Blind Stochastic Feature Transformation for Channel Robust Speaker Verification*

K. K. Yiu, M. W. Mak, M. C. Cheung and S. Y. Kung

October 12, 2004

Abstract

To improve the reliability of telephone-based speaker verification systems, channel compensation is indispensable. However, it is also important to ensure that the channel compensation algorithms in these systems surpress channel variations and enhance interspeaker distinction. This paper addresses this problem by a blind feature-based transformation approach in which the transformation parameters are determined online without any a priori knowledge of channel characteristics. Specifically, a composite statistical model formed by the fusion of a speaker model and a background model is used to represent the characteristics of enrollment speech. Based on the difference between the claimant's speech and the composite model, a stochastic matching type of approach is proposed to transform the claimant's speech to a region close to the enrollment speech. Therefore, the algorithm can estimate the transformation online without the necessity of detecting the handset types. Experimental results based on the 2001 NIST evaluation set show that the proposed transformation approach achieves significant improvement in both equal error rate and minimum detection cost as compared to cepstral mean subtraction, Znorm, and short-time Gaussianization.

*K. K. Yiu, M. W. Mak and M.C. Cheung are with the Center for Multimedia Signal Processing, Dept. of Electronic & Information Engineering, The Hong Kong Polytechnic University. S. Y. Kung is with the Dept. of Electrical Engineering, Princeton University. Correspondence should be sent to Dr. M.W. Mak, Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong. Tel: (852)2766-6257. Fax: (852)2362-8439. Email: enmwmak@polyu.edu.hk

Keywords: speaker verification; feature transformation; blind channel compensation; acoustic mismatch.

1 Introduction

The accuracy of speaker recognition systems that enroll client speakers under one acoustic environment (e.g., using a close-talk microphone in offices) but verify claimants under another environment (e.g., using mobile phones on the street) 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 presents one of the major technological challenges faced by speaker recognition researchers today. One cause of the mismatched conditions is transducer mismatch. Transducer mismatch 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 the matched condition without the need of a large amount of training data.

Channel compensation can be applied in feature space, model space, or score space. Feature-based compensation [1, 2] transforms channel-distorted speech features to fit clean speaker models, whereas model-based compensation [3, 4] adapts or transforms the parameters of clean models to fit a new acoustic environment. On the other hand, score-based compensation [5–7] 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 the speech signal during verification [2]. 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 [8, 9], it is difficult to find a threshold for detecting unseen handsets. On the

other hand, unsupervised (blind) compensation does not assume any knowledge of the channel characteristics. In particular, it adapts speaker models or transforms speaker features to accommodate channel variations based on verification utterances only. Therefore, handset detectors are no longer required.

In speaker verification, it is important to ensure that channel variations are suppressed so that the interspeaker distinction can be enhanced. In particular, given a claimant's utterance recorded in an environment different from that during enrollment, one aims to transform the features of the utterance so that they become compatible with the enrollment environment. Therefore, it is not appropriate to transform the claimant's utterance either to fit the speaker model only or to fit the background model only because the former will result in an unacceptably high FAR (false acceptance rate) and the latter an excessive FRR (false rejection rate). This paper proposes a feature-based *blind* transformation approach to solving this problem. Specifically, a feature-based transformation is estimated based on the statistical difference between a test utterance and a composite acoustic model formed by combining the speaker and background models. The transformation is then used to transform the test utterance before verification. The transformation is blind in that it compensates the handset distortion without a priori information about the channel's characteristics. Hereafter, this transformation approach is referred to as blind stochastic feature transformation (BSFT).

The paper is organized as follows. In Section 2, the procedures for estimating the parameters of BSFT are detailed and the philosophy behind this transformation approach is explained. In Section 3, speaker verification experiments that demonstrate the advantage of BSFT over other channel compensation approaches are presented. Finally, a conclusion of the paper is provided in Section 5.

