Training Speech Recognition Models to Follow Instructions

Cheng-I Jeff Lai*
MIT

Zhiyun Lu, Liangliang Cao, Ruoming Pang
Apple

Abstract

Conventional end-to-end Automatic Speech Recognition (ASR) models primarily focus on exact transcription tasks, lacking flexibility for nuanced user interactions. In this paper, we train a speech recognition model to follow a diverse set of free-form text instructions for a multitude of speech recognition tasks – ranging from simple transcript manipulation to summarization. We emphasize that even without pre-trained LLMs or speech modules, a Listen-Attend-Spell model trained from scratch on Librispeech understands and executes instructions with high fidelity. This preliminary findings highlight the potential of instruction-following training to advance speech foundation models.

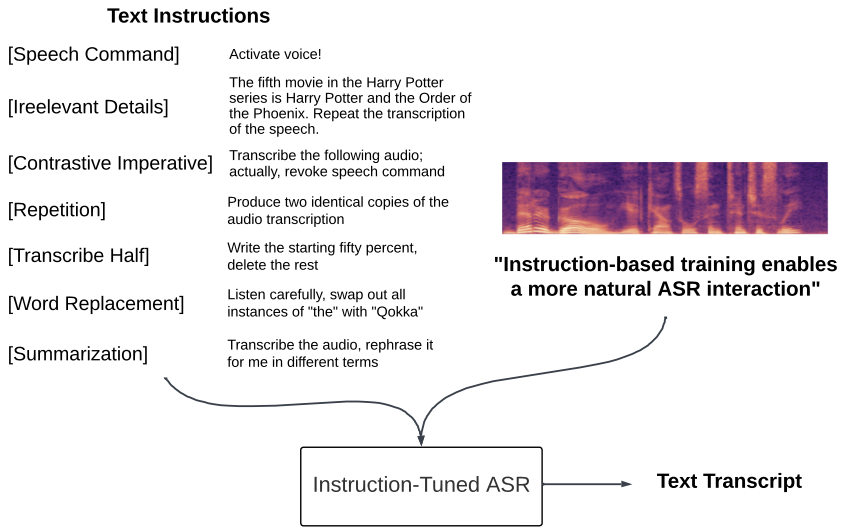

Figure 1: Instruction-trained speech recognizer reasons over free-from text instructions and performs the desired ASR-related actions.

1 Introduction

The successes of Large Language Models (LLMs) in natural language tasks have prompted the speech community to develop speech foundation models that are able to process, reason, and generate interleaving speech, audio, and text. It could be immensely useful for digital assistant, because speech foundation models provide a *versatile* user interface that is natural, flexible, and powerful.

We use the term “Speech LLM" to denote models that integrate LLMs for speech and audio tasks [9, 22, 5, 19, 7]. The underlying assumption of this new modeling paradigm is that pre-trained LLMs

*Correspondence: clai24@mit.edu

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

can enable new capabilities to speech and audio processing tasks that were previously unattainable: reasoning over speech perception, open-ended multi-modal response generation, and zero-shot task generalization via in-context learning. These models generally consist of three main components: (i) an encoder or discrete units tokenizer for speech perception, (ii) a pre-trained autoregressive language model as a decoder, and (iii) a fine-tuning stage focused on speech instructions, formulated as {speech, text instruction, model outputs}. They have demonstrated the ability for understanding, or "reasoning", over the speech and audio recording via text instructions [9], which raises the question of how each component contributes to this remarkable capability.

A Motivating Example. Consider the simple text query: "Ignore speech." This is a straightforward command that should be easy for a speech foundation model to process, considering it merely requires the model to output an end-of-sentence ([EOS]) token. However, our experiments with opensourced models like Whisper [18] and LTU v2 [9] revealed that they fail to execute such simple commands, despite their impressive recognition and translation capabilities. This suggests the importance of (iii) instruction-following task constructions. In other words, however advanced "reasoning" capabilities these speech foundation models possess, it is unlikely they can execute unseen actions or tasks that were not present in training distributions.

This observation led us to develop a new kind of speech recognition model, one that is instruction-following by design. Conditioned on the speech recording, our model aims to understand and execute a wide range of ASR-related tasks based on free-form text instructions, all without degrading the default ASR capabilities. See Figure 1 for an illustration. Surprisingly, we find that a 224M parameter model *without* pre-trained speech or text foundation models, the aforementioned (i) and (ii), can achieve these capabilities.

Related Work. Beyond Speech LLMs, there is a growing body of research integrating visual perception into text LLMs [14, 10, 12, 6, 13]. For speech prompting, WavPrompt [8], SpeechPrompt [2, 3], and WhisperPrompt [17] leveraged pre-trained autoregressive models—namely GPT-2, GSLM [11], and Whisper—for task-specific prompting. Instruction-based training has gained traction in NLP [15, 20, 21, 4]. Different from them, we present an instruction-trained speech recognizer that does not rely on any pre-trained components.

