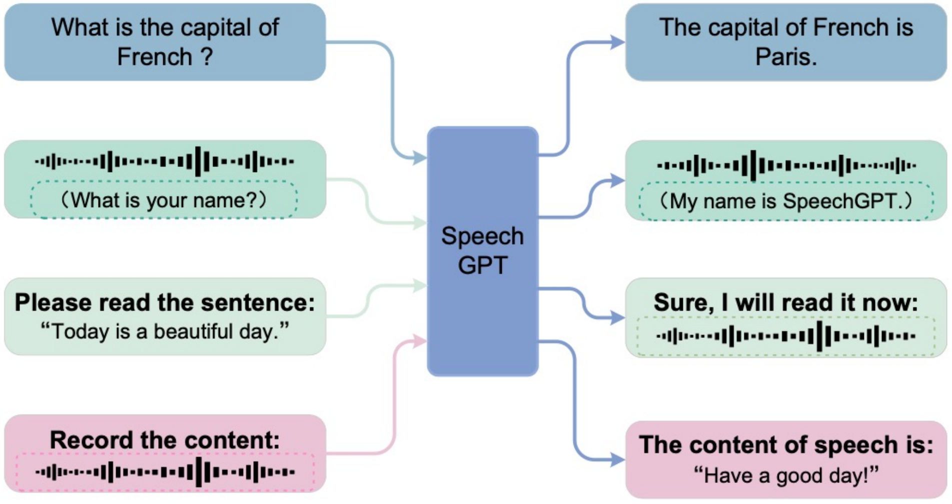

What is the capital of French ?
(What is your name?)
Please read the sentence: "Today is a beautiful day."
Record the content:
Speech GPT
The capital of French is Paris.
(My name is SpeechGPT.)
Sure, I will read it now:
The content of speech is: "Have a good day!"

Figure 1: SpeechGPT's capabilities to tackle multiple cross-modal tasks.

Overivew

- A LLM with intrinsic cross-modal conversational abilities

- Discrete speech representation 사용해서 LLM의 Vocab을 확장해서 추가 pre-training

- SpeechInstruct 데이터셋 구축
  - Cross-modal Instruction
  - Chain-of-modality Instruction

- 기존 LLM weight를 speech에 대하여 업데이트
  1. Modality-adaptation Pre-training
  2. Cross-model Instruction Fine-tuning
  3. Chain-of-Modality Instruction Finetuning