2 Blind Stochastic Feature Transformation

As discussed in the preceding section, one popular approach to compensating for handset distortion is to divide handsets into several broad categories according to the type of transducer (e.g., carbon button and electret). During operations, a handset selector is used to identify the most likely handset type from speech signals and handset distortion is compensated for based

on some a priori information about the identified type in the database. Although this method works well in landline phones, it may encounter difficulty in mobile handsets because they have a large number of categories, new handset models are frequently released, and models can become obsolete in a short time. Maintaining a handset database for storing the information of all possible handset models is a great challenge and updating the compensation algorithm whenever a new handset is released is also difficult. Therefore, it is imperative to develop a channel compensation method that does not necessarily require a priori knowledge of handsets. This section describes a blind compensation algorithm for this problem. The algorithm is designed to handle the situation in which no a priori knowledge about the channel is available (i.e., a handset model not in the handset database is being used). Because the algorithm does not require a handset selector, it is suitable for a broader scale of deployment than the conventional approaches.

2.1 Estimation of Transformation Parameters

Figure 1 illustrates a speaker verification system with BSFT, whose operations are divided into two separate phases: enrollment and verification.

1. *Enrollment Phase.* The speech of all client speakers are used to create a compact universal background model (UBM) Λ_b^M with M components. Then, for each client speaker, a compact speaker model Λ_s^M is created by adapting the UBM Λ_b^M using maximum a posteriori (MAP) adaptation [10]. Because verification decisions are based on the likelihood of the speaker model and background model, both models must be considered when the transformation parameters are computed. This can be achieved by fusing Λ_b^M and Λ_s^M to form a $2M$-component composite GMM Λ_c^{2M}. During the fusion process, the means and covariances remain unchanged but the value of each mixing coefficient is divided by 2. This step ensures that the output of the composite GMM represents a probability density function.

2. *Verification Phase.* Distorted features $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ extracted from a verification utterance are used to compute the transformation parameters $\nu = \{A, \mathbf{b}\}$. This is achieved by maximizing the likelihood of the composite GMM Λ_c^{2M} given the transformed features

$\hat{X} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_T\}$:

$$\hat{\mathbf{x}}_t = f_\nu(\mathbf{y}_t) = A\mathbf{y}_t + \mathbf{b}, \quad t = 1, \ldots, T, \tag{1}$$

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 $\mathbf{b}$ is a bias vector. The transformed vectors $\hat{X}$ are then fed to a full size speaker model Λ_s^N and a full size UBM Λ_b^N for computing verification scores in terms of likelihood ratio:

$$s(\hat{X}) = \log p(\hat{X}|\Lambda_s^N) - \log p(\hat{X}|\Lambda_b^N).$$

The transformation parameters $\nu = \{A, \mathbf{b}\}$ can be estimated by the EM algorithm. More specifically, given the current estimate $\nu' = \{A', \mathbf{b}'\}$, we compute

$$\nu = \arg\max_\nu Q(\nu|\nu') = \arg\max_\nu \sum_{t=1}^{T} \sum_{j=1}^{2M} h_j(f_{\nu'}(\mathbf{y}_t)) \log \left\{\omega_{c,j} p(f_\nu(\mathbf{y}_t)|\mu_{c,j}, \Sigma_{c,j}, \nu) \, |J_\nu(\mathbf{y}_t)|\right\},$$

where $\{\omega_{c,j}, \mu_{c,j}, \Sigma_{c,j}\}_{j=1}^{2M}$ are the parameters of Λ_c^{2M}, $h_j(f_{\nu'}(\mathbf{y}_t))$ is the posterior probability

$$h_j(f_{\nu'}(\mathbf{y}_t)) = P(j|\mathbf{y}_t, \Lambda_c^{2M}, \nu') = \frac{\omega_{c,j} p(f_{\nu'}(\mathbf{y}_t)|\mu_{c,j}, \Sigma_{c,j})}{\sum_{l=1}^{2M} \omega_{c,l} p(f_{\nu'}(\mathbf{y}_t)|\mu_{c,l}, \Sigma_{c,l})},$$

and $|J_\nu(\mathbf{y}_t)|$ is the determinant of a Jacobian matrix with (r, s)-th entry given by $J_\nu(\mathbf{y}_t)_{rs} = \partial f_\nu(\mathbf{y}_t)_s / \partial y_{t,r}$.

The main idea of BSFT is to transform the distorted features to fit the composite GMM Λ_c^{2M}, which ensures that the transformation compensates the acoustic distortion.

