A GENERALIZED FEATURE TRANSFORMATION APPROACH FOR CHANNEL ROBUST SPEAKER VERIFICATION

Donglai ZHU[1]*, Bin MA*[1]*, Haizhou LI*[1] *and Qiang HUO*[2]

[1] Institute for Infocomm Research, Singapore 119613
[2]Department of Computer Science, The University of Hong Kong, Hong Kong, China
(E-mails: {dzhu,mabin,hli}@i2r.a-star.edu.sg qhuo@cs.hku.hk)

ABSTRACT

In this paper we propose a generalized feature transformation approach to compensating for channel variation in speaker verification (SV) applications. Channel-dependent (CD) piecewise linear transformations are used for feature compensation. CD transformation parameters are estimated together with a channel-independent (CI) root Gaussian mixture model (GMM) from training data with a variety of channel conditions by using a maximum likelihood criterion. Experiments are conducted on the 2005 NIST Speaker Recognition Evaluation (SRE) corpus for several text-independent GMM-based SV systems. Experimental results show that the proposed approach achieves relative equal error rate (EER) reductions of 8.19% and 26.24% in comparison with a traditional feature mapping approach and a baseline system, respectively.

Index Terms— speaker verification, channel compensation, feature mapping, generalized feature transformation, maximum likelihood.

1. INTRODUCTION

How to deal with channel variation is one of the main challenges in speaker verification (SV) applications. Over the years, at least three types of compensation techniques have been studied for coping with this problem, namely feature-domain compensation (e.g., [1, 2, 3]), score-domain compensation (e.g., [4, 5]), and model-domain compensation (e.g., [6, 7, 8, 9]). In this study, we focus on the feature-domain compensation, because it is not tied to any particular SV approach. Besides the well-known techniques such as cepstral mean subtraction and RASTA filtering, a simple yet effective *feature mapping* (FM) approach was proposed in [3]. Actually, many effective feature compensation techniques for robust SV were inspired by, borrowed or adapted from the relevant techniques invented originally for robust automatic speech recognition (ASR). In the past decade, many interesting feature compensation techniques have been proposed and studied for robust ASR (e.g., [10, 11, 12, 13, 14, 15]), and some of them have not been tried out yet for robust SV. Inspired and encouraged by the promising results of the FM approach for SV in [3], we have adopted a feature compensation approach originally proposed in [15] for robust ASR and modified it to deal with the channel variability for robust SV. Experiments are conducted on the 2005 NIST Speaker Recognition Evaluation (SRE) corpus [16] to evaluate the effectiveness of the above approach. For the convenience of reference and the purpose of differentiation, our proposed approach is referred to as a *generalized feature transformation* (GFT) approach hereinafter. The main purpose of this paper is to report our study on this topic.

The rest of the paper is organized as follows. In Section 2, we review the FM approach of [3]. In Section 3, we present our GFT approach. Evaluation results are reported in Section 4. Finally, we conclude the paper in Section 5.

2. FEATURE MAPPING BASED SV SYSTEM

The feature mapping approach proposed in [3] works as follows. First a channel-independent (CI) root Gaussian mixture model (GMM) with parameters $\lambda = \{\omega_k, \mu_k, \Sigma_k; k = 1, \ldots, K\}$ is trained using an aggregation of data, $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_C\}$, from many different channels, where ω_k, μ_k and Σ_k are the mixture coefficient weight, the D-dimensional mean vector and $D \times D$ diagonal covariance matrix of the k-th Gaussian component respectively, and K is the number of Gaussian components. Next a set of channel-dependent (CD) GMMs, $\Lambda^{(c)} = \{\lambda^{(c)}, c = 1, \ldots, C\}$, are trained by adapting the root GMM, λ, using each set of channel-dependent data $\mathcal{Y}_c$ respectively, where $\lambda^{(c)} = \{\omega_k^{(c)}, \mu_k^{(c)}, \Sigma_k^{(c)}; k = 1, \ldots, K^{(c)}\}$, and $K^{(c)} = K$.