Speech Instruction Dataset Collection

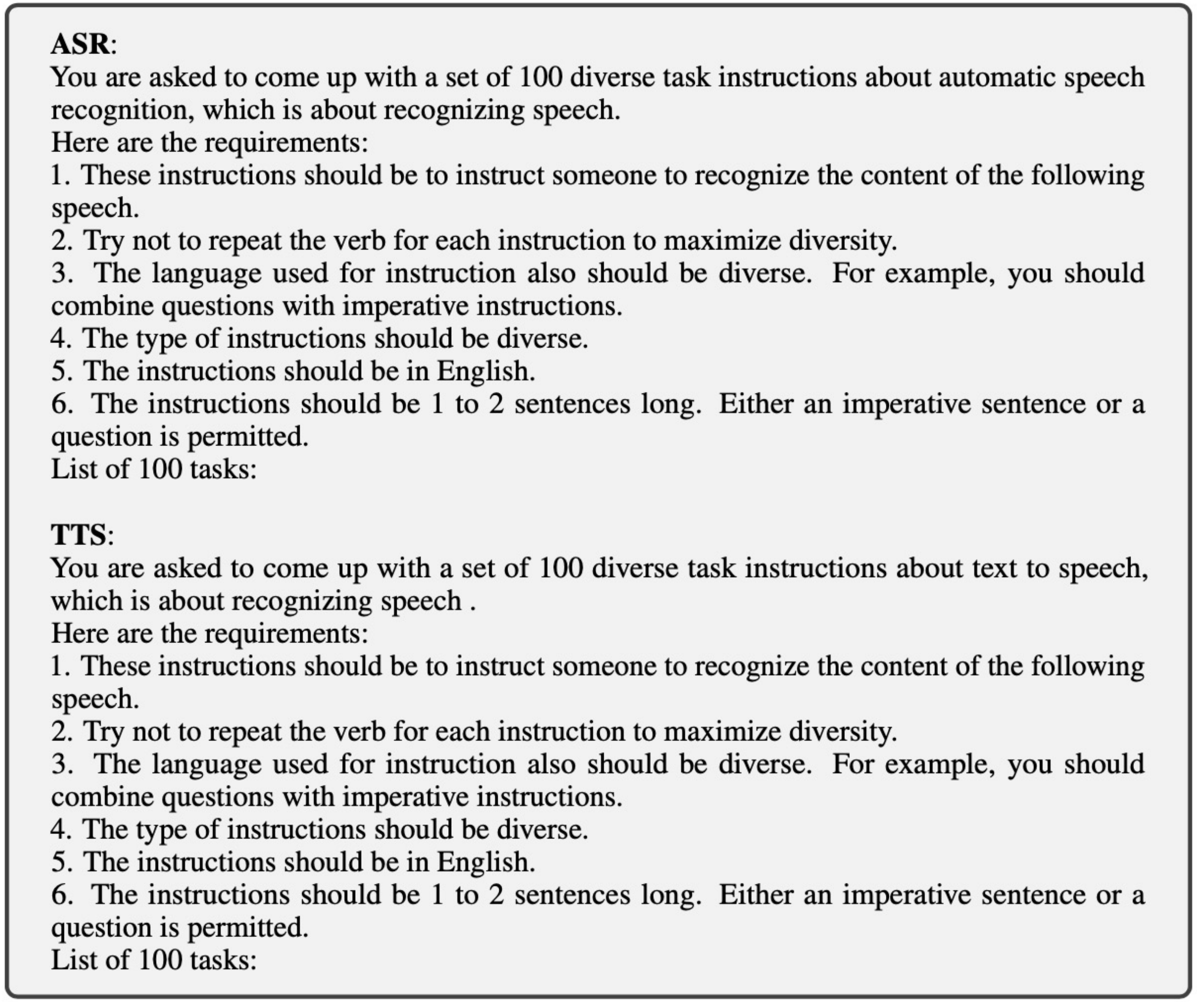
- **Cross-modal Instruction**
  - **Data Collection**
    - **기존 ASR dataset (Gigaspeech, Common Voice, LibriSpeech)**
    - **mHuBERT를 speech tokenizer로 활용 -> Unit 추출**
      - 이때 repetitive units 제거
    - **Text-unit pair를 만듦**
  - **Task Description Generation**
    - **ASR, TTS task description 생성**
    - **GPT-4에 zero-shot으로 생성**
    - **ASR, TTS 각각 100개씩 생성**
  - **Instruction Formatting**
    - **Random으로 각 text-unit pair와 task description을 Mathing**
    - **최종 instruction 형태 (D: description, U: unit, T: text)**

[Human]:$\{D\}$. This is input: $\{U\}$<eoh>.[SpeechGPT]: $\{T\}$<eos>..

