Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Human-AI Interaction (Sakriani Sakti (Professor))		
Student ID	2411806	Submission date	2025 / 1 / 21
Name	平野 雄太		
Thesis title	Advancing End-to-End Multi-Talker ASR by Conditioning the Decoder on Diarized Speaker Information ダイアライズされた話者情報でデコーダーを条件付けることによるEnd-to-End複数話者音声認識の改善		

Abstract

Since the development of the first speech recognition systems over 50 years ago, speech recognition technology has steadily advanced. Recent progress in deep learning has led to a significant improvement in the accuracy of speech recognition systems, with some claiming that these systems now exceed human capabilities. However, much of this progress has focused on the recognition of single-talker speech, and the recognition of multi-talker speech, which includes overlapping speech segments, remains a significant challenge.

There are two main approaches to multi-talker ASR: pipeline approach that consist of multiple modules such as speech separation, single-speaker ASR, and speaker diarization, and End-to-End (E2E) approach composed of a single neural network module. The pipeline approach offers an intuitive way to tackle the complex task of multi-talker speech recognition by solving each subtask sequentially. This method has been widely adopted since before the recent advances in deep learning and remains the mainstream approach for multi-talker speech recognition today.

This thesis addresses the challenges of multi-talker speech recognition by focusing on both of the above approaches. This study first focuses on developing a fast and accurate pipeline system. In order to achieve both high speed and high accuracy in the speech recognition module, this study combines feature extraction with the self-supervised learning model WavLM and a fast Zipformer transducer-based speech recognition model with in-module downsampling. For training data, several data augmentation techniques such as reverberation simulation and guided source separation (GSS) are applied. The proposed pipeline system was submitted to the CHiME-8 challenge for the meeting transcription task (NOTSOFAR-1), and it achieved up to a 6x speedup in inference compared to the baseline system. The proposed system also achieved 4th place (out of 10 entries) in the single-channel track and 3rd place (out of 6 entries) in the multi-channel track, demonstrating its effectiveness in terms of speech recognition accuracy.

While the proposed pipeline system achieved excellent results in the CHiME-8 challenge, the development process highlighted several limitations of the pipeline approach. First, it requires significant development effort due to the need to develop multiple models from multiple research areas. Second, the complete independence of each module makes overall optimization difficult. These drawbacks can be addressed by E2E approaches, which are implemented using a single neural network. Therefore, this study next focuses on improving the speech recognition accuracy of E2E models with overlapping speech. Through analysis of the baseline E2E model, it turned out that the decoder plays a crucial role in speaker separation. this work therefore proposes a method to improve the recognition accuracy of overlapping speech by conditioning the decoder with auxiliary information about "who spoke when." In addition, the speaker diarization task simultaneously is solved in the E2E model to obtain the "who spoke when" information without relying on external modules. The experimental results confirmed that the proposed method improves the speech recognition accuracy over the baseline model, with only a slight increase in the number of parameters.