Multi-Sample Fusion with Constrained Feature Transformation for Robust Speaker Verification

Ming Cheung Cheung, Kwok Kwong Yiu, Man Wai Mak

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 single-source multi-sample fusion approach to text-independent speaker verification. In conventional speaker verification systems, the scores obtained from claimant's utterances are averaged and the resulting mean score is used for decision making. Instead of using an equal weight for all scores, this paper proposes assigning a different weight to each score, where the weights are made dependent on the difference between the score values and a speaker-dependent reference score obtained during enrollment. Because the fusion weights depend on the verification scores, a technique called constrained stochastic feature transformation is applied to minimize the mismatch between enrollment and verification data in order to enhance the scores' reliability. Experimental results based on the 2001 NIST evaluation set show that the proposed fusion approach outperforms the equal-weight approach by 22% in terms of equal error rate and 16% in terms of minimum detection cost.

1. Introduction

Data fusion techniques have been used in many multi-modal biometric authentication systems to enhance system reliability [1]. While the fusion techniques are originally designed for fusing the scores of multiple modalities, they can easily be adapted for fusing the decisions or scores of a single modality. The idea is to consider the multiple samples extracted from a single modality as independent but coming from the same source. From the perspective of applications, single-source multi-sample (SSMS) fusion will not impose any burden on users because a single, long sample (e.g. an utterance or video shot) can always be divided into a number of short samples. The implementation cost of single-modal systems is also lower than that of multi-modal systems because only one sensor is required. Typically, SSMS fusion involves averaging the scores obtained from multiple samples (e.g., [2]). However, this approach is equivalent to the single-source single-sample case because the fusion weights are equal for all scores. The benefit of fusing multiple samples arises when the weights for individual scores from multiple samples are different.

To overcome the limitation of the equal-weight approach, we have recently proposed a fusion model [3][4] in which the fusion weights are dependent on the dispersion between the frame-based verification scores and the prior score statistics obtained from training data. As fusion weights depend on verification scores, the scores' reliability is detrimental to verification performance. One of the factors affecting the scores' reliability is the mismatch between the training and verification conditions. In [3], this problem is addressed by using a combination of handset detection and feature transformation. This handset-dependent transformation technique, however, assumes that the handset characteristics are known a priori, which may not be realistic in practical situations. To overcome this limitation, this paper proposes integrating a handset-independent transformation technique, namely constrained stochastic feature transformation (SFT)[5], into multi-sample fusion. In this transformation approach, channel-distorted test utterances are transformed to fit the clean speaker and background models before verification takes place, and the transformation is based on the statistical difference between the test-utterances and a composite Gaussian mixture model (GMM) formed by combining the client and background GMMs.

The constrained SFT is similar to feature mapping proposed by Reynolds [6] in that they both work in feature space. However, they also have two important differences. First, feature mapping is a supervised technique in that handset labels are required, while constrained SFT is an unsupervised technique in that it learns the statistical difference between the verification utterance and a composite GMM formed by a speaker and background model without any channel information. Second, feature mapping aims to map the features into a channel-independent space, whereas no such space is defined in constrained SFT.

The remainder of the paper is organized as follows. The data-dependent decision fusion for multi-sample speaker verification proposed in [3] is briefly reviewed

This work was supported by the Research Grant Council of Hong Kong SAR (Project Nos. PolyU 5131/02E and CUHK 1/02C).