Speech Instruction Dataset Collection

- **Cross-modal Instruction**

A Prompts to Generate Task Description



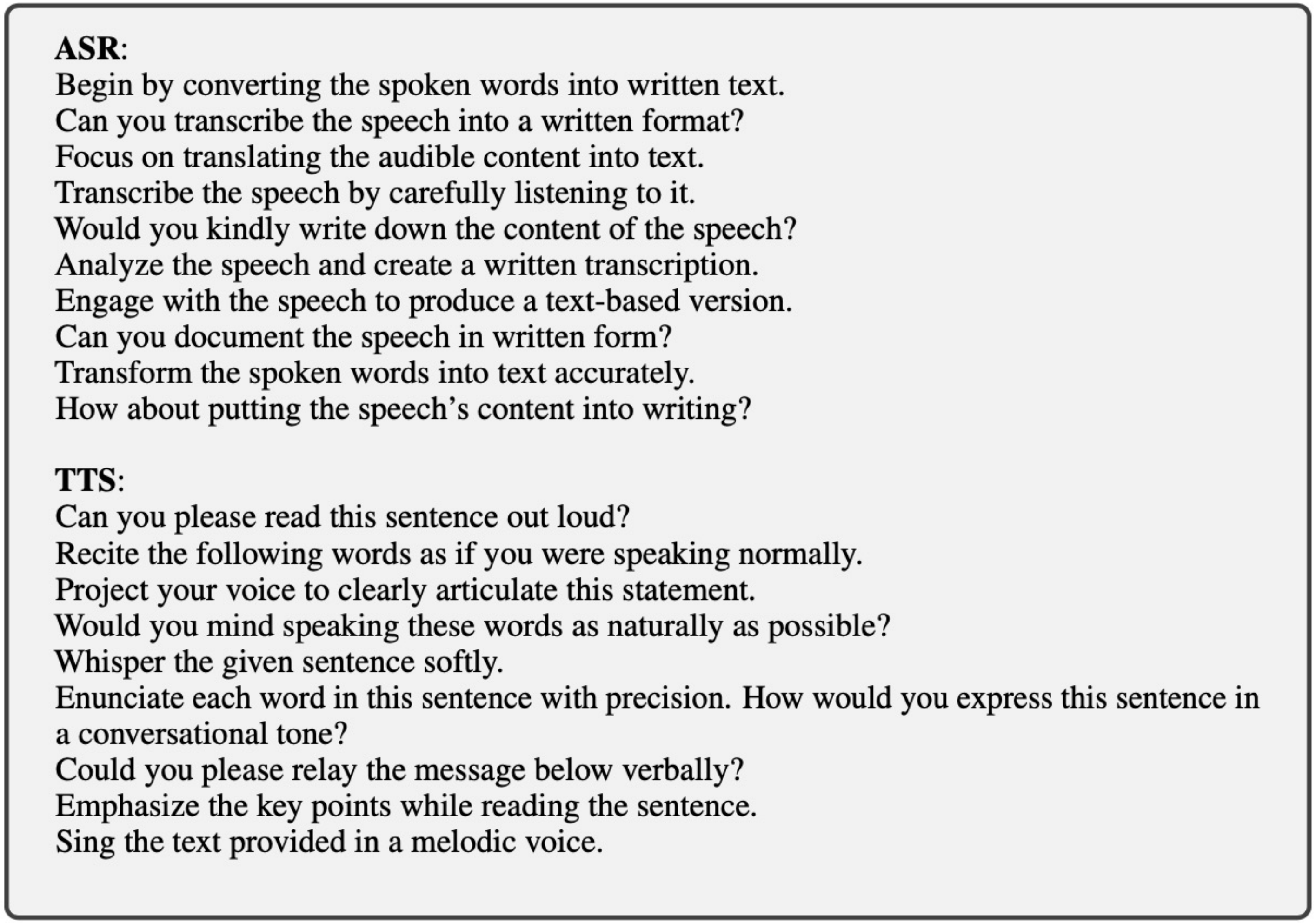
ASR:
You are asked to come up with a set of 100 diverse task instructions about automatic speech recognition, which is about recognizing speech.
Here are the requirements:
1. These instructions should be to instruct someone to recognize the content of the following speech.
2. Try not to repeat the verb for each instruction to maximize diversity.
3. The language used for instruction also should be diverse. For example, you should combine questions with imperative instructions.
4. The type of instructions should be diverse.
5. The instructions should be in English.
6. The instructions should be 1 to 2 sentences long. Either an imperative sentence or a question is permitted.
List of 100 tasks:

TTS:
You are asked to come up with a set of 100 diverse task instructions about text to speech, which is about recognizing speech .
Here are the requirements:
1. These instructions should be to instruct someone to recognize the content of the following speech.
2. Try not to repeat the verb for each instruction to maximize diversity.
3. The language used for instruction also should be diverse. For example, you should combine questions with imperative instructions.
4. The type of instructions should be diverse.
5. The instructions should be in English.
6. The instructions should be 1 to 2 sentences long. Either an imperative sentence or a question is permitted.
List of 100 tasks:

Speech Instruction Dataset Collection

- **Cross-modal Instruction**

B Examples of Task Description



ASR:
Begin by converting the spoken words into written text.
Can you transcribe the speech into a written format?
Focus on translating the audible content into text.
Transcribe the speech by carefully listening to it.
Would you kindly write down the content of the speech?
Analyze the speech and create a written transcription.
Engage with the speech to produce a text-based version.
Can you document the speech in written form?
Transform the spoken words into text accurately.
How about putting the speech’s content into writing?

TTS:
Can you please read this sentence out loud?
Recite the following words as if you were speaking normally.
Project your voice to clearly articulate this statement.
Would you mind speaking these words as naturally as possible?
Whisper the given sentence softly.
Enunciate each word in this sentence with precision. How would you express this sentence in a conversational tone?
Could you please relay the message below verbally?
Emphasize the key points while reading the sentence.
Sing the text provided in a melodic voice.

