Acoustic Model Adaptation for Codec Speech based on Learning-by-Doing Concept

Shingo Kuroiwa, Satoru Tsuge, Koji Tanaka, Kazuma Hara and Fuji Ren

Faculty of Engineering, The University of Tokushima,
Tokushimashi 770-8506, Japan
{kuroiwa,tsuge,ren}@is.tokushima-u.ac.jp,
WWW home page: http://a1-www.is.tokushima-u.ac.jp/

Abstract. Recently, personal digital assistants like cellular phones are shifting to IP terminals. The encoding-decoding process utilized for transmitting over IP networks deteriorates the quality of speech data. This deterioration causes degradation in speech recognition performance. Acoustic model adaptations can improve recognition performance. However, the conventional adaptation methods usually require a large amount of adaptation data. In this paper, we propse a novel acoustic model adaptation technique that generates "speaker-independent" HMM for the target environment based on the learning-by-doing concept. The proposed method uses HMM-based speech synthesis to generate adaptation data from the acoustic model of HMM-based speech recognizer, and consequently does not require any speech data for adaptation. By using the generated data after coding, the acoustic model is adapted to codec speech. Experimental results on G.723.1 codec speech recognition show that the proposed method improves speech recognition performance. A relative word error rate reduction of approximately 12% was observed.
Keywords: Speech Recognition, Model Adaptation, Codec Speech, Speech Synthesis, Learning-by-Doing

1 Introduction

In recent years, telephone speech recognition systems encompassing thousands of vocabularies have become practical and widely used [1, 2]. These systems are generally utilized by automatic telephone services for booking an airline ticket, inquiring about stock, receiving traffic information, and so on. However, the recognition accuracy of cellular phones is still inadequate due to compression coding or ambient noise [3–5].

Recently, personal digital assistants like cellular phones are shifting to IP terminals. For transmission over IP networks, speech data must be encoded at the sending end and subsequently decoded at the receiving end. This coding process deteriorates the quality of the voice data. Although most people can not notice this deterioration, it seriously affects the performance of those speech recognizers not designed for low-quality voice data[6]. The major causes of speech recognition performance degradation are : distortion in the transmission environment (transmission error and packet loss), and low bitrate speech coding (loss

Advances in Natural Language Proccesing
Research in Computing Science 18, 2006, pp. 105-114