（続紙 1）

京都大学	博士（情報学）	氏名	三村 正人
論文題目	End-to-end Transcription of Presentations and Meetings （講演・会議のend-to-end自動書き起こし）		

（論文内容の要旨）

This thesis presents a series of new techniques to improve the automatic transcription of real meetings and presentations with end-to-end models. End-to-end automatic speech recognition (ASR) has been intensively investigated because of its fast decoding and simplified architecture. It consists of a single neural network, unlike the hybrid HMM-based approach that uses modularized acoustic and language model components. However, it requires a considerable amount of paired data of speech and transcriptions for training, which are much more expensive to collect than unpaired speech and text resources for a target task or domain. To address this problem, we propose a framework for adapting end-to-end models using only untranscribed speech and text-only data to new domains with unseen acoustic and linguistic characteristics. Transcribing presentations and meetings is more challenging because they consist of spontaneous and long utterances. Moreover, faithfully transcribing these utterances word-by-word results in redundant and ungrammatical outputs that are not necessarily easy to read and comprehend. Therefore, we propose a method for robustly transcribing long utterances and a strategy for end-to-end generation of readable written-style texts directly from speech.

Chapter 2 gives a brief review of large vocabulary continuous speech recognition (LVCSR) and the introduction of end-to-end models.

In Chapter 3, we present an approach to cross-domain speech recognition based on acoustic feature mappings provided by a deep neural network, which is trained using non-parallel speech corpora from two domains without ground-truth labels. For training a target-domain acoustic model, we generate simulated target speech features from the labeled source domain features using a mapping. This forward mapping and the backward mapping are trained simultaneously with adversarial networks using a conventional adversarial loss and a cycle-consistency loss criterion that encourages the backward mapping to bring the translated feature back to the original as close as possible. In a highly challenging task of model adaptation only using the target-domain speech, this method achieved up to 16% relative improvements in word error rate (WER) in the evaluation using the CHiME3 real test data.

In Chapter 4, we investigate how we can leverage the latest speech synthesis technology to tailor the ASR system for a target domain by preparing only a relevant text corpus. From a set of target domain texts, we generate speech features using a sequence-to-sequence speech synthesizer. These synthesized speech features, together with real speech features from conventional speech corpora, are used to train an attention-based end-to-end ASR model. Experimental evaluations show that the proposed approach significantly improves the WER of presentation speeches of the CSJ from the baseline model trained only with real speech, although the synthetic part of

the training data comes only from a single speaker voice.

Chapter 5 investigates how bidirectional attention mechanisms can be integrated to improve the performance of ASR systems. The proposed approach decodes speech from left to right and right to left, utilizing forward and backward attention vectors. The best sentence hypothesis is chosen according to the combined probabilities provided by the decoders of two directions. The proposed bidirectional decoding improved the WER by up to 13% relative for presentation speeches of the CSJ.

In Chapter 6, we propose a novel approach that outputs clean, readable text directly from a meeting speech by removing fillers and disfluent regions, substituting colloquial expressions with formal ones, inserting punctuation, recovering omitted particles, and performing other types of appropriate corrections. We formalize this approach as end-to-end generation of written-style text from speech using a single neural network. We also propose a method to guide the training of this end-to-end model using automatically generated faithful transcripts and a novel speech segmentation strategy based on online punctuation detection. An evaluation using 700-hour Japanese Parliamentary speech demonstrates that the proposed direct approach successfully generates clean transcripts for human consumption more accurately at a faster decoding speed than the conventional cascade approach. We also conduct an in-depth analysis of the edit types that professional human editors perform in creating the official written records of Japanese Parliamentary meetings and evaluate the proposed system in terms of each edit type.

Chapter 7 concludes the thesis.

（続紙 2）

（論文審査の結果の要旨）

音声認識は、深層学習に基づくend-to-endモデルの導入により大きな進歩を遂げているが、音声と書き起こしテキストのペアデータが大規模に必要であり、講演や会議のような専門性と自発性の高い音声の対応には依然として大きな課題がある。本論文では、ペアデータを擬似的に生成することでその課題の解決を図るとともに、長くて冗長性の高い話し言葉音声から、会議録や字幕などの可読性の高い書き言葉スタイルのテキストを直接生成する方法を研究した成果をまとめたもので、主な成果は以下の通りである。

1. 新たな音響環境に対応するために、敵対的ネットワークを用いて、既存の大規模な音声データの特徴量をペアデータのない対象環境に変換する手法を提案した。サイクル一貫性制約に残差制約を導入することで、実際にクリーン音声から新たな雑音環境に対応できることを示し、雑音下の認識誤り率を大きく(約16%)改善した。

2. 新たな話題（ドメイン）に対応するために、音声合成技術を用いて、対象ドメインのテキストから音声を生成し、ペアデータを構成する方法を提案した。一般的な話し言葉音声のモデルから学会講演音声の対応を試み、認識誤り率を大きく(約22%)改善できることを示した。

3. 講演や会議のような長い入力音声に対応するために、通常行う前向き（時間）方向の仮説生成に加えて、終端から逆向き方向にも仮説生成を行い、両者の出力を統合する方法を提案した。本手法により、学会講演音声の認識誤り率を大きく(約13%)改善した。

4. 音声認識結果の出力テキストの可読性を向上させるために、フィラーや口語表現が多用される話し言葉音声から、会議録スタイルのテキストを直接生成するend-to-endモデルの構成法を提案した。会議録と音声のペアデータから、音声の忠実な書き起こしの別のペアデータを擬似的に生成することで、このモデル学習が効果的に実現できることを示し、実際に衆議院の会議音声から会議録テキストを高精度に生成できることを示した。

以上のように本論文は、end-to-end音声認識の本質的な課題の解決を図る方法を複数提案するとともに、国会の会議録作成支援に貢献できる技術を提示するもので、学術上・実用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和4年8月26日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。また、本論文のインターネットでの全文公開についても支障がないことを確認した。