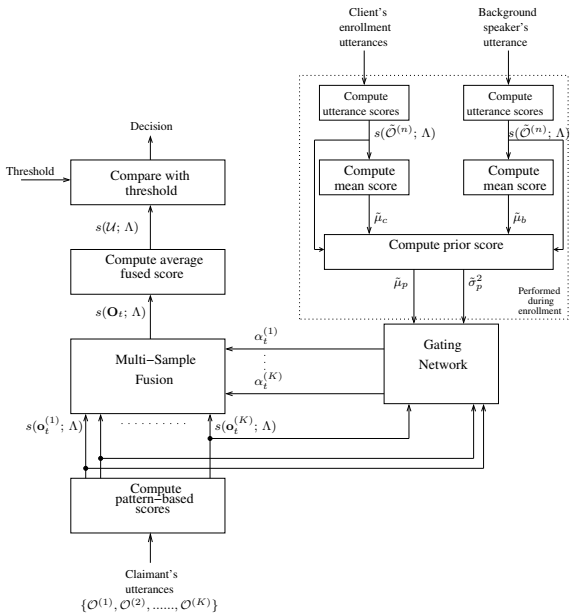

Figure 1: Architecture of the proposed speaker verification system.

in Section 2. This is followed by a brief explanation of constrained SFT in Section 3. The proposed method is further evaluated in Section 4 using the 2001 NIST evaluation set. Finally, in Section 5, concluding remarks are provided.

2. Multi-Sample Decision Fusion

Assume that K streams of speech vectors (e.g. MFCCs) can be extracted from K utterances $\mathcal{U} = \{\mathcal{U}_1, \ldots, \mathcal{U}_K\}$. Let us denote the observation sequence corresponding to utterance $\mathcal{U}_k$ by

$$\mathcal{O}^{(k)} = \{\mathbf{o}_t^{(k)} \in \Re^D; t = 1, \ldots, T_k\} \quad k = 1, \ldots, K \tag{1}$$

where D and T_k are respectively the dimensionality of $\mathbf{o}_t^{(k)}$ and the number of observations in $\mathcal{O}^{(k)}$, and t is the frame index. To simplify notation, let us assume that the K utterances contain the same number of feature vectors, i.e., $T_1 = T_2 = \cdots = T_K$. If it is not the case, we may append the tail of the longer utterances to the shorter ones to make the number of feature vectors equal.[1] We further define a normalized score function [7]

$$s(\mathbf{o}_t^{(k)}; \Lambda) = \log p(\mathbf{o}_t^{(k)}|\Lambda_{\omega_c}) - \log p(\mathbf{o}_t^{(k)}|\Lambda_{\omega_b}) \tag{2}$$

where $\Lambda = \{\Lambda_{\omega_c}, \Lambda_{\omega_b}\}$ contains the GMMs that characterize the client speaker (ω_c) and background speakers (ω_b), and $\log p(\mathbf{o}_t^{(k)}|\Lambda_\omega)$ is the output of GMM Λ_ω, $\omega \in \{\omega_c, \omega_b\}$, given observation $\mathbf{o}_t^{(k)}$.

In [3], frame-level fused scores are computed as

$$s(\mathbf{o}_t^{(1)}, \ldots, \mathbf{o}_t^{(K)}; \Lambda) = s(\mathbf{O}_t; \Lambda) = \sum\nolimits_{k=1}^{K} \alpha_t^{(k)} s_t^{(k)} \tag{3}$$

where $t = 1, \ldots, T$, $\mathbf{O}_t = \{\mathbf{o}_t^{(1)}, \ldots, \mathbf{o}_t^{(K)}\}$ contains the K observations from the K utterances at frame t and $\alpha_t^{(k)} \in [0, 1]$ represents the confidence (reliability) of the observation $\mathbf{o}_t^{(k)}$. Then, the mean fused score

$$s(\mathcal{U}; \Lambda) = \frac{1}{T} \sum\nolimits_{t=1}^{T} s(\mathbf{O}_t; \Lambda) \tag{4}$$

is compared against a decision threshold for decision making. By imposing different constraints on the values of $\alpha_t^{(k)}$, we can obtain two fusion models, namely equal-weight fusion ([2], which is our baseline) and zero-sum fusion:

- **equal-weight fusion:** $\alpha_t^{(k)} = \frac{1}{K} \quad \forall\, t = 1, \ldots, T$ and $k = 1, \ldots, K$;
- **zero-sum fusion:** $\sum_{k=1}^{K} \alpha_t^{(k)} = 1 \;\; \forall\, t = 1, \ldots, T$.