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, BSFT will become computationally intensive when the number of components is large. To perform rapid adaptation, we propose 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, we synthesize a compact composite GMM (Λ_c^{2M}) by fusing a compact speaker GMM (Λ_s^M) and a compact background GMM (Λ_b^M), both with M components where $M \ll N$. It was found that a good trade-off between performance and computation complexity can be maintained by using a suitable value of M.

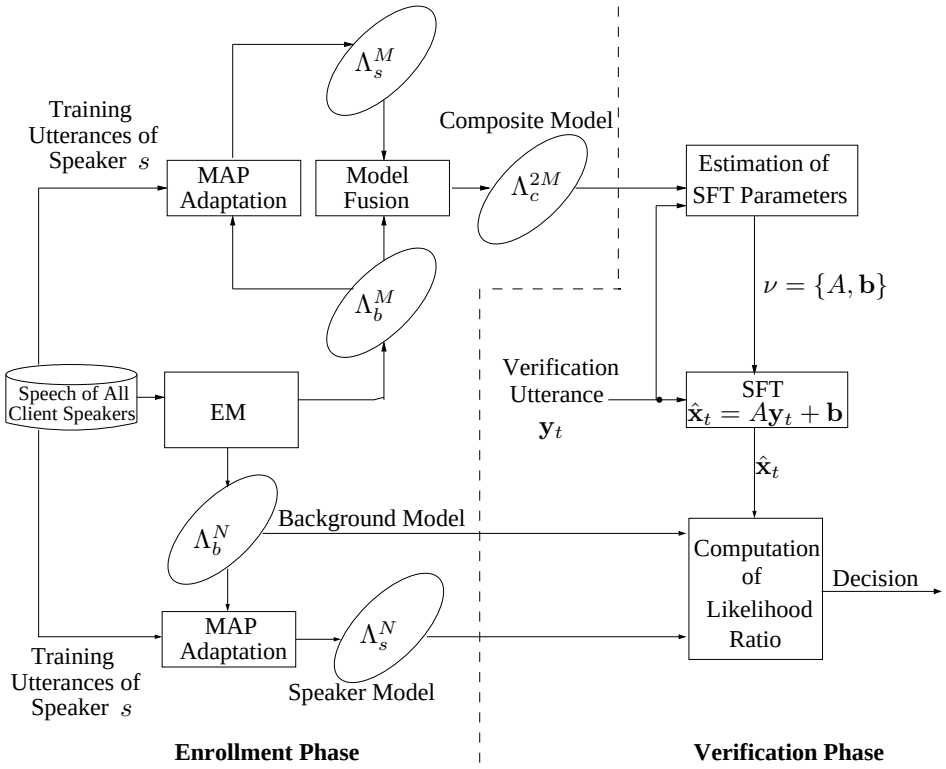

Figure 1: Estimation of BSFT parameters. The background model Λ_b^N, speaker model Λ_s^N, and composite model Λ_c^{2M}, produced during the enrollment phase, are subsequently used for verification purposes.

2.2 A Two-Dimensional Example

Figure 2 illustrates the idea of BSFT in a classification problem with two-dimensional input patterns. Figure 2(a) plots the clean and distorted patterns of Class 1 and Class 2. The upper right (respectively, lower left) clusters represent the clean (respectively, distorted) patterns. The ellipses show the corresponding equal density contours. Markers '♦' and '■' represent the centers of the clean models. Figure 2(b) illustrates a transformation matching the distorted data of Class 2 and the GMM of Class 1 (GMM1). Because the transformation only takes GMM1 into account, while ignoring GMM2 completely, it results in a high error rate. Similarly, the transformation in Figure 2(c) also has a high error rate. The transformation in Figure 2(d) was estimated from the distorted data of Class 1 and a composite GMM formed by fusing GMM1 and GMM2. In this case, the transformation adapts the data to a region close to both GMM1 and GMM2,