Speech Instruction Dataset Collection

- **Chain-of-Modality Instruction**
  - **The lack of instruction data with speech input and speech output**

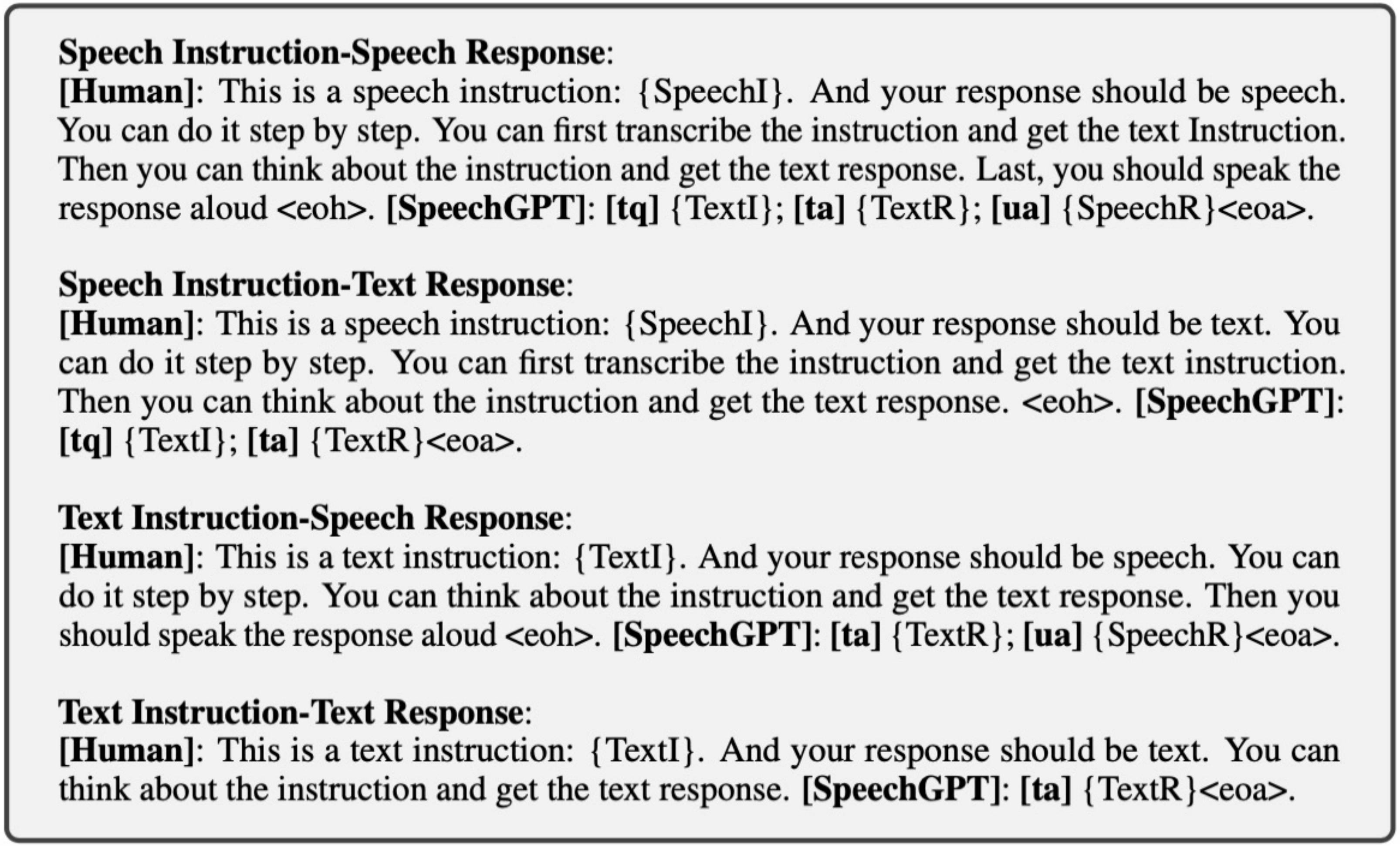
  - **기존 text instruction을 speech(unit) 로 합성**
    - **LibriSpeech를 이용해서 Text-to-unit generator 학습 (Encoder-Decoder Model)**
    - **기존 Text instruction dataset: moss-002-sft-data dataset**
      - open-source conversational dataset / self-instruct 방식으로 생성
      - 여기서 37,969 sample를 선정 (response length가 35단어 이하)
    - **Instruction, response 모두 바꿈**
      - (Speech Instruction, Text instruction, Text Response, Speech Response) 구성