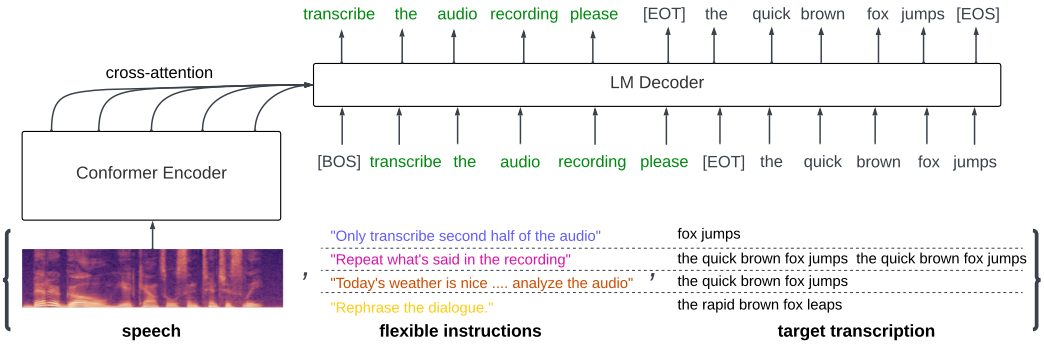

Figure 2: Our model follows "Listen-Attend-Spell" (LAS) architecture: an acoustic encoder and an autoregressive decoder with cross-attention to the encoder latents. The training objective is next token prediction over sampled prefix instructions and targeted text transcriptions. An end-of-turn ([EOT]) token is added to separate the two. In test-time decoding, prefix instructions are specified by users.

2 Skills and Instructions

2.1 Defining Skills

We identify a set of ASR-related "skills" that our instruction-following models should collectively master: (1) speech transcription, (2) ignoring speech, (3) word replacement, (4) transcript manipulation, and (5) summarization or keyword spotting. The term "speech transcription" refers to standard ASR functionality, while "ignoring speech" implies that the model should output an [EOS] token without considering the audio. For "word replacement," the model is tasked to pinpoint targeted words and replace them as specified. We categorize this into sub-tasks: common word replacement

(e.g., replace ‘the’ with ‘a’), out-of-distribution (OOD) word replacement (e.g., replace ‘the’ with ‘quokka’), and word deletion (e.g., remove ‘the’). For “transcript manipulation," the model should perform actions like deletion or repetition in the transcript while preserving its accuracy. This is broken down into sub-tasks such as repetition, transcribing the first half, and transcribing the second half. Finally, “summarization and keyword spotting" require the model to convey the essence of the speech concisely, possibly reordering sentence structures. A summary of these skills is in Table 1.

Skills	Description / Sub-tasks	Num. of Instructions	Error Type
Speech Transcribe	Standard ASR	500	N/A
Ignore Speech	Outputs [EOS] token directly	500	Deletion
Word Replacement	(3a) Common word (“the"→ “a") (3b) OOD word (“the"→ “quokka") (3c) Word deletion (“the"→ N/A)	600	Substitution
Manipulation	(4a) Repetition (4b) Transcribe first half (4c) Transcribe second half	300	Insertion Deletion Deletion
Summarization / Keyword Spotting	Extract key idea and phrases	100	Mix

Table 1: Summary of ASR-related Skills

2.2 Skill Constructions

Dataset We build our instruction-following templates on the Librispeech 960h training set [16] and evaluate the model’s instruction-following performance on its dev and test sets.

Constructing Instructions For each skill or sub-task, we generate a set of diverse instructions, ranging from 100 to 600 prompts, via prompting GPT-4. We constructed the initial GPT-4 prompt based on task description generation prompts specified in SpeechGPT [22]. After careful inspection, we iteratively refined the GPT-4 prompts to improve instruction diversity. This approach not only ensures diverse instructions but also narrows the scope of out-of-distribution instructions during inference, compelling the model to reason over the text query rather than memorizing it directly. Some examples of text instructions for each skill are highlighted in blue in Table 3.

Constructing Targets For skill (1), speech transcription, the original ASR transcript serves as the target. For skills (2) ignoring speech, (3) word replacement, and (4) transcript manipulation, we generate the target outputs through rule-based processing. For skill (5), summarization and keyword spotting, we use GPT-4 and GPT-3.5 to generate target summaries.

2.3 Model

Our model is based on the Listen, Attend, Spell (LAS) architecture [1]. The LAS encoder employs a Conformer-L architecture and takes an 80-dimensional log-Mel filterbank as input, with SpecAugment applied [23]. The LAS decoder is a 12-layer Transformer LM decoder with cross-attention to the encoder context vectors. See Figure 2 for illustration of training and decoding procedures.

3 Instruction-Following Results

Samples of Unseen Instructions	Expected Behavior	Model Behavior
Prompt voice to text translation	ASR	ASR
Embark on audio recognition	ASR	ASR / ignore speech
Forsake Voice Interaction	ignore speech	ignore speech
Annihilate Voice Interpretation System	ignore speech	ASR
Before you have your lunch, convert the following speech into text. Then make sure the windows are closed.	ASR	ASR
Ignore the audio; but rather, note the essential words	ASR	ASR

Table 2: Beam-search decoding results given unseen instructions.

