A New Approach to Channel Robust Speaker Verification via Constrained Stochastic Feature Transformation

Kwok-Kwong Yiu, Man-Wai Mak, Ming-Cheung Cheung

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University
USA

Abstract

This paper proposes a constrained stochastic feature transformation algorithm for robust speaker verification. The algorithm computes the feature transformation parameters based on the statistical difference between a test utterance and a composite GMM formed by combining the speaker and background models. The transformation is then used to transform the test utterance to fit the clean speaker model and background model before verification. By implicitly constraining the transformation, the transformed features can fit both models simultaneously. Experimental results based on the 2001 NIST evaluation set show that the proposed algorithms achieves significant improvement in both equal error rate and minimum detection cost when compared to cepstral mean subtraction and Z-norm. The performance of the proposed transformation approach is also slightly better than the short-time Gaussianization method proposed in [1].

1. Introduction

The accuracy of speaker recognition systems that enroll client speakers under one acoustic environment but verify claimants under another environment could be significantly lower than the ones that enroll and verify speakers under the same environment. This is mainly due to the acoustic mismatch between the training and recognition conditions, which is mainly caused by transducer variability. Transducer variability occurs when a system is trained with speech data obtained from one type of transducer and is subsequently tested on speech data recorded from other types of transducers. The goal of channel compensation is to achieve performance approaching that of a "matched condition" system while avoiding the need for a large amount of training data.

Channel compensation can be applied in feature space, model space or score space. Feature-based compensation [2], [3] transforms channel-distorted speech features to fit clean speaker models, whereas model-based compensation [4], [5] adapts or transforms the parameters of clean models to fit a new acoustic environment. On the other hand, score-based compensation [6] aims to minimize environment-dependent bias by normalizing the distribution of speaker scores.

Channel compensation can also be supervised or unsupervised. Supervised compensation assumes that the channel or handset characteristics are known a priori. Therefore, channel-specific compensation can be derived before recognition takes place. If handset labels are available during recognition, the corresponding channel-specific compensation can be applied to reduce the mismatch effect. Alternatively, one can detect the handset label from speech signals during verification [3]. However, this approach may not be practical because users may use a new handset, which is not well represented in the training set, during verification. While this problem can be partially resolved by using a handset classifier with out-of-handset rejection capability [7, 8], it is difficult to find a threshold for detecting unseen handsets. On the other hand, unsupervised compensation does not assume any knowledge of the channel characteristics. In particular, it adapts speaker models or transforms speaker features to accommodate the channel variation based on verification utterances only. Therefore, handset detectors are no longer required.

This paper addresses the problem of unsupervised compensation in which the features of a test utterance are transformed to better fit the clean speaker model and background model. In speaker verification, it is important to ensure that the transformation of either the features or models compensates channel variations instead of speaker variations. In particular, given a claimant's utterance recorded from an environment different from the enrollment environment, we aim to transform the features of the utterance as if it were recorded from the enrollment environment. Therefore, one cannot simply transform the claimant's utterance to fit the speaker model only because by doing so, the claimant will be accepted regardless of his/her genuineness. In speaker verification, the decision is based on the likelihood of the speaker model and background model. Given a test utterance from an unknown channel, we need to transform the features to fit both models simultaneously but not independently. This paper proposes a new feature-based transformation approach to address the above problem by implicitly constraining the transformation. Specifically, a feature-based transformation is estimated based on the statistical difference between a test utterance and a composite GMM formed by combining the speaker and background GMMs. The transformation is then used to transform the test utterance before verification. Hereafter, we refer to this transformation approach as constrained stochastic feature transformation (constrained SFT).

2. Constrained Stochastic Feature Transformation

As there is no handset nor channel label in the test utterances, we adopted the following unsupervised approach to environment adaptation. The speaker model Λ_s^N and universal background model (UBM) Λ_b^N—both with N components—were combined to form a composite GMM Λ_c^{2N} with $2N$ compo-

This project was supported by the Hong Kong Polytechnic University Grant No. G-W076 and the Research Grants Council of Hong Kong SAR (Project No. PolyU 5129/01E).

nents. During the combination, the means and covariances of the components were unchanged while the mixing coefficients were divided by two. This step ensures that the output of the composite GMM represents a probability density function.