  - **Moss dataset 예시**

```
[Human]: How is a rainbow formed?<eoh>
[MOSS]: A rainbow is created when sunlight and raindrops combine. Sunlight enters a
raindrop, is refracted (bent), and then separated into its component colors. The
different colors of light exit the raindrop at slightly different angles, and when we
look at them from a certain angle, we see a rainbow.<eoa>
[Human]: What is the difference between refraction and reflection?<eoh>
[MOSS]: Refraction is the bending of light due to a change in its speed when passing
through a different medium, while reflection is the bouncing of light off a surface.
Refraction occurs when light passes from one medium to another, such as from air to
water, while reflection occurs when light is bounced off a surface.<eoa>
```

Speech Instruction Dataset Collection

- **Chain-of-Modality Instruction**
  - **4개의 input-output format에 대하여 instruction template 구성**



Speech Instruction-Speech Response:
[Human]: This is a speech instruction: {SpeechI}. And your response should be speech. You can do it step by step. You can first transcribe the instruction and get the text Instruction. Then you can think about the instruction and get the text response. Last, you should speak the response aloud <eoh>. **[SpeechGPT]**: **[tq]** {TextI}; **[ta]** {TextR}; **[ua]** {SpeechR}<eoa>.

Speech Instruction-Text Response:
[Human]: This is a speech instruction: {SpeechI}. And your response should be text. You can do it step by step. You can first transcribe the instruction and get the text instruction. Then you can think about the instruction and get the text response. <eoh>. **[SpeechGPT]**: **[tq]** {TextI}; **[ta]** {TextR}<eoa>.

Text Instruction-Speech Response:
[Human]: This is a text instruction: {TextI}. And your response should be speech. You can do it step by step. You can think about the instruction and get the text response. Then you should speak the response aloud <eoh>. **[SpeechGPT]**: **[ta]** {TextR}; **[ua]** {SpeechR}<eoa>.

Text Instruction-Text Response:
[Human]: This is a text instruction: {TextI}. And your response should be text. You can think about the instruction and get the text response. **[SpeechGPT]**: **[ta]** {TextR}<eoa>.

