Available online at www.sciencedirect.com

ScienceDirect

Neurocomputing 71 (2007) 137–146

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Probabilistic feature-based transformation for speaker verification over telephone networks

Man-Wai Mak[a,*], Kwok-Kwong Yiu[a], Sun-Yuan Kung[b]

[a] Center for Multimedia Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

[b] Department of Electrical Engineering, Princeton University, USA

Available online 15 August 2007

Abstract

Feature transformation aims to reduce the effects of channel- and handset-distortion in telephone-based speaker verification. This paper compares several feature transformation techniques and evaluates their verification performance and computation time under the 2000 NIST speaker recognition evaluation protocol. Techniques compared include feature mapping (FM), stochastic feature transformation (SFT), blind stochastic feature transformation (BSFT), feature warping (FW), and short-time Gaussianization (STG). The paper proposes a probabilistic feature mapping (PFM) in which the mapped features depend not only on the top-1 decoded Gaussian but also on the posterior probabilities of other Gaussians in the root model. The paper also proposes speeding up the computation of PFM and BSFT parameters by considering the top few Gaussians only. Results show that PFM performs slightly better than FM and that the fast approach can reduce computation time substantially. Among the approaches investigated, the fast BSFT (fBSFT) strikes a good balance between computational complexity and error rates, and FW and STG are the best in terms of error rates but with higher computational complexity. It was also found that fusion of the scores derived from systems using fBSFT and STG can reduce the error rate further. This study advocates that fBSFT, FW, and STG have the highest potential for robust speaker verification over telephone networks because they achieve good performance without any *a priori* knowledge of the communication channel.

Keywords: Speaker verification; Feature transformation; Channel compensation; Biometrics; EM algorithm

1. Introduction

Speaker verification is a biometric technology that aims to authenticate users via their voice patterns. Among the biometric traits that are currently under intensive investigation, speaker verification is apparently the best candidate for identifying or authenticating users over the telephone networks. Although commercial speaker verification systems that aim at securing financial transactions and remote information access are now available, the lack of robustness to channel variability and the acoustic mismatch between enrollment and verification conditions remain a major practical challenge. Currently, this problem is addressed by a technique called channel mismatch compensation.

The goal of channel compensation is to achieve performance approaching that of a “matched condition” system. Channel compensation can be applied in feature space [15,19], model space [2,20] or score space [11]. One advantage of feature-space compensation is that it is not necessary to modify the speaker models after training.

This paper focuses on feature-based compensation and compares several closely related feature compensation techniques under the 2000 NIST SRE framework [16]. Techniques compared include feature mapping (FM) [12], stochastic feature transformation (SFT) [8], blind stochastic feature transformation (BSFT) [19], feature warping (FW) [9], and short-time Gaussianization (STG) [18]. The first three techniques attempt to transform distorted spectral features to fit the clean acoustic models. The last two, FW and STG, attempt to make the features less channel-dependent by normalizing the feature distribution.

*Corresponding author. Tel.: +852 27666257; fax: +852 23628439.
E-mail address: enmwmak@polyu.edu.hk (M.-W. Mak).
URL: http://www.eie.polyu.edu.hk/~mwmak/mypage.htm (M.-W. Mak).

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.08.003

The paper proposes improving the performance of FM by introducing a probabilistic term in the mapping function so that the mapped features depend not only on the winner mixture but also on the posterior probabilities of other mixtures in the root model. The resulting mapping is referred to as probabilistic feature mapping (PFM). Because both BSFT and PFM require the posterior probabilities of all mixtures in the parameter estimation process, computation time can become excessively long for large model size. The paper therefore proposes speeding up the computation of FM and BSFT's parameters by considering the top few components only in the parameter estimation process.

The paper is organized as follows. Section 2 discusses two main types of channel compensation: blind and non-blind. Section 3 explains the BSFT and its fast version, which is followed by Section 4 where FM and fast FM are outlined. These methods are then compared in Sections 5 and 6 under the NIST00 evaluation protocol.

2. Feature transformation for channel compensation

There are two main types of channel compensation: blind and non-blind. The former adapts speaker models or transforms speaker features during recognition to accommodate the channel variation without *a priori* knowledge of the channel characteristics. Non-blind compensation, on the other hand, estimates channel-specific compensation based on *a priori* knowledge of all possible channels. Specifically, during recognition, the channel type is identified and is used to select the pre-computed channel compensation to reduce the acoustic mismatch caused by mismatched channels.

2.1. Blind compensation

Blind compensation can be categorized into four types. The first type exploits the temporal variability of feature vectors. For example, cepstral mean subtraction (CMS) [1] subtracts the cepstral mean of an utterance from each of the cepstral vectors. RASTA [7] applies a bandpass filter to the sequence of cepstral vectors to remove the slow varying components corresponding to the channel. It has been shown that both mean normalization and bandpass filtering can minimize the filtering effect of linear channels [1,7]. However, these techniques may cause performance degradation when both training and recognition are derived from the same acoustic environment [20].

The second type of blind compensation transforms the distorted features such that acoustic environments have minimum effect on the distribution of the transformed features. For example, in feature warping [9], observed features are mapped to a target distribution (e.g., standard normal) such that they follow the target distribution over a sliding window of feature vectors. Specifically, given a sequence of feature vectors, a sliding window of 3 s is applied to the sequence to compute the cumulative distribution function (cdf) of each feature component. For each feature component, the original feature value at the middle of the sliding window is then mapped to a target value such that the cdf at the original feature value is equal to the target cdf at the target value. The warping can be viewed as a nonlinear feature transformation from the original feature to a warped feature. Feature warping has been shown to be robust to channel variations and background noise. In short-time Gaussianization [18], a linear transformation is applied to the distorted feature before mapping them to a normal distribution. The transformation aims to decorrelate the feature vectors, making them more amendable to diagonal covariance Gaussian mixture models (GMMs). Short-time Gaussianization has shown advantages over feature warping, especially at low false acceptance rate.

The third type makes use of the statistical difference between the clean acoustic models and the distorted speech to estimate a transformation matrix to map the distorted vectors to fit the clean model. This type of technique include the blind stochastic feature transformation [19] to be detailed in Section 3.

The fourth type, namely discriminative feature design [6], trains a neural network discriminatively to maximize speaker recognition performance on the training set. Because the training set consists of different types of acoustic distortions that the system may encounter during recognition, the neural network is able to "recall" the compensation required during recognition to reduce the effects of handset distortions on speaker discrimination.

2.2. Non-blind compensation

One typical property of non-blind techniques is the requirement of channel detection during recognition. Typical examples are feature mapping [12], spectral-magnitude matching [10], and stochastic feature transformation [8].

In feature mapping, the handset type of the testing utterance is identified by a handset detector; feature vectors are then mapped to the channel-independent space based on the closest Gaussian in the channel-dependent GMM. In spectral-magnitude matching [10], a nonlinear polynomial mapper is trained to minimize the mean-squared spectral magnitude error between speech arising from electret and carbon-button handsets. The mapper is shown to be good at minimizing mismatches caused by phantom formants, bandwidth widening, and spectral flattening due to channel nonlinearity. Stochastic feature transformation (SFT) [8] is derived from the stochastic matching method of Sankar and Lee [14], which was originally proposed for robust speech recognition. SFT aims to transform the distorted features to fit the clean speech models by selecting the most appropriate pre-computed transformation matrix. It has been shown that SFT can be extended to nonlinear feature transformation to overcome the nonlinear distortion [8].