Chapter 16
Improvement of Lecture Speech Recognition by Using Unsupervised Adaptation

Tetsuo Kosaka
Yamagata University, Japan

Takashi Kusama
Yamagata University, Japan

Masaharu Kato
Yamagata University, Japan

Masaki Kohda
Yamagata University, Japan

ABSTRACT
The aim of this work is to improve the recognition performance of spontaneous speech. In order to achieve the purpose, the authors of this chapter propose new approaches of unsupervised adaptation for spontaneous speech and evaluate the methods by using diagonal-covariance and full-covariance hidden Markov models. In the adaptation procedure, both methods of language model (LM) adaptation and acoustic model (AM) adaptation are used iteratively. Several combination methods are tested to find the optimal approach. In the LM adaptation, a word trigram model and a part-of-speech (POS) trigram model are combined to build a more task-specific LM. In addition, the authors propose an unsupervised speaker adaptation technique based on adaptation data weighting. The weighting is performed depending on POS class. In Japan, a large-scale spontaneous speech database “Corpus of Spontaneous Japanese (CSJ)” has been used as the common evaluation database for spontaneous speech and the authors used it for their recognition experiments. From the results, the proposed methods demonstrated a significant advantage in that task.

DOI: 10.4018/978-1-61520-871-5.ch016

Copyright © 2011, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
1. INTRODUCTION

High recognition accuracy has been achieved for read speech with a large vocabulary continuous speech recognition (LVCSR) system. The recognition accuracy higher than 95% has been reported using the state-of-the-art speech recognition system (Furui, 2005a). However, it is well known that rather poor performance is reported for spontaneous speech recognition. Compared with read speech, spontaneous speech has repairs, hesitations, filled pauses, unknown words, unclear pronunciation, and so on, and those phenomena cause poor performance.

In order to improve the recognition performance of spontaneous speech, it is necessary to develop more accurate acoustic and language models which are suitable for spontaneous speech. Since difference between read speech and spontaneous speech is very large, it is difficult to use these models developed for read speech as models for spontaneous speech. Currently, state-of-the-art speech recognition systems use statistical-based models for both acoustic and language modeling. Development of accurate statistical models requires a large amount of training data. For this reason, a large-scale spontaneous speech corpus, Corpus of Spontaneous Japanese (CSJ) was developed and released in 2004 (Maekawa, 2003). By taking this opportunity, the research on this field becomes active in Japan.

The development of CSJ has brought steady progress in the research of Japanese spontaneous speech recognition. However, the absolute performance of recognition is still insufficient. This might result from that utterances of spontaneous speech have a vast variation both acoustically and linguistically. Even if accurate language and acoustic models are created by using a large amount of training data such as the CSJ, it is difficult to cover such an acoustic and linguistic variation in spontaneous speech.

In order to solve the problem, adaptation techniques for acoustic or language models have been investigated. These techniques are referred to as acoustic model (AM) adaptation or language model (LM) adaptation respectively. Many efforts have been made with these issues over the years. For example, maximum-likelihood linear regression (MLLR) technique is well known as the effective speaker adaptation method (Leggetter et al., 1995; Gales, 1998). Regarding the language model adaptation, class-based language model adaptation methods have been proposed to cope with the sparseness of language model space (Moore et al., 2000; Yamamoto et al., 2001).

The adaptation techniques can be classified into two categories. One is supervised adaptation and the other is unsupervised adaptation. In general, the adaptation techniques require both utterance data and transcription data in which the contents of utterances are described by speech units such as phonemes. Since the transcription data does not provide in the unsupervised mode, recognition results are used in place of them. Then, the performance of the unsupervised adaptation is lower than that of the supervised adaptation because of the influence of recognition errors. Generally the transcription data cannot be obtained in advance. Then, the unsupervised adaptation technique can be used for wide range of application areas.

The aim of this work is to improve the performance of spontaneous speech recognition by using both AM adaptation and LM adaptation techniques in the unsupervised mode. This chapter focuses on recognizing lecture speech. However, the proposed methods described below are expected to be applicable to another speaking style. We have an assumption that the recognition is performed off-line. After recognizing the lecture speech data, the model adaptation is carried out and the data are recognized again by using the adapted models. This kind of adaptation is called batch-type adaptation.

For the unsupervised adaptation, it is important to reduce the adverse effect of recognition errors. In order to solve the problem, we investigate two types of approaches for the unsupervised adapta-