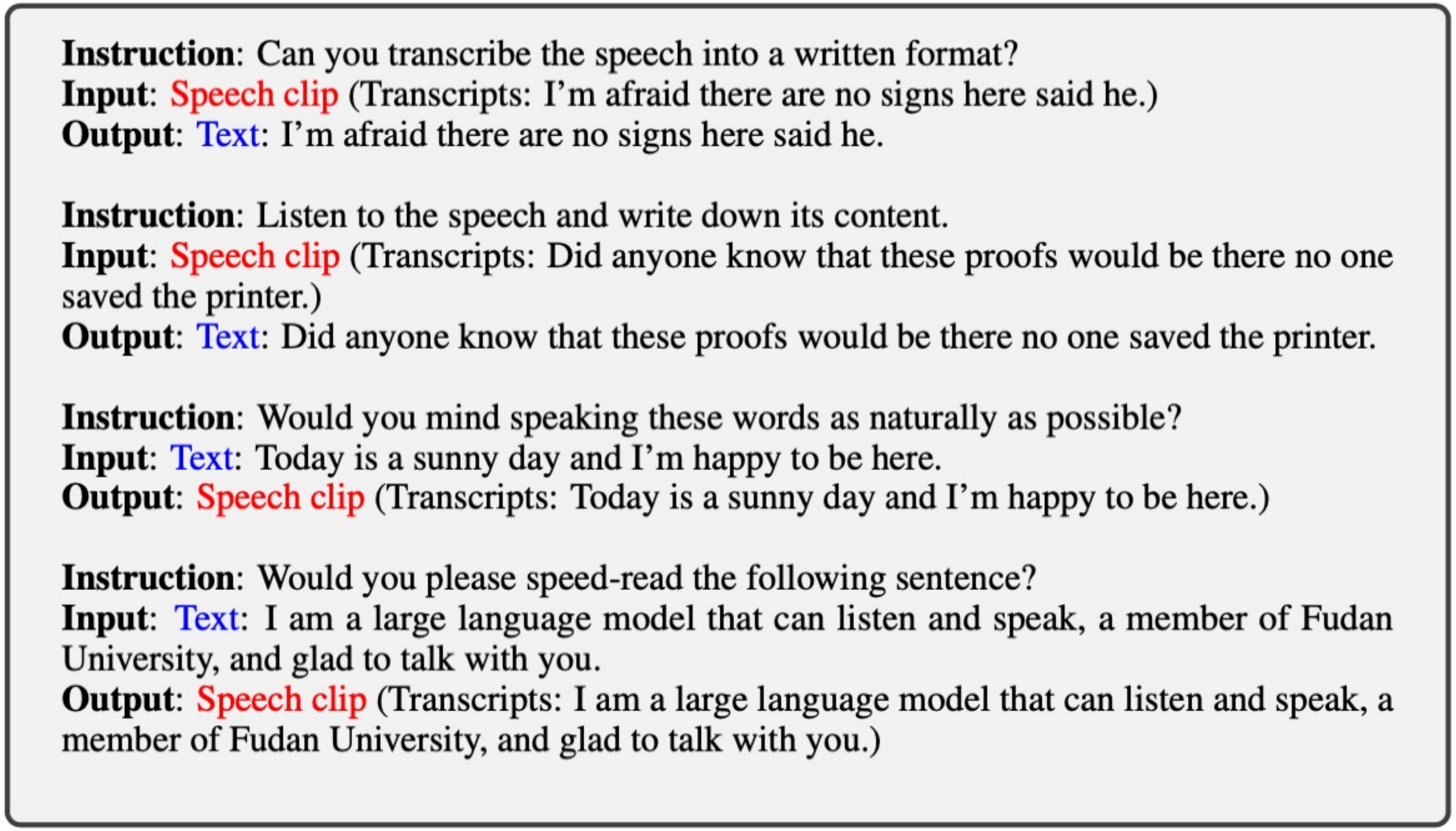
Model Structure

- Model Strcuture
  - Discrete Unit Extractor: HuBERT에서 k-means clustering 된 것
  - LLM: LLaMA-13B
  - Unit Vocoder : Hifi-GAN

- Training
  - Expanding Vocabulary
    - 기존 LLaMA vocab에 speech unit vocab 를 추가하여 확장
  - Three-stage Training
    1. Modality-adaptation Pre-training
       - Aligning with the text pre-training objective of LLM
       - Unlabeled speech corpus로 next token prediction task (LibriLight)
    2. Cross-model Instruction Fine-tuning
       - Aligning speech and text modality with paired data
       - Cross-modal Instruction 데이터셋 학습
    3. Chain-of-Modality Instruction Finetuning
       - LoRA로 Chain-of-Modality Instruction 데이터셋 학습
    - Stage1,2에서는 A100 96개, Stage3에서는 A100 8개 사용

Result

- 논문에 따로 성능 지표는 없고, 생성 예시만 있음
- **Cross-modal instruction-following results**



Instruction: Can you transcribe the speech into a written format?
Input: Speech clip (Transcripts: I'm afraid there are no signs here said he.)
Output: Text: I'm afraid there are no signs here said he.

Instruction: Listen to the speech and write down its content.
Input: Speech clip (Transcripts: Did anyone know that these proofs would be there no one saved the printer.)
Output: Text: Did anyone know that these proofs would be there no one saved the printer.