Given an input speech utterance $Y = \{y_1, y_2, \ldots, y_T\}$ of a speaker, the most likely channel condition is first identified as the one maximizing the likelihood of channel-dependent GMM:

$$c = \arg\max_{c'} p(Y|\lambda^{(c')}) \ . \tag{1}$$

For each feature vector y_t in the utterance, the "top-1" Gaussian component in GMM $\lambda^{(c)}$ is then selected:

$$k_t = \arg\max_{k'} p(k'|y_t, \lambda^{(c)}) \ . \tag{2}$$

The feature vector y_t is finally mapped to a channel-independent space as follows:

$$x_t = \Sigma_{k_t}(\Sigma_{k_t}^{(c)})^{-1}(y_t - \mu_{k_t}^{(c)}) + \mu_{k_t} \ . \quad (3)$$

Although the above feature mapping can be treated as a part of the front-end processing module and is independent of the follow-on speaker verification (SV) system, in this study, we adopt the same strategy as in [3] to share the verification model with the mapping model for greater efficiency. More specifically, our text-independent GMM-based SV system is essentially the same as the ones described in [4, 3]. For each speaker, we train a GMM by using the mapped features from enrollment speech for MAP adaptation of the above-mentioned channel-independent root GMM. The same root GMM is then used as a universal background model (UBM). During verification, the log likelihood scores of the mapped features from the input speech are computed against the speaker and root GMMs respectively, and their difference (i.e., the log likelihood ratio score) is calculated and compared to a threshold to decide whether to accept or reject the putative speaker claim.

3. OUR APPROACH

3.1. Generalized Feature Transformation

In this paper, we propose to use the following piecewise linear transformation, which is borrowed from [15], for feature mapping:

$$x = \mathcal{F}(y;\Theta^{(c)}) = A^{(c)}y + \sum_{k=1}^{K^{(c)}} p(k|y,\lambda^{(c)})b_k^{(c)} \ , \quad (4)$$

where $A^{(c)}$ is a nonsingular $D\times D$ matrix, $b_k^{(c)}$ is a D-dimensional vector, and c denotes the corresponding channel condition to which y belongs. For the convenience of notation, we use $\Theta^{(c)} = \{A^{(c)}, b_k^{(c)}; k = 1,\ldots,K^{(c)}\}$ to denote the set of trainable parameters of the above CD transformation function $\mathcal{F}(y;\Theta^{(c)})$. In order to make sure the transformation parameters can be well trained with sufficient data, we use a single $A^{(c)}$ for each channel condition while allow the bias vector $b_k^{(c)}$ being Gaussian component dependent. Other options of parameter tying are possible, but they are not explored further in this study.

Given the set of training data $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_C\}$, the above transformation parameters $\Theta = \{\Theta^{(c)}, c = 1,\ldots,C\}$ and the parameters of the CI root GMM λ can be estimated by maximizing the following likelihood function

$$\mathcal{L}(\Theta,\lambda) = \prod_{Y_i\in\mathcal{Y}} p(\mathcal{F}(Y_i;\Theta)|\lambda) \quad (5)$$

defined on the training data $\mathcal{Y}$. In the following subsection, we describe in detail an approximate ML training procedure, which again is adapted from the relevant procedure originally proposed in [15] for robust ASR.

3.2. ML Training Procedure

Our ML training procedure is as follows:

Step 1: *Initialization*

First, the initial parameters of the root GMM λ are trained from multi-channel training data $\mathcal{Y}$. The initial values of transformation matrices $A^{(c)}$ are set to be identity matrices and the initial values of bias vectors $b_k^{(c)}$ are set to be zero vectors.