Note that for zero-sum fusion, scores from different utterances *compete* with each other because the fusion weights from different utterances sum to one; whereas there is no competition among the scores in equal-weight fusion, as all weights are equal.

To use the prior information about the scores, the fusion weights $\alpha_t^{(k)}$ are made dependent on both the training data (prior information) and recognition data. Specifically, using enrollment data, the prior score $\tilde{\mu}_p$ and prior variance $\tilde{\sigma}_p^2$ are computed as follows:

$$\tilde{\mu}_p = \frac{K_c \tilde{\mu}_c + K_b \tilde{\mu}_b}{K_c + K_b}, \tag{5}$$

$$\tilde{\sigma}_p^2 = \frac{1}{K_c + K_b} \sum_{k=1}^{K_c + K_b} \Big[\tilde{s}(\mathcal{O}^{(k)}; \Lambda) - \tilde{\mu}_p\Big]^2 \tag{6}$$

where K_c and K_b are respectively the numbers of client speaker's utterances and background speakers' utterances, $\tilde{\mu}_c$ and $\tilde{\mu}_b$ are respectively the score means of client's and background speakers' utterances and $\tilde{s}(\mathcal{O}^{(k)}; \Lambda)$ denotes the mean score of the k-th utterance. Then, during verification, the claimant is asked to utter K utterances and the fusion weights are computed as

$$\alpha_t^{(k)} = \frac{\exp\{(s(\mathbf{o}_t^{(k)}) - \tilde{\mu}_p)^2 / 2\tilde{\sigma}_p^2\}}{\sum_{k=1}^{K} \exp\{(s(\mathbf{o}_t^{(k)}) - \tilde{\mu}_p)^2 / 2\tilde{\sigma}_p^2\}} \tag{7}$$

where $t = 1, \ldots, T$ and $k = 1, \ldots, K$. Fig. 1 illustrates the flow of the fusion algorithm.

[1] As it is likely that the utterances are obtained from the same speaker under the same environment in a verification session, moving feature vectors from utterances to utterances will have the same effect as partitioning a long utterance into a number of equal-length short utterances. In fact, the equal-weight approach concatenates several utterances into one and determines the mean score of the concatenated utterance. The idea is identical to moving the feature vectors among the utterances here.

3. Constrained Stochastic Feature Transformation

Because neither handset nor channel labels are available, an unsupervised approach to channel compensation was adopted [5]. In this approach, in addition to the regular N-component speaker and universal background models (UBM), where $N = 1024$, small speaker and UBMs with M-components (where typically $M = 64$) are also trained. During verification, the small speaker model (Λ_s^M) of the claimed identity is combined with an M-component UBM (Λ_b^M) to form a composite GMM (Λ_c^{2M}) with $2M$ components. During the combination, the means and covariances remain unchanged while the mixing coefficients are divided by two. This step ensures that the composite GMM represents a probability density function.

Another UBM (Λ_b^{2M}) of $2M$ components are trained using the training utterances of all client speakers. Then, for each test utterance, a testing GMM (Λ_t^{2M}) with $2M$ components is created by adapting the UBM (Λ_b^{2M}) using maximum a posteriori (MAP) adaptation [7]. Using the test utterance, Λ_c^{2M} and Λ_t^{2M}, a set of feature transformation parameters ν are computed based on the stochastic feature transformation technique [8]. In this work, We used first-order stochastic feature transformation with $K = 1$ in Eq. (2) of [8]. More specifically, given a distorted vector $\mathbf{x}$, the transformed feature vector is

$$\hat{\mathbf{x}} = f_\nu(\mathbf{x}) = A\mathbf{x} + \mathbf{b} \tag{8}$$

where $A = \text{diag}\{a_1, a_2, \ldots, a_D\}$ and $\mathbf{b}$ is a bias vector. The main idea is to transform the test data, which is modelled by the testing GMM Λ_t^{2M}, to fit the composite GMM Λ_c^{2M}. See [5] for the details of finding the transformation parameters ν.