Another UBM (Λ_b^{2N}) of $2N$ components was trained using the training utterances of all client speakers. Then, for each test utterance, a testing GMM (Λ_t^{2N}) with $2N$ components was created by adapting the UBM (Λ_b^{2N}) using maximum a posteriori (MAP) adaptation [9]. Using the test utterance, a set of feature transformation parameters ν were computed based on the stochastic feature transformation technique [3]. We used stochastic feature transformation of zeroth- and first-order with $K = 1$ in Eq. 2 of [3]. More specifically, given a D-dimensional distorted vector $\mathbf{x}$, the transformed feature vector is

$$\hat{\mathbf{x}} = f_\nu(\mathbf{x}) = A\mathbf{x} + \mathbf{b} \quad (1)$$

where A is a $D \times D$ identity matrix for zeroth-order transformation and $A = \text{diag}\{a_1, a_2, \ldots, a_D\}$ for first-order transformation, and $\mathbf{b}$ represents a bias vector. Fig. 1 illustrates the process of determining the transformation parameters ν. The main idea is to transform the test data, which is modelled by the testing GMM Λ_t^{2N}, to fit the composite GMM Λ_c^{2N}. By transforming test utterances to fit the composite GMM, constraints on the feature transformation will be automatically and implicitly imposed.

Because the computation complexity of estimating SFT parameters grows with the amount of adaptation data and the total number of mixture components in the GMMs, the constrained SFT will become computationally intensive when the number of components is large. To perform rapid adaptation, this paper proposes adopting a light-weight approach to computing transformation parameters. One of the positive properties of SFT is that the transformation can be estimated using GMMs with only a few components. In the light-weight approach, a small, composite GMM (Λ_c^{2M}) is synthesized from another small speaker GMM (Λ_s^M) and background GMM (Λ_b^M), both with M components where $M \ll N$. Similarly, the testing GMM (Λ_t^{2M}) was adapted from another UBM with $2M$ components. It was found that a good trade-off between performance and computation complexity can be maintained by using a suitable value of M.

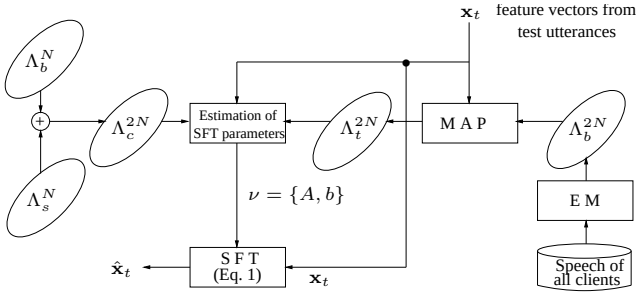


Figure 1: *Estimation of constrained stochastic-feature transformation parameters*

Fig. 2 illustrates the idea of environment adaptation in a two-class problem. Fig. 2(a) plots the clean and distorted patterns of Class 1 and Class 2. The upper right (resp. lower left) clusters represent the clean (resp. distorted) patterns. The ellipses show the corresponding equal density contours. Single, first-order stochastic feature transformation was used in these figures. In Figures 2(b), 2(c) and 2(d), markers '♦' and '■' represent the centers of the clean models.

In Fig. 2(b), the transformation is derived from the distorted data of Class 2 and the GMM of Class 1 (GMM1). The transformation adapts all data to a region around GMM1 since the transformation only takes account of GMM1. As a result, all distorted data, especially for those not belonging to Class 1, will be transformed to a region around GMM1, which results in a high error rate. Similarly, in Fig. 2(c), the transformation is estimated from the distorted data of Class 1 and the GMM of Class 2 (GMM2). Again, this transformation results in a high error rate. Fig. 2(c) illustrates the idea of constrained SFT. In this case, the transformation is estimated from the distorted data of Class 1 and a composite GMM formed by combining GMM1 and GMM2. The constrained transformation adapts the data to a region between GMM1 and GMM2. Hence, the transformed data will not be biased towards Class 1 nor Class 2. The adaptation capability of constrained SFT is also demonstrated in a speaker verification evaluation to be described next.