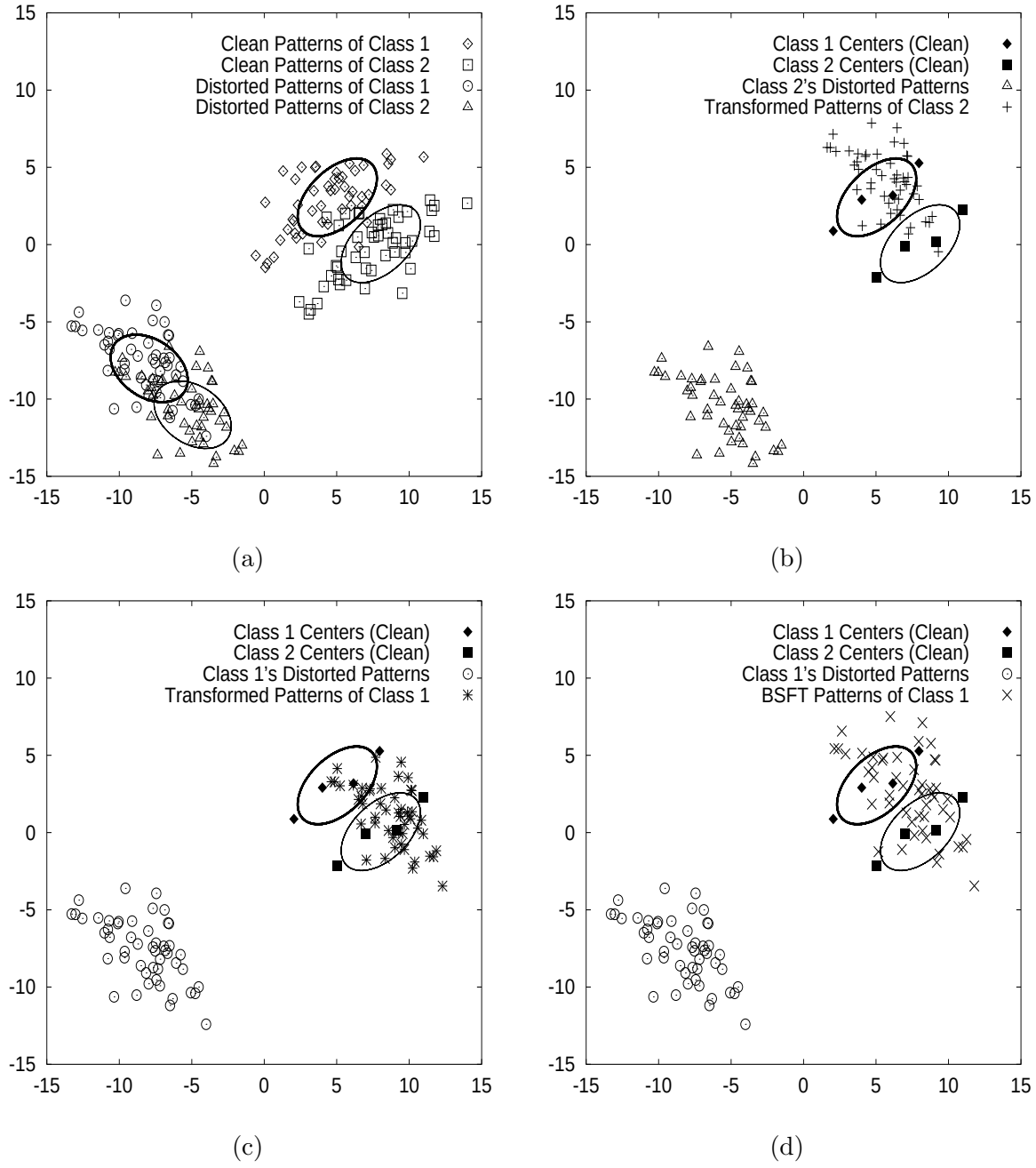

Figure 2: A Two-class problem illustrating the idea of BSFT. (a) Scatter plots of the clean and distorted patterns corresponding to Class 1 and Class 2. The thick and thin ellipses represent the equal density contours of Class 1 and Class 2, respectively. The upper right (respectively, lower left) clusters contain the clean (respectively, distorted) patterns. (b) Distorted patterns of Class 2 were transformed to fit Class 1's clean model. (c) Reversely, distorted patterns of Class 1 were transformed to fit Class 2's clean model. (d) Distorted data of Class 1 were transformed to fit the clean models of both Class 1 and Class 2 using first-order BSFT. For clarity, only the distorted patterns before and after transformation were plotted in (b) through (d).

because it takes both GMMs into account. Therefore, instead of transforming the distorted data to a region around GMM1 or GMM2 as in Figures 2(b) and 2(c), the transformation in Figure 2(d) attempts to compensate the distortion. The capability of BSFT is also demonstrated in a speaker verification task to be described next.