Step 2: *Estimating feature transformation parameters* Θ

Given the root GMM parameters λ, for each channel c, we estimate the channel-dependent feature transformation parameters $\bar{\Theta}_c$ to increase the likelihood function $\mathcal{L}(\Theta,\lambda)$. It can be achieved by iterating the following two sub-steps N_Θ times:

Step 2-1: *Estimating* $A^{(c)}$ *by fixing* $b_k^{(c)}$

By fixing $b_k^{(c)}$'s, the updating formula for $A^{(c)}$ can be derived in a similar way as CMLLR in [11]. Let's use $A_r^{(c)}$ to denote the r-th row of $A^{(c)}$. $A^{(c)}$ is updated N_A times as follows:

$$A_r^{(c)} = \alpha_r^{(c)}p_r^{(c)}G_r^{(c)-1} + v_r^{(c)}G_r^{(c)-1} \ , \quad (6)$$

where $p_r^{(c)}$ is the cofactor row vector $[c_{r1}^{(c)} \ \ldots \ c_{rD}^{(c)}]$ with $c_{rl}^{(c)} = cof(A_{rl}^{(c)})$, and

$$G_r^{(c)} = \sum_{i\in I_c}\sum_t\sum_m \frac{1}{\sigma_{mr}^2}\zeta_{it}(m)y_{it}y_{it}^{Tr} \ , \quad (7)$$

$$v_r^{(c)} = \sum_{i\in I_c}\sum_t\sum_m \frac{1}{\sigma_{mr}^2}\zeta_{it}(m)(\mu_{mr} - b_{k_t r}^{(c)})y_{it}^{Tr} , \quad (8)$$

$$\alpha_r^{(c)} = -\frac{\varepsilon_2}{2\varepsilon_1} \pm \frac{\sqrt{\varepsilon_2^2 + 4\beta^{(c)}\varepsilon_1}}{2\varepsilon_1} \ , \quad (9)$$

$$\beta^{(c)} = \sum_{i\in I_c}\sum_t\sum_m \zeta_{it}(m) \ , \quad (10)$$

$$\varepsilon_1 = p_r^{(c)}G_r^{(c)-1}p_r^{(c)Tr} \ , \quad (11)$$

$$\varepsilon_2 = p_r^{(c)}G_r^{(c)-1}v_r^{(c)Tr} \ . \quad (12)$$

The value of $\alpha_r^{(c)}$ is selected that maximizes

$$\mathcal{Q}_c = \beta^{(c)}\log|\alpha_r^{(c)}\varepsilon_1 + \varepsilon_2| - \frac{1}{2}\alpha_r^{(c)2}\varepsilon_1 \ . \quad (13)$$

In the above equation, I_c denotes the subset of the subscript of training utterance Y_i which belongs to the channel c, k_t is calculated by using Eq. (2), and

$$\zeta_{it}(m) = \frac{\omega_m\mathcal{N}(A^{(c)}y_{it} + b_{k_t}^{(c)};\mu_m,\Sigma_m)}{\sum_{j=1}^K \omega_j\mathcal{N}(A^{(c)}y_{it} + b_{k_t}^{(c)};\mu_j,\Sigma_j)} \ . \quad (14)$$

Step 2-2: *Estimating* $b_k^{(c)}$ *by fixing* $A^{(c)}$

After $A^{(c)}$ is updated in the above step, $b_k^{(c)}$'s are updated N_b times as follows:

$$b_{kd}^{(c)} = \frac{\sum_{i \in I_c} \sum_{t,m} \delta_{kit} \zeta_{it}(m)(\mu_{md} - A_d^{(c)} \cdot y_{it})/\sigma_{md}^2}{\sum_{i \in I_c} \sum_{t,m} \delta_{kit} \zeta_{it}(m)/\sigma_{md}^2}, \quad (15)$$

where

$$\delta_{kit} = \begin{cases} 1 & \text{if } k = \arg\max_{k'} p(k'|y_{it}, \lambda^{(c)}) \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

and $\zeta_{it}(m)$ can be calculated by using Eq. (14).