4. Speaker Verification Experiments

To demonstrate the capability of the proposed fusion algorithm under practical situations, we have performed evaluations using the 2001 NIST speaker recognition evaluation set [9].

4.1. Speech Corpora and Feature Extraction

The 2001 NIST evaluation set contains cellular phone speech of 174 target speakers (clients) with 74 male and 100 female. Another set of 60 speakers is also available for development. Each target speaker has 2 minutes of speech for training, and the duration of test utterances varies from 15 to 45 seconds. The evaluation set provides a total of 20,380 gender-matched verification trials. The ratio between target and impostor trials is roughly 1:10.

In the experiments, 12 MFCCs and their delta coefficients were extracted from the utterances at a rate of 71Hz using a 28ms Hamming window. Cepstral mean subtraction (CMS) [10] was applied to all MFCCs before they were appended to the delta MFCCs.

4.2. Enrollment Procedures

A 1024-component UBM was created based on the speech of 174 client speakers. Then, for each client speaker, a speaker-dependent GMM was created by adapting the UBM using maximum a posteriori (MAP) adaptation [7]. The speaker-dependent prior scores $\tilde{\mu}_p$ and variances $\tilde{\sigma}_p^2$ in (7) were then estimated from the speech of 60 speakers in the development set of the corpus. This was achieved by considering these 60 speakers as pseudo-impostors and presenting their feature vectors to the speaker and background models to obtain a sequence of speaker-dependent pseudo-impostor scores. The utterances used for training the speaker models were also presented to the speaker and background models to obtain a sequence of client scores. The averages of these scores were then used to estimate the prior scores and variances as in (5) and (6). The dotted box in Fig. 1 illustrates the process of estimating the prior scores and variances.

4.3. Verification Procedures

For each verification session, the feature sequence Y obtained from a claimant's utterance was transformed by constrained transformation $f_\nu(\cdot)$ to form a sequence of transformed vectors Y_ν. First-order constrained SFT with 64 components ($M = 64$ in Section 3) was used. The transformed vectors were then fed to a 1024-component speaker model (Λ_s^{1024}) and a 1024-component UBM (Λ_b^{1024}) to obtain a sequence of normalized scores

$$S(Y_\nu) = \log p(Y_\nu|\Lambda_s^{1024}) - \log p(Y_\nu|\Lambda_b^{1024}).$$

The average of these scores was used for decision making. This traditional method (referred to as equal-weight approach) is the baseline for comparison.

The proposed fusion algorithm was applied to compute the weight of each score. Since there is only one test utterance in each verification session, the utterance was split into two segments and then fusion was performed. In this case, the length was calculated according to

$$T = \lfloor \frac{L}{2} \rfloor$$

where T is a positive integer and L represents the length of the test utterance.

4.4. Performance Measures

Detection error trade-off (DET) curves and detection cost function (DCF) were used as performance measures. The DCF is defined as follows:

$$\text{DCF} = C_{FA}Pr(FA|I)Pr(I) + C_{FR}Pr(FR|T)Pr(T) \tag{9}$$

where $Pr(I)$ and $Pr(T)$ are the prior probability of impostors and target speakers respectively, and where C_{FA} and C_{FR} are the costs of false alarm and false rejection

