論 文 内 容 の 要 旨

博士論文題目　Direct End-to-end Speech Translation for Distant Language Pairs
（遠い言語間のための End-to-end 音声翻訳の提案）

氏　　名　　叶　高朋

（論文内容の要旨）

Directly translating spoken utterances from a source language to a target language is a challenging task as it requires a fundamental transformation in both linguistic and para/non-linguistic features. The traditional speech-to-speech translation approaches concatenated automatic speech recognition (ASR), text-to-text machine translation (MT), and text-to-speech synthesizer (TTS) via text information. The traditional speech translation performance is worse than that of the MT because the translation results are affected by the ASR errors. The end-to-end speech translation system has a potential to recover ASR errors and achieves higher performance than that of the cascade speech translation system. Recent state-of-the-art models for ASR, MT, and TTS have mainly been built using Deep Neural Networks (DNN), in particular, an encoder-decoder model with an attention mechanism. Several works have attempted to construct an end-to-end direct speech translation using DNN. However, the usefulness of these models has only been investigated on language pairs of similar syntax and word order (e.g., English-French or English-Spanish). For syntactically distant language pairs (e.g., English-Japanese), the speech translation requires distant word reordering. This thesis addresses how to build a speech translation system for syntactically distant language pairs that suffer from long-distance reordering. I focus mainly on English (subject-verb-object (SVO) word order) and Japanese ((SOV) word order) language pairs. First, I propose a neural speech translation without requiring significant changes in the cascade ASR and MT structure. Specially, I construct a neural network model that passes the ASR all candidate scores to the MT part. The MT part could consider the

ASR hypothesis in the translation process. Therefore the MT model can learn how to recover the ASR error during translation. I demonstrate how the acoustic information helps to recover the ASR error and improves the translation quality in the proposed model. Next, I propose the first attempt to build an end-to-end speech translation system for syntactically distant language pairs that suffer from long-distance reordering. To guide the encoder-decoder attentional model for this challenging problem, I construct an end-to-end speech-to-text translation module with transcoder and utilize Curriculum Learning (CL) strategies that gradually train the network for the end-to-end speech translation tasks by adapting the decoder or encoder parts. I then focus on the text-to-speech translation tasks and apply speech information to the target text decoding process. My proposed approach shows the speech information helps target text generation, and the generated results are much closer to the reference sentence. Finally, I propose a complete end-to-end speech-to-speech translation system and compare the performance with that of the state of the art system. The experiment results show that the proposed approaches provide significant improvements in comparison with the conventional end-to-end speech translation models.

氏　名	叶　高朋

（論文審査結果の要旨）

Directly translating spoken utterances from a source language to a target language is a challenging task as it requires a fundamental transformation in both linguistic and para/non-linguistic features. The traditional speech-to-speech translation approaches concatenated automatic speech recognition (ASR), text-to-text machine translation (MT), and text-to-speech synthesizer (TTS) via text information. The traditional speech translation performance is worse than that of the MT because the translation results are affected by ASR errors. The end-to-end speech translation system has the potential to recover ASR errors and achieves higher performance than that of the cascade speech translation system. Several works have attempted to construct an end-to-end direct speech translation using DNN to tackle these problems. However, the usefulness of these models has only been investigated on language pairs of similar syntax and word order (e.g., English-French or English-Spanish). This thesis first proposes a neural speech translation without requiring significant changes in the cascade ASR and MT structure. The proposed model passes the ASR all candidate scores to the MT part. Next, the thesis proposes a first attempt to build an end-to-end speech translation system for syntactically distant language pairs by a transcoder module and Curriculum Learning (CL) strategies. The thesis then focuses on text-to-speech translation tasks and apply speech information to the target text decoding process. Finally, the thesis proposes a complete end-to-end speech-to-speech translation system and compare the performance with that of the state of the art system.

The research proposed solutions to the problems which haven't been solved and a series of his research resulted in two journal papers, four peer-reviewed international conference papers, and some domestic conference papers. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.