3 Experimental Evaluations

3.1 Enrollment and Verification

Per discussion earlier, the experiments were divided into two phases: enrollment and verification.

1. *Enrollment Phase.* A 1,024-component UBM Λ_b^{1024} (i.e., N = 1,024 in Figure 1) was trained using the training utterances of all target speakers. The same set of data was also used to train an M-component UBM (Λ_b^M in Figure 1). For each target speaker, a 1,024-component speaker-dependent GMM Λ_s^{1024} was created by adapting Λ_b^{1024} using MAP adaptation [10]. Similarly, Λ_s^M was created by adapting Λ_b^M, and the two models were fused to form a composite GMM Λ_c^{2M}. The value of M was varied from 2 to 64 in the experiments.

2. *Verification Phase.* For each verification session, a feature sequence Y was extracted from the utterance of a claimant. The sequence was used to determined the BSFT parameters (A and $\mathbf{b}$ in Eq. 1) to obtain a sequence of transformed vectors $\hat{X}$. The transformed vectors were then fed to Λ_s^{1024} and Λ_b^{1024} to obtain verification scores for decision making.

3.2 Speech Data and Features

The 2001 NIST speaker recognition evaluation set [11], which contains cellular phone speech of 74 male and 100 female target speakers extracted from the SwitchBoard-II Phase IV Corpus, was used in the evaluation. Each target speaker has 2 minutes of speech for training (i.e., enrollment); a total of 850 male and 1,188 female utterances are available for testing (i.e., verification). Each verification utterance has a length of between 15 and 45 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 amounts to a total of 2,038 target trials and 20,380 impostor attempts for 2,038 verification utterances.

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

3.3 Performance Measures

Detection error trade-off (DET) curves and equal error rates (EERs) were used as performance measures. They were obtained by pooling all scores of both sex from the speaker and impostor trials. In addition to DET curves and EERs, decision cost function (DCF) was also used as 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 the prior probability of target and nontarget speakers, respectively, and where C_{Miss} and $C_{FalseAlarm}$ are the costs of miss and false alarm errors, respectively. Following NIST's recommendation [14], these parameters were set as follows: $P_{Target} = 0.01$, $P_{Nontarget} = 0.99$, $C_{Miss} = 10$, and $C_{FalseAlarm} = 1$.

4 Results and Discussions

4.1 Verification Performance

Figure 3 and Table 1 show the results of the baseline (CMS only), Znorm [5], and BSFT with different order and number of components M.[1] Evidently, all cases of BSFT show significant

[1]Theoretically, the larger the value of M, the better the results. However, setting M larger than 64 will result in unacceptably long verification time.

reduction in error rates when compared to the baseline. In particular, Table 1 shows that first-order BSFT with Znorm achieves the largest error reduction. The DET curves also show that BSFT with Znorm performs better than the baseline and Znorm alone for all operating points.

Because the evaluation trials in NIST01 are gender-matched, gender-dependent background models can also be used for enrollment and estimation of BSFT parameters. In another experiment, speaker models were adapted from gender-dependent background models using MAP adaptation. A compact gender-dependent background model (with 64 components) was used to estimate the BSFT parameters. As shown in Table 1 and Figure 4, using gender-dependent background model helps to reduce the EERs and minimum DCF further for all cases of BSFT. However, Znorm and BSFT with Znorm seem to perform better when the background model is gender-independent. This may be attribute to the fact that less data are available for determining the Znorm parameters (score mean and variance) for each speaker when gender-dependent background models were used, which results in less reliable Znorm scores for verification. For the gender-independent case, the training utterances of 60 speakers from the "devtest" section of NIST2001 were used for estimating the Znorm parameters. For the gender-dependent case, however, the Znorm parameters of each speaker were estimated from the respective gender of these 60 speakers. Among these 60 speakers, 38 are male and 22 are female, and each of them has one training utterance. As a results, the Znorm parameters of the female speakers were determined by 22 utterances only.

