Adaptive Decision Fusion for Multi-Sample Speaker Verification over GSM Networks

Ming-Cheung Cheung and 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

In speaker verification, a claimant may produce two or more utterances. In our previous study [1], we proposed to compute the optimal weights for fusing the scores of these utterances based on their score distribution and our prior knowledge about the score statistics estimated from the mean scores of the corresponding client speaker and some pseudo-impostors during enrollment. As the fusion weights depend on the prior scores, in this paper, we propose to adapt the prior scores during verification based on the likelihood of the claimant being an impostor. To this end, a pseudo-imposter GMM score model is created for each speaker. During verification, the claimant's scores are fed to the score model to obtain a likelihood for adapting the prior score. Experimental results based on the GSM-transcoded speech of 150 speakers from the HTIMIT corpus demonstrate that the proposed prior score adaptation approach provides a relative error reduction of 15% when compared with our previous approach where the prior scores are non-adaptive.

1. Introduction

Speaker verification is to verify a speaker's claimed identity based on his/her voice. A speaker claiming an identity is called a *claimant*, and an unregistered speaker posing as a registered speaker is an *impostor*. A speaker verification system should avoid rejecting registered speakers or accepting impostors as registered speakers, for the former will result in false rejection and the latter will lead to false acceptance.

Recently, there has been increasing interest in using fusion techniques to improve the performance of speaker verification systems. For example, in [2], different feature extraction methods (e.g. Mel Frequency Cepstral Coefficients (MFCC) and Maximum Auto-Correlation Values (MACV)) are applied to the same utterance in order to create multiple features. Fusion is done by applying different weights to the normalized scores (likelihood ratios) obtained from different features, and the final opinion is based on the linear weighted sum of these normalized scores. However, the fusion weights are optimal only for training data and kept fixed during verification. In [3], the scores of multiple utterances from a claimant are averaged, meaning that the fusion weights are equal and non-adaptive. In [1], fusion weights are adapted according to the verification scores and the prior scores statistics determined from enrollment data. Although the fusion weights in [1] are adaptive, the prior score determined from enrollment data may not be optimum for verification data.

In this work, we investigated the fusion of scores from multiple utterances to improve the performance of speaker verification from GSM-transcoded speech. In addition to computing the adaptive fusion weights as in [1], we also adapted the prior scores based on the score distribution of the verification utterances in order to further improve the system performance. As the variation of handset characteristics and the encoding/decoding process will introduce substantial distortion to the speech signals [4], we also applied stochastic feature transformation [5] to the feature vectors extracted from the GSM-transcoded speech before presenting them to the speaker models.

2. Multi-sample speaker verification

2.1. Data-dependent decision fusion

Assume that K streams of features vectors (e.g. MFCCs) can be extracted from K independent utterances $U = \{U_1, \ldots, U_K\}$. Let us denote the observation sequence corresponding to utterance U_k by

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

where D and T_k are respectively the dimensionality of $\mathbf{o}_t^{(k)}$ and the number of observations in $O^{(k)}$. We further define a normalized score function

$$s(\mathbf{o}_t^{(k)}; \Lambda) \equiv \log p(\mathbf{o}_t^{(k)}|\Lambda_{\omega_c}) - \log p(\mathbf{o}_t^{(k)}|\Lambda_{\omega_b}) \qquad (1)$$

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

In [1], frame-level fused scores are computed as

$$s(\mathbf{o}_t^{(1)}, \ldots, \mathbf{o}_t^{(K)}; \Lambda) = \sum_{k=1}^{K} \alpha_t^{(k)} s(\mathbf{o}_t^{(k)}; \Lambda) \qquad (2)$$

where the fusion weight $\alpha_t^{(k)} \in [0, 1]$ represents the reliability of the observation $\mathbf{o}_t^{(k)}$ and $\sum_k \alpha_t^{(k)} = 1$. The fusion weights are computed according to

$$\alpha_t^{(k)} = \frac{\exp\{(s_t^{(k)} - \tilde{\mu}_p)^2 / 2\tilde{\sigma}_p^2\}}{\sum_{l=1}^{K} \exp\{(s_t^{(l)} - \tilde{\mu}_p)^2 / 2\tilde{\sigma}_p^2\}} \qquad k = 1, \ldots, K \quad (3)$$

where $s_t^{(k)} \equiv s(\mathbf{o}_t^{(k)}; \Lambda)$ and the speaker-dependent prior score $\tilde{\mu}_p$ and the prior variance $\tilde{\sigma}_p^2$ are respectively equal to the score

This work was supported by The Hong Kong Polytechnic University, Grant No. A442 and the RGC of HKSAR Grant No. PolyU 5131/02E. [♯]S.Y. Kung was also a Distinguished Chair Professor of The Hong Kong Polytechnic University.