3. Experiments

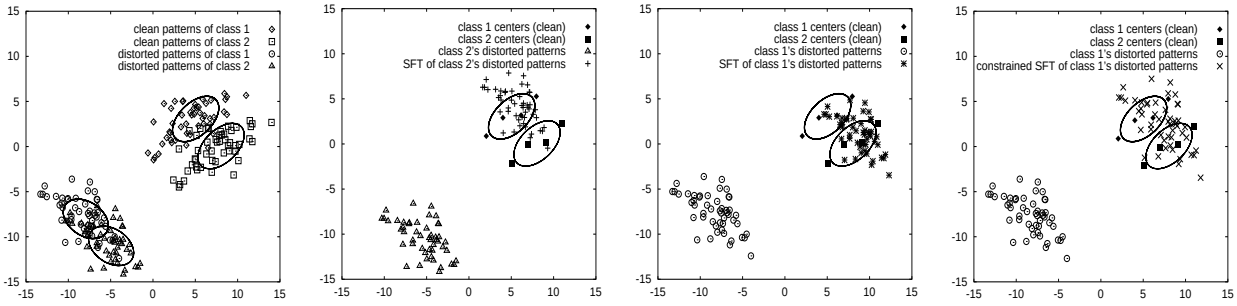
The constrained SFT was applied to the one-speaker detection task specified in the 2001 NIST speaker recognition evaluation set [10]. The evaluation set contains cellular phone speech extracted from the SwitchBoard-II, Phase IV Corpus. The evaluation includes 74 male and 100 female target speakers. For each speaker, approximately 2 minutes of speech is available for enrollment. There are 850 male and 1188 female verification utterances. Each verification utterance has length not exceeding 60 seconds and is evaluated against 11 hypothesized speakers of the same sex as the speaker of the verification utterance. Out of these 11 hypothesized speakers, one is the target speaker who produced the verification utterance. Therefore, there are one target and 10 impostor trials for each verification utterance, which amount to a total of 2,038 target trials and 20,380 impostor attempts for 2,038 verification utterances.

Detection Error Trade-off (DET) plots, which show the trade-off between miss probability and false acceptance, were used to present the experimental results. From these plots, systemwise equal error rates (EERs), at which the chance of false acceptance is equal to that of false rejection, were also obtained. Performance was evaluated based on all speaker scores and impostor scores, which were obtained by pooling all scores of both sex from the speaker and impostor trials respectively. In addition to DET plots and EERs, a minimum decision cost function (DCF), defined as the weighted sum of the miss and false alarm error probabilities, was also used as a performance measure. The DCF is defined as

$$\begin{aligned}\text{DCF} = {} & C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ & + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \\ & \times P_{NonTarget}\end{aligned}$$

where P_{Target} and $P_{NonTarget}$ are respectively the *a priori* probability of target and impostor speakers, and where C_{Miss} and $C_{FalseAlarm}$ are respectively the costs of miss and false alarm errors. Following the NIST recommendation [11], we set $P_{Target} = 0.01$, $P_{NonTarget} = 0.99$, $C_{Miss} = 10$ and $C_{FalseAlarm} = 1$.

Mel-frequency cepstral coefficients (MFCCs) [12] and their first-order derivatives were computed every 14ms using a Hamming window of 28ms. Cepstral mean subtraction (CMS) [13] was applied to the MFCCs to remove linear channel effects.



(a) 2-class problem

(b) Distorted class 2 data adapted to class 1

(c) Distorted class 1 data adapted to class 2

(d) constrained SFT

Figure 2: *(a) Scatter plots of the clean and distorted patterns corresponding to Class 1 and Class 2 in a two-class problem. The thick and thin ellipses represent the equal density contours of Class 1 and Class 2, respectively. The upper right (resp. lower left) clusters contain the clean (resp. distorted) patterns. (b) Distorted patterns of Class 2 were blindly transformed to fit Class 1's clean model using SFT. (c) Distorted patterns of Class 1 were blindly transformed to fit Class 2's clean model using SFT. (d) Distorted Class 1 data were transformed to fit the clean models of both Class 1 and Class 2 using first-order constrained SFT. For clarity, only the distorted patterns before and after adaptation were plotted in (b)–(d).*

The MFCCs and delta MFCCs were concatenated to form 24-dimensional feature vectors.

