論 文 内 容 の 要 旨

博士論文題目　　Machine Speech Chain
（機械によるスピーチチェインモデルに関する研究）

氏　　名　　Andros Tjandra

（論文内容の要旨）

Despite the close relationship between speech perception and production, research in automatic speech recognition (ASR) and text-to-speech synthesis (TTS) has progressed more or less independently without exerting much mutual influence. In human communication, on the other hand, a closed-loop speech chain mechanism with auditory feedback from the speaker's mouth to her ear is crucial. We take a step further and develop a closed-loop machine speech chain model based on deep learning. The sequence-to-sequence model in closed-loop architecture allows us to train our model on the concatenation of both labeled and unlabeled data. While ASR transcribes the unlabeled speech features, TTS attempts to reconstruct the original speech waveform based on the text from ASR. In the opposite direction, ASR also attempts to reconstruct the original text transcription given the synthesized speech. To the best of our knowledge, this is the first deep learning framework that integrates human speech perception and production behaviors. Our experimental results show that the proposed approach significantly improved performance over that from separate systems that were only trained with labeled data. In this thesis, first I present a study about end-to-end speech modeling in general and followed by their application for ASR and TTS. Later, the basic of machine speech chain is described in detail in Chapter 3. Next, we integrate speech chain with speaker embedding model in Chapter 4 to achieve multi-speaker speech chain and improve the ASR and TTS performance on multi-speaker dataset settings. In Chapter 5, we identify the issue where the output of ASR is discrete variables, therefore we proposed a way to fully backpropagate the loss from TTS to the ASR model by using the straight-through estimator. In Chapter 6, we propose an alternative ASR training with reinforcement learning to solve the discrepancy between training and inference stage.

氏　名	Andros Tjandra

（論文審査結果の要旨）

Despite the close relationship between speech perception and production, research in automatic speech recognition (ASR) and text-to-speech synthesis (TTS) has progressed more or less independently without exerting much mutual influence. In human communication, on the other hand, a closed-loop speech chain mechanism with auditory feedback from the speaker's mouth to her ear is crucial. Mr. Tjandra takes a step further and develop a closed-loop machine speech chain model based on deep learning.

Mr. Tjandra contributed in the following research results.

1) Mr. Tjandra first proposed a deep learning framework that integrates human speech perception and production behaviors. The experimental results show that the proposed approach significantly improved performance over that from separate systems that were only trained with labeled data.

2) Mr. Tjandra integrates speech chain with speaker embedding model to achieve multi-speaker speech chain and improve the ASR and TTS performance on multi-speaker dataset settings.

3) Mr. Tjandra identifies the issue where the output of ASR is discrete variables, therefore he proposed a way to fully backpropagate the loss from TTS to the ASR model by using the straight-through estimator. In addition, he proposed an alternative ASR training with reinforcement learning to solve the discrepancy between training and inference stage.

The research proposed solutions to the problems which haven't been solved and a series of his research resulted in 3 journal papers, 23 peer-reviewed international conference papers, and 4 pee-reviewed international workshop papers and pre-prints. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.