別紙４－１（課程博士（英文））

Date of Submission (month day, year) : Jan 11, 2019.

Department of Computer Science and Engineering	Student ID Number	123346
Applicant's name	Hiroshi Seki	

Supervisors	Tomoyosi Akiba Hitoshi Isahara

Abstract (Doctor)

Title of Thesis	Rapid adaptation of deep neural network for speech recognition (深層学習に基づく音声認識システムの高速適応)

Approx. 800 words

In recent years, automatic speech recognition (ASR) has become a popularized technology for us. The ASR system models conversion from speech to text by using training data. During operation, the ASR system recognizes input speech uttered by various speakers which are recorded in various conditions. However, since the ASR system cannot identify speaker characteristics and background noise information in advance, there is an acoustic mismatch between speech data used for the development of the ASR system (training data) and the input speech (evaluation data) which leads to performance degradation. Therefore, recover/improvement of recognition performance is required to accurately recognize the speech of various speakers uttered in various conditions.

Adaptation is one approach to suppress this acoustic mismatch. This method first collects speech data (adaptation data) under the same condition as the evaluation data, and then the ASR system is tuned to match the evaluation data based on the adaptation data. However, the data collection depending on each speaker and each condition takes tedious time and leads to the poor user experience. In other words, users have to utter additional words before the recognition of intended words which leads to undesired high latency. Therefore, it is important to adapt the ASR system rapidly and robustly using a small amount of adaptation data. In this research, we focus on the rapid adaptation of the ASR system based on a small size adaptation data and propose two adaptation methods.

Auxiliary-feature based adaptation method extracts information which degrades the recognition performance, e.g., speaker characteristic and background noise, and uses it as an auxiliary feature. However, earlier works assume the availability of speech from several ten seconds to several minutes. A widely used speaker representation called i-vector requires more than 5.0 seconds for robust estimation of speaker characteristic. Therefore, the usage of i-vector is not applicable to the recognition of a short time utterance such as keyword and key-phrase. In this thesis, we propose a speaker representation, speaker-class information, which is defined as a set of similarities between pre-defined speech clusters and the input speech. The average word error rate (WER) of the model without the input of speaker class information was 11.2%. The WER of our proposed model was decreased to 10.4% and the relative error reduction rate of 7.0% was obtained. These results

demonstrated that speaker class, which was estimated from only the 0.5 second in the utterance, provided important information to suppress the diversity of speakers, and it is applicable to the recognition of short time utterances, i.e., speech retrieval, speech assistance, and speech command input.

Model adaptation, another commonly used adaptation method, re-estimates parameters of the ASR system using the given adaptation data for the suppression of acoustic mismatch. Earlier works on model adaptation update sub-modules of neural network architecture and introduce various restriction methods and regularization methods for prevention of an overfitting problem. However, there is a trade-off between complexity and expressiveness. In other words, the update of a large number of parameters causes the overfitting problem or requirement of much adaptation data and the extreme restriction (low expressiveness) makes it difficult to adapt to the given adaptation data. Therefore, it is important to maximize expressiveness while minimizing the number of free parameters to robustly update the ASR system using the small size adaptation data. In this thesis, we proposed a neural network based filterbank layer which minimizes the number of free parameters while maximizing the expressiveness by focusing on a human auditory system and physical characteristics of speakers. We conducted experiments for speaker adaptation and noise adaptation, and showed that the proposed adaptation method obtained better recognition performance than the conventional adaptation methods under the scenario that the available adaptation data is limited. By adapting the filterbank layer using 15 utterances, WER was improved from 12.1% to 11.2%, and a word error reduction rate (WERR) of 7.4% was obtained. This WERR was better than the unadapted model at a significance level of 0.005 under a statistical sign test. These results showed that the adjustment of the filter shapes can handle the diversity of speakers. The recognition performance of our proposed model was better than the other adaptation methods, although other adaptation methods also showed performance improvements.

Through the above research, we developed the easy-to-use robust adaptation methods for the ASR system by reducing the required speech duration and the cost of the data collection procedure.