Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 21 (2007) 231–246
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Environment adaptation for robust speaker verification by cascading maximum likelihood linear regression and reinforced learning ☆

K.K. Yiu [a], M.W. Mak [a,*], S.Y. Kung [b]

[a] *Department of Electronic and Information Engineering, Center for Multimedia Signal Processing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong*
[b] *Department of Electrical Engineering, Princeton University, United States*

Received 12 October 2005; accepted 18 May 2006
Available online 22 June 2006

Abstract

In speaker verification over public telephone networks, utterances can be obtained from different types of handsets. Different handsets may introduce different degrees of distortion to the speech signals. This paper attempts to combine a handset selector with (1) handset-specific transformations, (2) reinforced learning, and (3) stochastic feature transformation to reduce the effect caused by the acoustic distortion. Specifically, during training, the clean speaker models and background models are firstly transformed by MLLR-based handset-specific transformations using a small amount of distorted speech data. Then reinforced learning is applied to adapt the transformed models to handset-dependent speaker models and handset-dependent background models using stochastically transformed speaker patterns. During a verification session, a GMM-based handset classifier is used to identify the most likely handset used by the claimant; then the corresponding handset-dependent speaker and background model pairs are used for verification. Experimental results based on 150 speakers of the HTIMIT corpus show that environment adaptation based on the combination of MLLR, reinforced learning and feature transformation outperforms CMS, Hnorm, Tnorm, and speaker model synthesis.

1. Introduction

The state-of-the-art approach to text-independent speaker verification consists in using mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) as speaker features and Gaussian mixture models (GMMs) (Reynolds and Rose, 1995) for statistical speaker modeling. To increase the discrimination between

☆ Paper No. CSL034-03. (Revised Version). This work was supported by the Hong Kong Polytechnic University Grant Nos. PolyU 5214/04E and PolyU 5230/05E.
* Corresponding author. Tel.: +852 2766 6257; fax: +852 2362 8439.
E-mail address: enmwmak@polyu.edu.hk (M.W. Mak).

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2006.05.001

client speakers and impostors, a GMM-based background model (Reynolds et al., 2000) is typically used to represent impostors' characteristics. During verification, the ratio between the likelihood that the claimant is a genuine speaker and the likelihood that the claimant is an impostor is compared against a decision threshold for decision making. In case enrollment data for individual speakers are scarce, speaker-dependent GMMs can be adapted from the background model using maximum a posteriori (MAP) techniques (Reynolds et al., 2000). Because almost perfect verification has become achievable for clean and well-matched speech, researchers have focused on the problems of transducer mismatches and robustness in recent years. Environmental robustness is an important issue in telephone-based speaker verification because users of speaker verification systems tend to use different handsets in different situations. It has been noticed that recognition accuracy degrades dramatically when users use different handsets for enrollment and verification. This lack of robustness with respect to handset variability makes speaker verification over telephone networks a challenging task.

When sufficient speech data is available from a new acoustic environment, it is sensible to retrain the speaker and background models to accommodate the new environment. However, retraining on corrupted speech requires a large amount of data from each of the possible environments. An alternative is to use speech data from different acoustic environments to train the models. This is known as multi-style training (Lippmann et al., 1987). However, fine speaker characteristics will be blurred by pooling multiple training environments.

With some modifications, standard speaker adaptation methods, such as maximum a posteriori (MAP) (Lee et al., 1991) and maximum-likelihood linear regression (MLLR) (Leggetter and Woodland, 1995), can be used for environment adaptation. One of the positive properties of MAP is that its performance approaches that of maximum-likelihood-based methods provided that sufficient adaptation data are available. However, MAP is an unconstrained method in that adaptation is performed only on those model parameters who have "seen" the adaptation data. MLLR, on the other hand, applies a transformation matrix to a group of acoustic centers so that all the centers are transformed. As a result, MLLR provides a quick improvement, but its performance quickly saturates as the amount of adaptation data increases.

This paper investigates two model adaptation/transformation techniques – reinforced/anti-reinforced learning of probabilistic decision-based neural networks (PDBNNs) (Lin et al., 1997) and maximum-likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) – in the context of telephone-based speaker verification. These techniques adapt or transform the model parameters to compensate for the *mismatch* between the training and testing conditions. We have reported in Yiu et al. (2003) some preliminary results on PDBNN and MLLR adaptation for robust speaker verification. Here, we extend the results in Yiu et al. (2003) by combining these two techniques with stochastic feature transformation (SFT) (Mak and Kung, 2002). Specifically, precomputed MLLR transformation matrices are used to transform clean models to handset-dependent MLLR-adapted models. Then, PDBNN adaptation is performed on the MLLR-adapted models using handset-dependent, stochastically transformed patterns to obtain the final adapted models. The paper also compares the proposed channel compensation methods with state-of-the-art techniques, including CMS (Atal, 1974), Hnorm (Reynolds, 1997a), Tnorm (Auckenthaler et al., 2000), and speaker model synthesis (SMS) (Teunen et al., 2000). Experimental results based on 150 speakers of the HTIMIT corpus show that the proposed methods outperform these classical techniques.

The paper is organized as follows. In Section 2, we describe the techniques of model adaptation, including PDBNN and MLLR adaptation. Our proposed methods are detailed in Sections 2.4 and 2.5. Section 3 outlines the architecture of a handset selector that enables the verification system to select the most appropriate handset-dependent model for verification. In Section 4, speaker verification experiments that compare the effectiveness of different adaptation approaches are presented and the results are discussed in Section 5. Finally, a conclusion of the paper is provided in Section 6.

2. Model transformation and adaptation

To address the acoustic mismatch between training and recognition conditions, a number of compensation techniques have been proposed in the literature. These techniques can be roughly categorized into three classes: feature transformation (Mak and Kung, 2002; Sankar and Lee, 1996, Reynolds, 2003), model transforma-