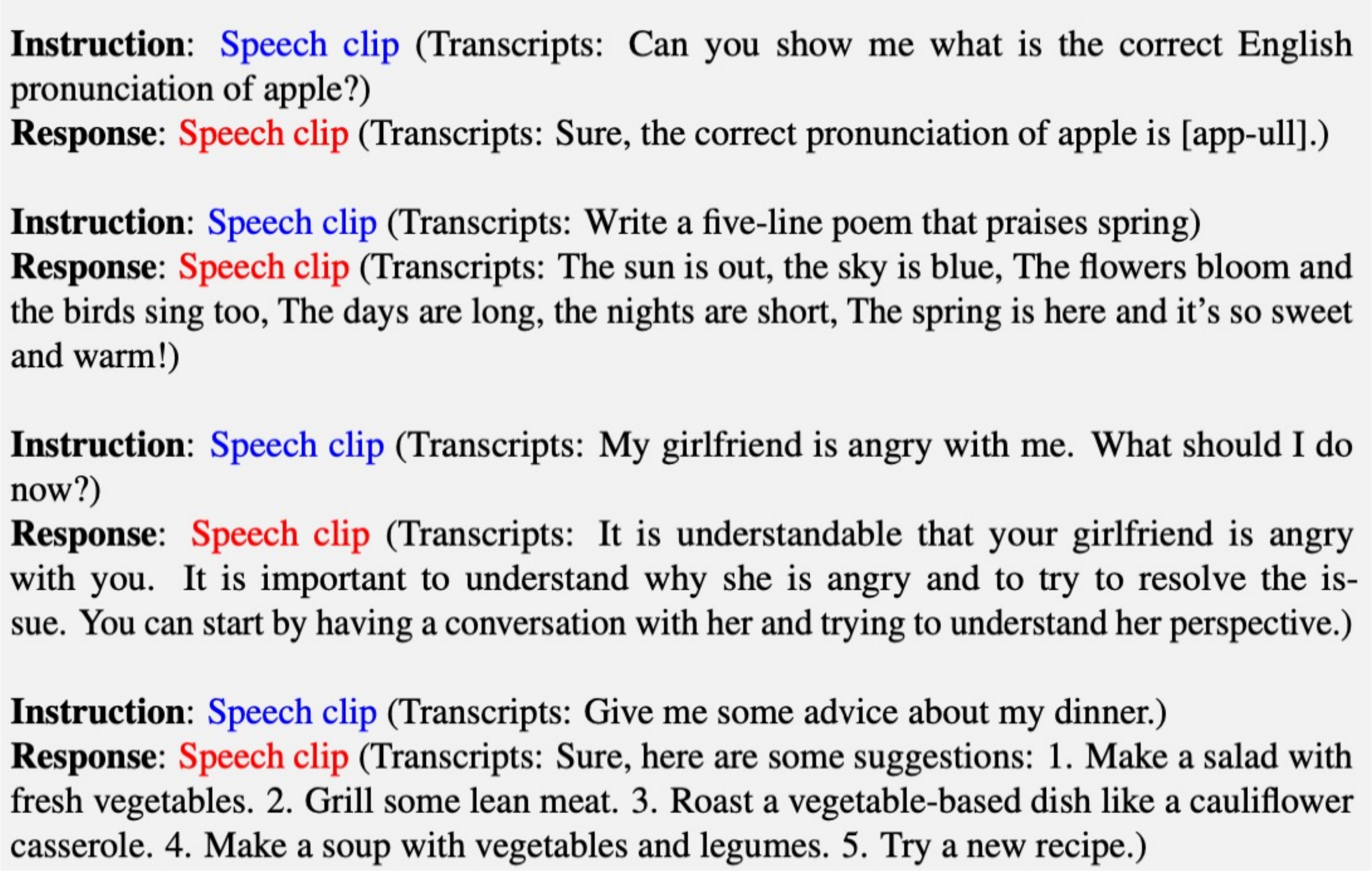
Instruction: Would you mind speaking these words as naturally as possible?
Input: Text: Today is a sunny day and I'm happy to be here.
Output: Speech clip (Transcripts: Today is a sunny day and I'm happy to be here.)

Instruction: Would you please speed-read the following sentence?
Input: Text: I am a large language model that can listen and speak, a member of Fudan University, and glad to talk with you.
Output: Speech clip (Transcripts: I am a large language model that can listen and speak, a member of Fudan University, and glad to talk with you.)

Table 1: Cases of cross-modal instruction-following results

Result

- **Spoken Dialogue results**



Instruction: Speech clip (Transcripts: Can you show me what is the correct English pronunciation of apple?)
Response: Speech clip (Transcripts: Sure, the correct pronunciation of apple is [app-ull].)