4.2 Comparison with Other Models

It is of interest to compare BSFT with the short-time Gaussianization approach proposed in Xiang et al. [15] because both methods transform distorted features in the feature space and their transformation parameters are estimated by the EM algorithm [16]. The short-time Gaussianization achieves an EER of 10.84% in the NIST 2001 evaluation set [15], whereas BSFT achieves an EER of 9.26%, which represent an error reduction of 14.58%.[2] The minimum decision cost of BSFT is also lower than that of short-time Gaussianization (0.0384 versus 0.0440).

[2]Because Xiang et al. did not use Znorm in [15], their results should be compared with the one without Znorm here.

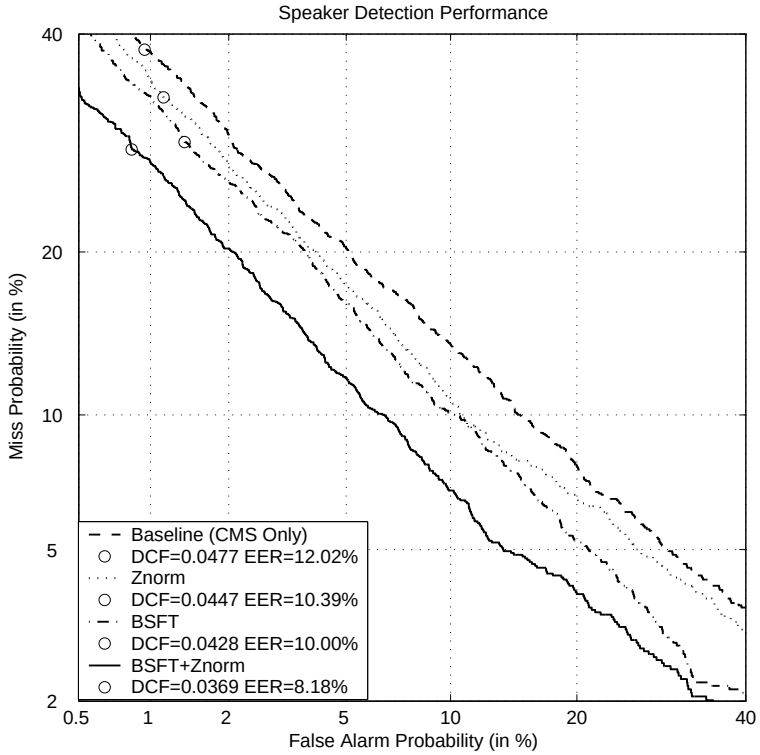

Figure 3: *DET curves comparing speaker verification performance using CMS (dashed), Znorm (dotted), first-order BSFT (dash-dot), and first-order BSFT with Znorm (solid). For BSFT, the number of components M in the compact GMMs was set to 64. The circles represent the errors at which minimum decision costs occur. A gender-independent background model was used in all cases.*

4.3 Computation Consideration

In BSFT, a set of transformation parameters ν is computed by the EM algorithm in which the likelihood function of a composite GMM given the transformed test data is maximized. In short-time Gaussianization, a linear, global transformation matrix, which aims to decorrelate the distorted features, is estimated by the EM algorithm using the training data of all background speakers. The distorted features are then transformed and 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 a warped feature so that its cumulative density function (CDF) matches the CDF of a standard normal distribution.

