論文内容の要旨

申請者氏名　　胡　尤佳

Speech translation (ST), which automates the translation of spoken language, is a crucial technology that bridges linguistic barriers by converting spoken language in real-time. This dissertation addresses for achieving low latency and high robustness, especially in scenarios involving disfluencies, and spontaneous speech. Firstly, this study focuses on a novel multi-task end-to-end ST incorporating automatic speech recognition (ASR) posterior distribution-based loss for improving robustness against ASR ambiguities. Experiments demonstrate the improvements over baseline methods, showing robustness of the disfluent inputs. Secondly, this study proposes a effective multi-stage fine-tuning methods for training disfluent-to-fluent speech translation models by integrating augmented disfluency-tagged data. Experimental results show that the approach effectively identifies and removes disfluencies, leading to more fluent and accurate translations in spontaneous speech. Thirdly, this study proposes a fine-tuning approach using both offline and simultaneous speech translation data to tackle the problem of small amount of simultaneous interpretation (SI) data. This method achieves a balance between latency and translation quality, providing practical solutions for real-time applications. These contributions provide practical solutions to address the key challenges techniques ST for low latency and high robustness in real-world applications. The methods proposed in this thesis represent a significant step forward in delivering accurate and high-quality speech translation ensuring both robustness and low latency.

論文審査結果の要旨

申請者氏名　　胡　尤佳

This dissertation aims to achieve low-latency and robust speech translation, particularly for disfluent and spontaneous speech. First, it introduces a multi-task end-to-end speech translation (ST) model with an ASR posterior distribution-based loss, enhancing robustness against ASR ambiguities and improving performance on disfluent inputs. Second, a multi-stage fine-tuning approach is proposed, using augmented disfluency-tagged data to train disfluent-to-fluent translation models. This method effectively removes disfluencies, resulting in more fluent and accurate translations. Third, a fine-tuning strategy combining offline and simultaneous speech translation data is introduced to address the scarcity of SI data. This approach balances latency and translation quality, making it practical for real-time applications. The findings of this research have been published in two journals, The Journal of Natural Language Processing and IEICE Transaction, along with nine papers presented at international conferences. These contributions fulfill the requirements for a Ph.D. degree.