Step 3: *Estimating the Root GMM Parameters* λ

Given the updated parameters of feature transformations $\bar{\Theta}$, the training feature vectors are compensated using Eq. (4). Using the compensated training feature vectors, N_λ EM iterations are performed to re-estimate the root GMM parameters $\bar{\lambda}$, with an increase of the likelihood function $\mathcal{L}(\bar{\Theta}, \lambda)$.

Step 4: *Repeat* **Step 2** *and* **Step 3** N_{total} *times.*

3.3. Discussions

Compared with the feature mapping approach in [3], our generalized feature transformation approach incurs additional computational overhead only during the offline training of transformation parameters and the root GMM parameters, but not for the enrollment of target speakers and scoring of testing trials. In both approaches, the main online computational costs come from the calculation of probability density function (PDF) values of each feature vector against Gaussian components of CD-GMMs $\Lambda^{(c)}$. One of the advantages of our approach is that we can use a much smaller value of $K^{(c)}$ than that of K, therefore the online computational costs can be reduced dramatically.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

We perform SV experiments on the 2005 NIST SRE corpus [16] to evaluate the performance of different feature mapping approaches. We focus on the task of "one two-channel conversation training" and "one two-channel conversation testing" combination. The testing consists of 1,941 true trials and 29,477 false trials.

In feature extraction, speech frames are obtained using 30ms window size and 20ms frame rate. Each frame of feature vector has 36 coefficients including 12 MFCCs, their first and second derivatives. An energy-based voice activity detection is performed to discard vectors from low-energy frames. RASTA, utterance-based mean and variance normalization are applied to the features to mitigate channel effects.

Table 1. Equal Error Rates (EERs) of different systems in Fig. 1

Systems	EER (in %)	Relative EER Reduction (in %)
Baseline	12.46	-
FM	10.01	19.66
GFT-32	9.48	23.92
GFT-512	9.19	26.24

Two gender-dependent root GMMs are trained for female and male speakers respectively. Each GMM consists of 512 Gaussian components with diagonal covariance matrices. They are trained using the single-channel conversation training data in the 2004 NIST SRE task [16]. The training data are recorded from 170 male and 275 female speakers, and labeled with recording devices including 3 types of phones (Cellular, Landline, Cordless) and 4 types of microphones (Speakerphone, Headset, Ear-bud, Regular). There are a total of 12 channel conditions in combination.

In the GMM-based baseline SV system, each target speaker's GMM is adapted from the root GMM of the same gender using the MAP adaptation. In the feature mapping (FM) approach of [3], for each gender, twelve CD-GMMs are adapted from the corresponding gender-dependent root GMM. For our generalized feature transformation (GFT) approach, we implemented two systems with different numbers (namely 512 and 32) of Gaussian components for each gender-dependent CD-GMM. For the joint ML training of the parameters of feature transformations and the gender-dependent root GMM, the relevant control parameters are set as $N_{total} = 1$, $N_\Theta = 1$, $N_A = 1$, $N_b = 1$, and $N_\lambda = 2$. In both feature mapping based approaches, feature vectors are compensated by the respective feature mapping functions before sent to the GMM-based SV systems.

4.2. Experimental Results

Fig. 1 illustrates the DET curves of the GMM-based baseline system, the FM-based system, and the two GFT-based systems (GFT-32 and GFT-512) with 32 and 512 mixture components in CD-GMMs respectively. It is observed that the FM approach improves the DET curve compared with the baseline system. The two GFT-based systems yield similar DET curves. It indicates that the number of mixture components in the CD-GMMs can be set to relatively small values, which helps save computational costs dramatically. Table 1 lists the equal error rates (EERs) corresponding to the DET curves in Fig. 1. It is observed that in comparison with the baseline system, the FM-based system and the GFT-512 system achieve relative EER reductions of 19.66% and 26.24% respectively. In comparison with the FM-based system, the GFT-512 system and the GFT-32 system achieve relative EER reductions of 8.19% and 5.29% respectively.