As mentioned in Section 2, both the speaker and background models are Gaussian mixture models (GMMs). During training, a 1024-component UBM was trained using the training utterances of all 174 target speakers. Then, for each target speaker, a speaker-dependent GMM was created by adapting the UBM using maximum a posteriori (MAP) adaptation [9].

During verification, the feature sequence Y obtained from a claimant was transformed by the feature transformation parameters $f_\nu(\cdot)$ to form a sequence of transformed vectors Y_ν. The transformed vectors were then fed to a 1024-center GMM speaker model (Λ_s^N) and the 1024-center UBM (Λ_b^N) to obtain the score

$$S(Y_\nu) = \log p(Y_\nu|\Lambda_s^N) - \log p(Y_\nu|\Lambda_b^N).$$

$S(Y_\nu)$ was compared with a global, speaker-independent threshold for decision making. In this work, the threshold was adjusted to determine an equal error rate (EER).

4. Results and Discussions

Fig. 3 and Table 1 show the results of different environment adaptation approaches, including cepstral mean subtraction (CMS), Znorm [6] and the constrained stochastic feature transformation (SFT) with different order and numbers of components (M). Evidently, all cases of SFT show significant reduction in error rates when compared to CMS. The DET curves also show that the constrained SFT performs better than CMS in all operating points. In particular, first-order constrained SFT with 64 components achieves the largest error reduction.

It is of interest to compare the proposed method with the short-time Gaussianization approach proposed in [1] because both methods attempt to transform distorted features in the feature space and their transformation parameters are determined by the EM algorithm. Short-time Gaussianization achieves an EER of 10.84% in the NIST 2001 evaluation set, whereas constrained SFT achieves an EER of 10.00%, which represent an error reduction of 8.40%. The minimum decision cost of constrainted SFT is also slightly lower than that of short-time Gaussianization (0.0428 vs. 0.0440).

Adaptation Method	**order**	**M**	**EER (%)**	**min. DCF**
CMS	NA	NA	12.02	0.0477
SFT	0	2+2	11.90	0.0473
SFT	1	2+2	12.00	0.0506
SFT	0	4+4	11.82	0.0458
SFT	1	4+4	11.55	0.0471
SFT	0	8+8	11.39	0.0449
SFT	1	8+8	10.70	0.0464
SFT	0	16+16	11.24	0.0450
SFT	1	16+16	10.47	0.0454
SFT	0	32+32	11.22	0.0450
SFT	1	32+32	10.43	0.0446
SFT	0	64+64	11.16	0.0443
SFT	1	64+64	10.00	0.0428
Znorm	NA	NA	10.39	0.0447

Table 1: *Equal error rates (in %) and minimum decision cost achieved by cepstral mean subtraction (CMS), Znorm, and zeroth- and first-order constrained stochastic feature transformation (SFT) with different order and numbers of components (M). In the third column, $M + M$ means the composite GMM (Λ_c^{2M}) was created by combining a reduced size speaker model and a reduced size background model, both with M components. Note that the number of components in the full-size speaker and background models is 1024.*

In constrained SFT, a set of transformation parameters ν is computed by the EM algorithm in which the likelihood of a composited GMM given the transformed test data is maximized. In short-time Gaussianization, a linear, global transformation matrix is estimated by the EM algorithm using the training data from all background speakers. The global transformation aims to decorrelate the features in the new feature space, and is applied to the distorted features before they are mapped to fit a normal distribution. The linearly transformed features are divided into a number of overlapping segments, with each segment containing a number of consecutive transformed vectors. The consecutive vectors in a segment are then sorted in ascending order. The rank of the central frame is used to find