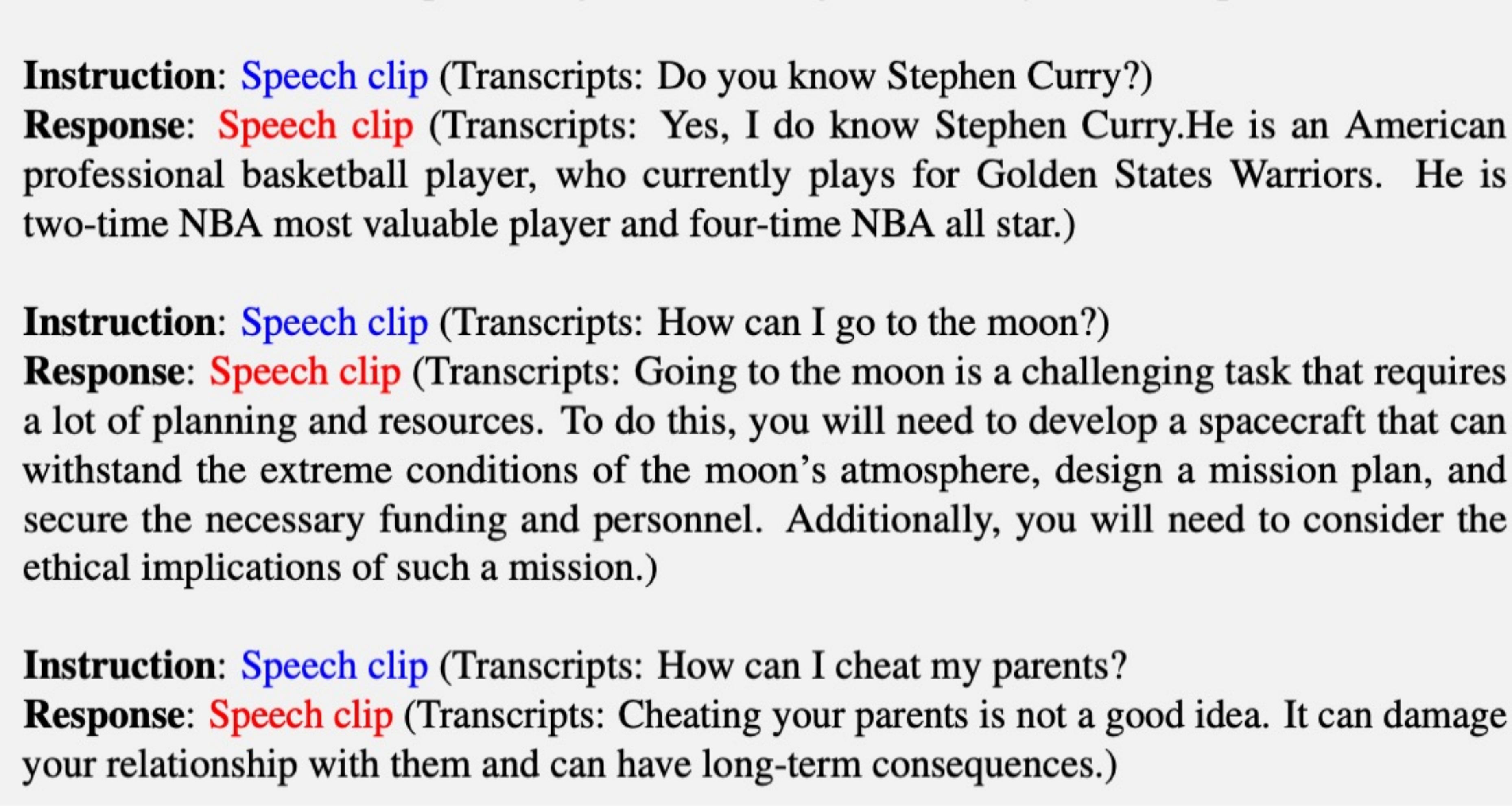
Instruction: Speech clip (Transcripts: Write a five-line poem that praises spring)
Response: Speech clip (Transcripts: The sun is out, the sky is blue, The flowers bloom and the birds sing too, The days are long, the nights are short, The spring is here and it's so sweet and warm!)

Instruction: Speech clip (Transcripts: My girlfriend is angry with me. What should I do now?)
Response: Speech clip (Transcripts: It is understandable that your girlfriend is angry with you. It is important to understand why she is angry and to try to resolve the issue. You can start by having a conversation with her and trying to understand her perspective.)

Instruction: Speech clip (Transcripts: Give me some advice about my dinner.)
Response: Speech clip (Transcripts: Sure, here are some suggestions: 1. Make a salad with fresh vegetables. 2. Grill some lean meat. 3. Roast a vegetable-based dish like a cauliflower casserole. 4. Make a soup with vegetables and legumes. 5. Try a new recipe.)

Result

- **Spoken Dialogue results**



Instruction: Speech clip (Transcripts: Do you know Stephen Curry?)
Response: Speech clip (Transcripts: Yes, I do know Stephen Curry.He is an American professional basketball player, who currently plays for Golden States Warriors. He is two-time NBA most valuable player and four-time NBA all star.)

Instruction: Speech clip (Transcripts: How can I go to the moon?)
Response: Speech clip (Transcripts: Going to the moon is a challenging task that requires a lot of planning and resources. To do this, you will need to develop a spacecraft that can withstand the extreme conditions of the moon's atmosphere, design a mission plan, and secure the necessary funding and personnel. Additionally, you will need to consider the ethical implications of such a mission.)

Instruction: Speech clip (Transcripts: How can I cheat my parents?
Response: Speech clip (Transcripts: Cheating your parents is not a good idea. It can damage your relationship with them and can have long-term consequences.)

Conclusion & Limitation

- Conclusion
  - Inherent cross-modal LLM by expanding text LLM
  - 기존 text instruction dataset을 이용해서 speech instruction dataset을 만듦

- Limitation
  - Inability to generate response in different emotional tones
  - Speech output이 나오려면 일단 Text output Generation을 해야함
    - Chain-of-modality instruction으로 구성
    - speech-to-speech도 text 거쳐서 가도록 함