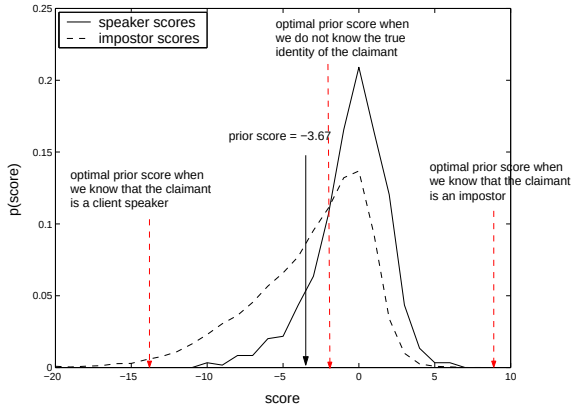

Figure 1: Distributions of pattern-by-pattern speaker scores and impostor scores corresponding to speaker "mdac0".

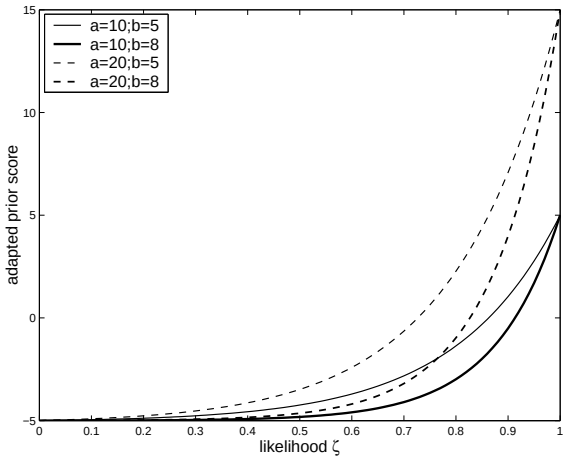

Figure 2: The influence of varying a and b of (4) on the adapted prior scores $\widehat{\mu}_p^{(k)}$, assuming that $\tilde{\mu}_p = -5$.

mean and variance of the client speaker and the background speakers, determined using enrollment data. Note that we have assumed that the K utterances contain the same number of feature vectors.

2.2. Adaptation of prior scores

In (3), we use the overall mean $\tilde{\mu}_p$ as the prior score. However, as the number of background speaker's utterances is much larger than the number of speaker's utterances during the training phase, the overall mean is very close to the mean score of the background speakers $\tilde{\mu}_b$, i.e. $\tilde{\mu}_p \approx \tilde{\mu}_b$. This causes most of the client-speaker scores larger than the prior score as $\tilde{\mu}_b$ is typically smaller than the client speaker scores; moreover, only part of the impostor scores will be smaller than the prior score. An example of this situation is illustrated in Fig. 1 in which the distributions of the pattern-by-pattern speaker scores and impostor scores corresponding to speaker "mdac0" in the HTIMIT corpus are shown. We can see from Fig. 1 that if the claimant is a client speaker, the fusion algorithm will favor large scores as most of the speaker scores in Fig. 1 are larger than the prior score. On the other hand, the algorithm will have almost the same preference on both small and large scores if the claimant is an impostor because the prior score lies on the middle of the impostor scores distribution.

The ultimate goal of (3) is to favor large scores when the claimant is a client speaker; on the other hand, if the claimant is an impostor, the fusion algorithm should favor small scores. In other words, the goal is to increase the separation between client speaker's scores and impostors' scores. As a result, when the claimant is a true speaker, the prior score should be smaller than all possible speaker scores so that (3) only favors larger scores; on the other hand, when the claimant is an impostor, the prior score should be larger than all possible impostor scores so that (3) only favors smaller scores. However, achieving these conditions are almost impossible in practice as we never know the true identity of the claimant. Therefore, the optimal prior score should be equal to the intersection point of speaker score distribution and impostor score distribution (see Fig. 1). At that point, the number of speaker scores smaller than the prior score plus the number of impostor scores larger than the prior score is kept to a minimum.