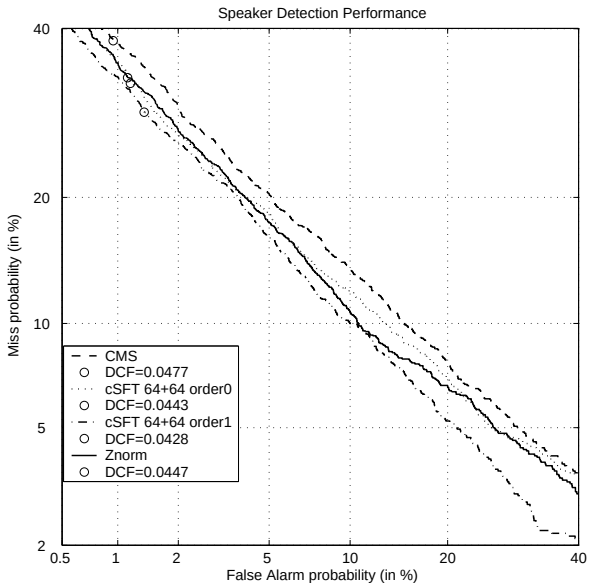


Figure 3: *DET curves comparing speaker verification performance using cepstral mean subtraction (dashed curve), Znorm (solid curve) and stochastic feature transformation with M=64 (64+64 in Table 1). Dotted curve: 0th-order; dashdot curve: 1st-order. The circles represent the errors at which minimum DCF occurs.*

a warped feature so that its cumulative density function (CDF) matches the CDF of a standard normal distribution. Theoretically, short-time Gaussianization is more computationally intensive than constrained SFT during the training phase because it uses all training data to estimate the global transformation matrix.

However, short-time Gaussianization is more efficient during the verification phase. In [1], the global transformation matrix was estimated based on speaker-independent data. The matrix was then applied to transform the features of all speakers. This can greatly reduce the verification time as the global transformation matrix can be estimated offline. On the other hand, in constrained SFT, it is necessary to estimate one set of transformation parameters for each test utterance. Fortunately, the estimation is computationally cheap because unlike short-time Gaussianization the transformation matrix in constrained SFT is diagonal. Additional computation saving can also be obtained by a light-weight approach as suggested in this work.

5. Conclusions

We have presented a new approach to channel robust speaker verification and provided experimental results based on the 2001 NIST evaluation set. The proposed algorithm computes the feature transformation parameters based on the statistical difference between a test utterance and a composite GMM formed by combining the speaker and background models. The transformation is then used to transform the test utterance to fit the clean speaker model and background model before verification. Experimental results show that the proposed algorithms achieves significant improvement in both equal error rate and minimum detection cost when compared to cepstral mean subtraction, Z-norm and short-time Gaussianization proposed in [1].

6. References

[1] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. IEEE ICASSP'02*, 2002, vol. 1, pp. 681–684.

[2] A. C. Surendran, C. H. Lee, and M. Rahim, "Nonlinear compensation for stochastic matching," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 6, pp. 643–655, 1999.

[3] M. W. Mak and S. Y. Kung, "Combining stochastic feautre transformation and handset identification for telephone-based speaker verification," in *Proc. ICASSP'02*, 2002, pp. I701–I704.

[4] F. Beaufays and M. Weintraub, "Model transformation for robust speaker recognition from telephone data," in *ICASSP'97*, 1997, vol. 2, pp. 1063–1066.

[5] K. K. Yiu, M. W. Mak, and S. Y. Kung, "Environment adaptation for robust speaker verification," in *Eurospeech'03*, 2003, pp. 2973–2976.

[6] D. A. Reynolds, "Comparison of background normalization methods for text independent speaker verification," in *Eurospeech'97*, 1997, pp. 963–966.

[7] C. L. Tsang, M. W. Mak, and S. Y. Kung, "Divergence-based out-of-class rejection for telephone handset identification," in *Proc. Int. Conf. on Spoken Language Processing*, 2002, pp. 2329–2332.

[8] M.W. Mak, C.L. Tsang, and S.Y. Kung, "Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification," *J. on Applied Signal Processing*, vol. 4, pp. 452–465, 2004.

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

[10] "The NIST year 2001 speaker recognition evaluation plan," in *http://www.nist.gov/speech/tests/spk/2001/doc*.

[11] M. Przybocki A. Martin, "NIST's assessment of text independent speaker recognition performance 2002," in *The Advent of Biometircs on the Internet, A COST 275 Workshop*, Rome, Italy, Nov. 2002.

[12] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, no. 4, pp. 357–366, August 1980.

[13] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. ASSP-29, no. 2, pp. 254–272, 1981.