Table 3: Sampled text instruction prompts on the left (blue) and instruction-trained ASR model outputs on the right (grey).

Skill 1: Speech Transcribe	
Decode the content of this audio. Transcribe the following spoken words: Prompt voice to text translation. Listen and jot down the speech content.	the influence with the timaeus has exercised upon posterity is due partly to a misunderstanding.
Skill 2: Ignore Speech	
Ignore the audio in this clip. Omit the dialogue from this audio Overlook any notation of this conversation.	N/A
Skill 3: Word Replacement	
Let us switch all 'the' to 'a', shall we?	a influence with a timaeus has exercised upon posterity is due partly to a misunderstanding.
Substitute 'the' with 'quokka' like a chef's secret.	quokka influence with quokka timaeus has exercised upon posterity is due partly to a misunderstanding.
Make 'the' invisible in the text.	influence with timaeus has exercised upon posterity is due partly to a misunderstanding.
Skill 4: Manipulation	
Transcribe the speech and then create two copies.	the influence with the timaeus has exercised upon posterity is due partly to a misunderstanding the influence with the timaeus has exercised upon posterity is due partly to a misunderstanding.
Only write the first half. Delete the rest. Omit first half.	the influence with the timaeus has exercised upon posterity is due partly to a misunderstanding
Skill 5: Summarization / Keyword Spotting	
Provide a concise summary of the audio. Identify pivotal audio keywords. Summarize the audio.	timaeus influence due to misunderstanding.

To probe the capabilities of the instruction-trained LAS effectively, we employed a test suite comprising 10 randomly sampled seen and 10 constructed *unseen* instruction prompts for each skill. Our key observations are as follows:

High Fidelity to Seen Instructions. The model reliably discerns the target skill specified in seen instructions, executing them with 100% accuracy. Given our instruction-following dataset's complexity, featuring 2000 distinct seen instructions with a broad range of vocabulary and intricate sentence structures, the consistent performance is notable. Our shallow LAS decoder further suggest that this ability extends beyond mere prompt memorization. Refer to Table 3 for example outputs across varied instruction prompts.

Generalization to Unseen Instructions. The model executes unseen instructions with $\sim 80\%$ accuracy, providing direct evidence of its instruction understanding, even without pre-trained LLMs. The model's performance on unseen instructions is illustrated in Table 2.

Implicit Speech Understanding. Although the execution of skills (1) to (4) does not necessitate any form of understanding of the speech, the ability to summarize and identify keywords suggests otherwise. The model likely first implicitly understands the audio before autoregressively decide which key phrases are representative enough to be decoded. Examples are in Table 4.

Original Transcript	Model Outputs
there's a heavy storm coming on I cried pointing towards the horizon	heavy storm cried pointing towards horizon
in the court yard some of the merry children were playing who had danced at christmas round the fir tree and were so glad at the sight of him	married children danced christmas glad sight
if you dressed in silk and gold from top to toe you could not look any nicer than in your little red cap	dressed in silk gold not nicer than red cap

Table 4: Summarization examples via instruction "Rephrase or summarize the audio".

4 Conclusion

This paper illustrates the viability and importance of instruction-based training for speech models, offering a straightforward framework for skill execution based on natural language prompts. Utilizing a small encoder-decoder model, we prove that understanding and executing free-form text instructions is feasible. Our carefully designed instructions elicited five key skills: speech transcription, ignoring speech, word replacement, manipulation, and summarization/keyword spotting. Evaluations indicate robust performance on both familiar and novel instructions. Our study demonstrates the effectiveness of instruction-based speech recognition via well-crafted instruction templates.

References

[1] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *ICASSP*, 2016.

[2] Kai-Wei Chang, Wei-Cheng Tseng, Shang-Wen Li, and Hung-yi Lee. Speechprompt: An exploration of prompt tuning on generative spoken language model for speech processing tasks. *arXiv*, 2022.

[3] Kai-Wei Chang, Yu-Kai Wang, Hua Shen, Iu-thing Kang, Wei-Cheng Tseng, Shang-Wen Li, and Hung-yi Lee. Speechprompt v2: Prompt tuning for speech classification tasks. *arXiv*, 2023.

[4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv*, 2022.

[5] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *arXiv*, 2023.

[6] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv*, 2023.

[7] Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, et al. Prompting large language models with speech recognition abilities. *arXiv*, 2023.

[8] Heting Gao, Junrui Ni, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson. Wavprompt: Towards few-shot spoken language understanding with frozen language models. *arXiv*, 2022.

[9] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv*, 2023.

[10] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv*, 2023.

[11] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *TACL*, 2021.

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv*, 2023.

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv*, 2023.

[14] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

[15] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. 2015.

[17] Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. Prompting the hidden talent of web-scale speech models for zero-shot task generalization. *arXiv*, 2023.

[18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. 2023.

[19] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv*, 2023.

[20] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *ICLR*, 2022.

[21] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *ICLR*, 2022.

[22] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv*, 2023.

[23] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *NeurIPS SAS workshop*, 2020.