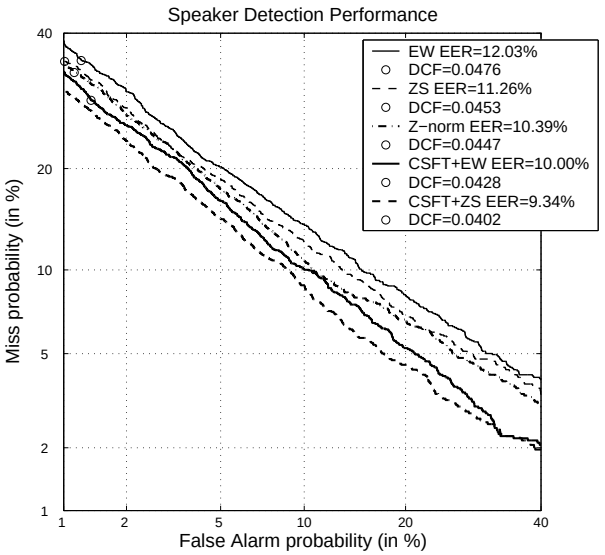

Figure 2: Speaker detection performance for zero-sum (ZS) fusion and equal-weight (EW) fusion. CSFT stands for constained stochastic feature transformation. For ease of comparison, the labels in the legend are arranged in decreasing EER.

respectively. In this work, $Pr(I) = 0.99$, $Pr(T) = 0.01$, $C_{FA} = 1$ and $C_{FR} = 10$ were set, as recommended by Przybocki and Martin [11].

4.5. Results and Discussions

Fig. 2 depicts the speaker detection performance of equal-weight fusion and zero-sum fusion. The figure shows that multi-sample decision fusion achieves lower error rates. In particular, the equal error rate (EER) achieved by zero-sum fusion (without constrained SFT) is 11.26%. When compared to equal-weight fusion (which achieves an EER of 12.03%), a relative error reduction of 6% was obtained. Zero-sum fusion also achieves a 5% improvement in terms of minimum DCF as compared with the baseline. Without constrained SFT, the proposed fusion algorithm is inferior to Z-norm, which attains 10.39% in EER and 0.0447 in minimum DCF. This may be due to the low reliability of verification scores. However, with constrained SFT, the performance of zero-sum fusion is significantly better than Z-norm. More specifically, the fusion system achieves an EER of 9.34%, which represents a 22% relative error reduction as compared to the baseline and 10% error reduction as compared with Z-norm. Regarding minimum DCF, a 16% and 10% improvement has been achieved as compared with the baseline and Z-norm, respectively.

Since NIST 2001 has a standard evaluation protocol, in addition to the baseline and Z-norm, we can also compare our results with the short-time Gaussianization approach proposed in [12]. The work in [12] focuses on feature-level processing to minimize channel distortions. Our approach, on the other hand, first reduces channel distortions at the feature level and then weighs more heavily on those useful scores at the score level. This two-level optimization approach helps reduce the error rate even further: 9.34% EER (our fusion approach) versus 10.84% EER (short-time Gaussianization). In addition, multi-sample fusion can achieve a slightly lower minimum detection cost (0.0402) than that of short-time Gaussianization (0.0440).

5. Conclusions

A decision fusion algorithm that makes use of prior score statistics and the distribution of the recognition data was presented. Constrained stochastic feature transformation was used to remove channel mismatch. The 2001 NIST evaluation set was used to evaluate the proposed speaker verification approach. Results show that the proposed fusion algorithm with constrained SFT can reduce equal error rate by 22%.

6. References

[1] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[2] N. Poh, S. Bengio, and J. Korczak, "A multi-sample multi-source model for biometric authentication," in *Proc. IEEE 12th Workshop on Neural Networks for Signal Processing*, 2002, pp. 375–384.

[3] M. W. Mak, M. C. Cheung, and S. Y. Kung, "Robust speaker verification from GSM-transcoded speech based on decision fusion and feature transformation," in *Proc. IEEE ICASSP'03*, 2003, pp. II745–II748.

[4] M. C. Cheung, M. W. Mak, and S. Y. Kung, "Multi-sample data-dependent fusion of sorted score sequences for biometric verification," in *Proc. IEEE ICASSP'04*, 2004, pp. V681–V684.

[5] K. K. Yiu, M. W. Mak, M. C. Cheung, and S. Y. Kung, "A new approach to channel robust speaker verification via constrained stochastic feature transformation," in *Proc. ICSLP'04*.

[6] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE ICASSP'03*, 2003, pp. II53–II56.

[7] 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.

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

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

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

[11] 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.

[12] 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.