（ 続紙 1 ）

京都大学	博士（情報学）	氏名	上乃 聖
論文題目	Data Augmentation Approaches for Automatic Speech Recognition Using Text-to-Speech （音声認識のための音声合成を用いたデータ拡張手法）		

（論文内容の要旨）

Automatic speech recognition (ASR) systems are widely used as an aid and basis for speech-based human-human and human-robot communications. ASR systems need to achieve high accuracy with small latency, and they should be customized for the application domains and topics. Thanks to the development of deep neural networks (DNNs), end-to-end ASR models have been intensively investigated recently. The end-to-end ASR models convert speech into words faster with a simpler architecture than the conventional models. However, they require a large amount of paired data of speech and transcription for training.

To alleviate this problem, this thesis addresses data augmentation for the ASR model using a text-to-speech (TTS) system. In this framework, artificial speech data are generated from text-only data using a TTS system to prepare pseudo paired datasets. In the naive implementation, however, we observe that the improvement of ASR performance is limited compared to the case using real speech data. This is because there are serious mismatches between synthesized data and real data. In this study, we investigate three data augmentation approaches to solve the problem.

In Chapter 3, we adopt a waveform-based approach. In general, a TTS system is composed of two models: a text-to-mel network to generate log Mel-scale filterbank (lmfb) features and a vocoder network to convert the generated lmfb features into a waveform. We observe that the lmfb feature produced by the text-to-mel model is blurry, particularly on the time dimension. This problem is mitigated by introducing the vocoder to generate speech of better quality or spectrogram of better time-resolution. This makes it possible to train waveform-input end-to-end ASR. Here we use CNN filters and apply a masking method similar to SpecAugment. We compare the waveform-input model with two kinds of lmfb-input models: (1) lmfb features are directly generated by TTS, and (2) lmfb features are converted from the waveform generated by TTS. Experimental evaluations show the effectiveness of the combination of waveform-output TTS and the waveform-input end-to-end ASR model for improving the ASR performance.

In Chapter 4, we propose a data augmentation approach via a discrete speech representation. In general TTS, a text-to-mel network predicts continuous value (lmfb features), which is not an easy task. It is also not guaranteed that the generated lmfb features exist in the real world. In this work, we introduce a discrete speech representation, which TTS model predicts instead of lmfb features. We expect that the use of the discrete representation based on vq-wav2vec not only makes TTS training easier but also mitigates the mismatch with real data. The ASR model

also uses the discrete representation as its input. Experimental evaluations show that the proposed method outperforms the data augmentation method using the conventional TTS. We found that it reduces speaker dependency, and the generated features are distributed more closely to the real features.

In Chapter 5, we propose a phone-informed post-processing network that refines lmfb features without using the vocoder. The widely-used procedure, as presented in Chapter 3, first generates an lmfb feature from text data, then converts it into a waveform, and converts it again to an lmfb feature. These conversions take a long time and are not necessary for data augmentation. In this work, we propose a mel-to-mel network that directly refines the lmfb features. The proposed network consumes not only lmfb features but also phone information for refinement. This approach takes less time than converting to the waveform domain. Experimental evaluations in domain adaptation show that the proposed network achieves better improvement of ASR performance than using the vocoder network with much faster processing time. It is also shown that the use of phone information is critical for the improvement.

Chapter 6 concludes this thesis with a comparison of the three works and a brief look at future work.

（続紙 ２ ）

（論文審査の結果の要旨）

音声認識は、ニューラルネットワークに基づくEnd-to-Endモデルにより大きな進歩を遂げているが、その学習には音声と書き起こしテキストのペアデータを大規模に必要とする。本論文は、テキストのみのデータを活用するために、音声合成を用いてデータ拡張を行う方法の研究成果をまとめたもので、主な成果は以下の通りである。

1. 音声合成は通常、テキストから周波数特徴量を生成する過程と周波数特徴量から音声波形を生成する過程からなる。通常の音声認識は周波数特徴量を入力とするので、前者のみで十分であるが、音声合成で生成される周波数特徴量は品質が十分でなく、いったん波形を生成することで、周波数特徴量の品質及びそれを用いて学習した音声認識の性能が改善されることを明らかにした。さらに、音声波形を入力とするEnd-to-Endモデルによる音声認識も実現した。

2. 周波数特徴量や音声波形のような数値ベクトル／系列でなく、それらの符号を入力とする音声認識、及びその符号を生成する音声合成を設計・実装することにより、効率的なデータ拡張の枠組みを提案した。これにより、周波数特徴量レベルでデータ拡張を行う場合と比べて高い性能を実現できることを示した。

3. 周波数特徴量を入力とする音声認識に直接的に対応するために、音声合成で生成される周波数特徴量を洗練する方法を検討した。通常の音声強調では発話内容は未知の設定が一般的であるが、音声合成ではテキストの音素系列が既知であることを活用することで効果的な音声強調を行う。音声波形をいったん生成する場合と比べてはるかに高速にデータ拡張が実現でき、同等以上の認識精度の改善を得ることができた。

以上のように本論文は、音声認識のための音声合成を用いたデータ拡張の方法を、音声波形、周波数特徴量、及び離散符号レベルで検討しており、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 2月24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。なお、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。