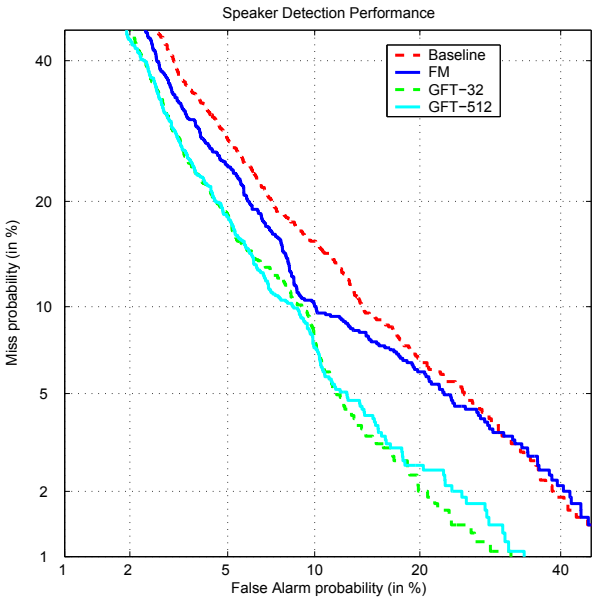

Fig. 1. DET curves of several SV systems on the 2005 NIST SRE task of "1-conversation train and 1-conversation test": GMM-based baseline system, feature mapping (FM)-based system, generalized feature transformation based systems with 32 (GFT-32) and 512 (GFT-512) Gaussian mixture components for each CD-GMM.

5. SUMMARY

In this paper, we have proposed a generalized feature transformation approach to compensating for channel variation in speaker verification (SV) applications. Channel-dependent (CD) piecewise linear transformations are used for feature compensation. CD transformation parameters are estimated together with a channel-independent (CI) root Gaussian mixture model (GMM) from training data with a variety of channel conditions by using a maximum likelihood criterion. Evaluation results on the 2005 NIST SRE task demonstrate that our proposed approach outperforms a traditional feature mapping approach reported in [3].

6. REFERENCES

[1] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *Proc. Odyssey*, 2001, pp.213-218.

[2] M. W. Mak and S. Y. Kung, "Combining stochastic feature transformation and handset identification for telephone-based speaker verification," *Proc. ICASSP*, 2002, pp.701-704.

[3] D. A. Reynolds, "Channel robust speaker verification via feature mapping," *Proc. ICASSP*, 2003, pp.II-53-56.

[4] D. A. Reynolds, T. Quatieri and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp.19-41, 2000.

[5] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp.42-54, 2000.

[6] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," *Proc. ICSLP*, 2000.

[7] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," *Proc. Odyssey*, 2004, pp.219-226.

[8] R. Vogt and S. Sridharan, "Experiments in session variability modelling for speaker verification," *Proc. ICASSP*, 2006, pp.897-900.

[9] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," *Proc. ICASSP*, 2005, pp.629-632.

[10] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp.190-202, 1996.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp.75-98, 1998.

[12] L. Deng, A. Acero, M. Plumpe, and X.-D. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," *Proc. ICSLP*, 2000, pp.806-809.

[13] J. Wu and Q. Huo, "An environment compensated minimum classification error training approach and its evaluation on AURORA2 database," *Proc. ICSLP*, 2002, pp.453-456.

[14] J. Wu, Q. Huo, and D. Zhu, "An environment compensated maximum likelihood training approach based on stochastic vector mapping," *Proc. ICASSP*, 2005, pp.429-432.

[15] Q. Huo and D. Zhu, "A maximum likelihood training approach to irrelevant variability compensation based on piecewise linear transformations," *Proc. Interspeech — ICSLP*, 2006, pp.1129-1132.

[16] NIST, "The NIST year 2004/2005 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk, 2004/05.