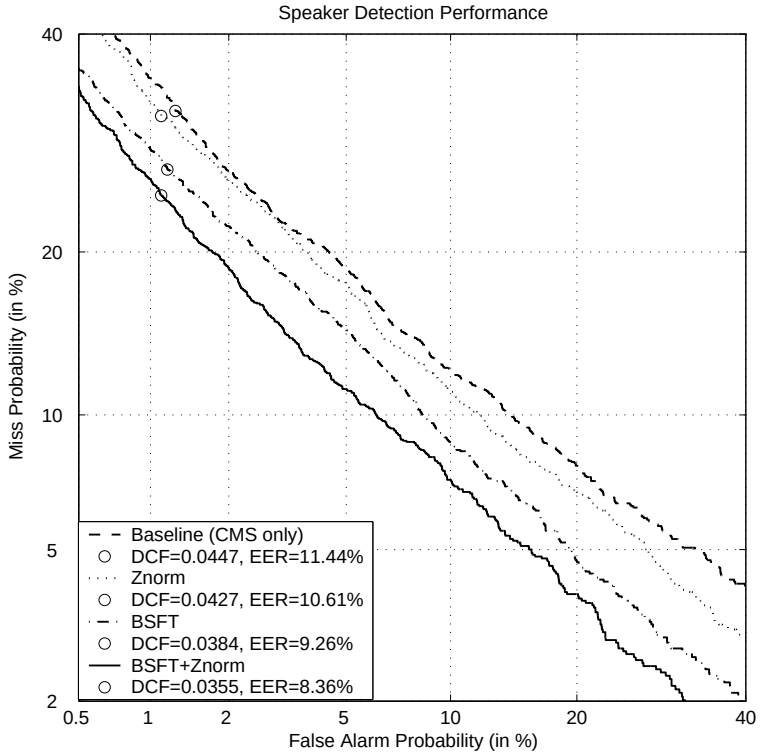

Figure 4: *DET curves comparing speaker verification performance using CMS (dashed), Znorm (dotted), first-order BSFT (dash-dot), and first-order BSFT with Znorm (solid). For BSFT, the number of components M in the compact GMMs was set to 64. The circles represent the errors at which minimum decision costs occur. Gender-dependent background models were used in all cases.*

It can be argued that the inferior performance of Gaussianization is due to the nonadaptive nature of its transformation parameters. However, the adaptive nature of BSFT comes with a computational price: different transformation parameters have to be computed for each speaker. Therefore, it is vital to have a cost effective computation approach for BSFT. Note that the computation complexity of estimating BSFT parameters grows with the amount of adaptation data (i.e., the value of T in Eq. 1) and the number of mixture components in the compact GMMs (i.e., the value of M). To reduce computation time, M should be significantly smaller than N, the number of components in the full size speaker and background models. This is particularly important for the computation of BSFT parameters during the verification phase because the computation time of this phase is a significant part of the overall verification time. The evaluations suggest that a good tradeoff between performance and computation complexity can be achieved by using a suitable value of M.

5 Conclusions

We have presented a new approach, namely blind stochastic feature transformation, to channel robust speaker verification and provided experimental results on the 2001 NIST evaluation set. The algorithm computes 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, Znorm, and short-time Gaussianization.

References

[1] 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.

[2] 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.

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

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

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

[6] 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.

[7] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *ICASSP-88*, 1988, vol. 1, pp. 595–598.

[8] 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.

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

[10] 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.

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

[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.

[14] M. Przybocki and 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.

[15] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. IEEE ICASSP02*, 2002, vol. 1, pp. 681–684.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of Royal Statistical Soc., Ser. B.*, vol. 39, no. 1, pp. 1–38, 1977.

Compensation Method	SFT Order	M	Background Model (Λ_b^N)			
			Gender-Independent		Gender-Dependent	
			Equal Error Rate (%)	Minimum Decision Cost	Equal Error Rate (%)	Minimum Decision Cost
Baseline	NA	NA	12.02	0.0477	11.44	0.0477
BSFT	Zeroth	2	11.90	0.0473	11.49	0.0440
BSFT	Zeroth	4	11.82	0.0458	11.16	0.0427
BSFT	Zeroth	8	11.39	0.0449	10.89	0.0428
BSFT	Zeroth	16	11.24	0.0450	10.79	0.0420
BSFT	Zeroth	32	11.22	0.0450	10.80	0.0422
BSFT	Zeroth	64	11.16	0.0443	10.61	0.0414
BSFT	First	2	12.00	0.0506	11.29	0.0445
BSFT	First	4	11.55	0.0471	10.27	0.0425
BSFT	First	8	10.70	0.0464	9.77	0.0409
BSFT	First	16	10.47	0.0454	9.48	0.0394
BSFT	First	32	10.43	0.0446	9.38	0.0395
BSFT	First	64	10.00	0.0428	9.26	0.0384
Znorm	NA	NA	10.39	0.0447	10.61	0.0427
BSFT+Znorm	First	64	8.18	0.0369	8.36	0.0355

Table 1: Equal error rates and minimum decision cost achieved by the baseline (CMS only), Znorm, and zeroth- and first-order BSFT with different order and number of components M in the compact GMMs. The number of components in the full size speaker and background models is 1,024. The columns "Gender-Independent" and "Gender-Dependent" represents the types of background models being used for obtaining the results.