Here, we propose a method to adapt the prior score during verification in order to achieve the goal mentioned above. Let us denote the adapted prior score corresponding to utterance U_k as $\widehat{\mu}_p^{(k)}$. During recognition, the prior score is adapted according to

$$\widehat{\mu}_p^{(k)} = \tilde{\mu}_p + f(\zeta) \tag{4}$$

where $f(\zeta)$ is a positive monotonic increasing function of ζ, and ζ is a normalized likelihood computed by

$$\zeta = \frac{p(\tilde{s}; \Omega_{score})}{\max_{-\infty \le s \le \infty} p(s; \Omega_{score})} \tag{5}$$

where

$$p(\tilde{s}; \Omega_{score}) = \sum_{j=1}^{M} \pi_j p(\tilde{s}|j) = \sum_{j=1}^{M} \pi_j \mathcal{N}(\tilde{s}; \mu_j, \Sigma_j) \tag{6}$$

and

$$\tilde{s} = \frac{1}{T_k} \sum_{t=1}^{T_k} s(\mathbf{o}_t^{(k)}; \Lambda). \tag{7}$$

Here, we denote $\Omega_{score} = \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{M}$ as a single-dimension GMM that characterizes the mean pseudo-impostor scores $\tilde{s}$, and $p(\tilde{s}; \Omega_{score})$ is denoted as the output of the GMM given a specific $\tilde{s}$. Notice that (4) increases the prior score rather than decreasing it because we found that the optimal prior score is always greater than the unadapted prior score $\tilde{\mu}_p^{(k)}$. In this work, we use $f(\zeta) = b(e^{a\zeta} - 1)/(e^a - 1)$ as the monotonic increasing function, where a and b respectively control the rate of increase of $f(\zeta)$ and the maximum amount of adaptation. Fig. 2 shows the effect of varying a and b on the adapted prior scores, where we can notice that the larger the normalized likelihood, the greater the degree of adaptation. We used an exponential increasing function rather than linear increasing functions because a large normalized likelihood means that the claimant is likely to be an impostor; as a result, the prior score should be increased by a larger amount to de-emphasize the large scores of this claimant. With the large scores being de-emphasized, the claimant's mean fused score will become smaller, thus increasing the chance of rejecting this potential impostor.

To demonstrate the effect of the proposed prior score adaptation on fused score distributions, we arbitrarily select a client speaker (mdac0) from GSM-transcoded HTIMIT and plot the

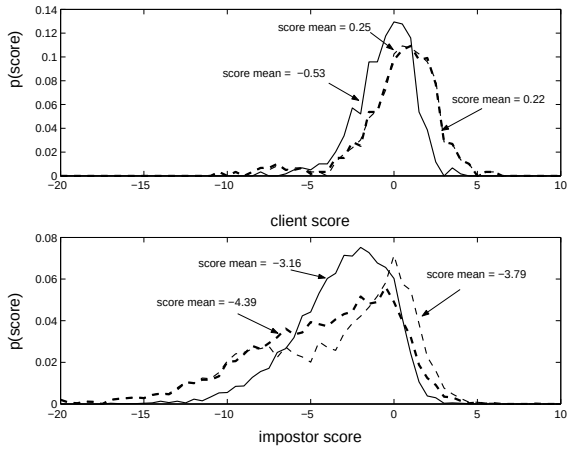

Figure 3: Distributions of pattern-by-pattern speaker scores (upper figure) and impostor scores (lower figure) based on equal weight fusion and data-dependent fusion. Solid: equal weight fusion. Thin Dotted: Data-dependent fusion without prior score adaptation. Thick Dotted: Data-dependent fusion with prior score adaptation.

distributions of the fused speaker scores and fused impostor scores in Fig. 3, using equal weight fusion ($\alpha_t^{(1)} = \alpha_t^{(2)} = 0.5$ $\forall t$), data-dependent fusion without prior score adaptation (3) and data-dependent fusion with prior score adaptation (4). Evidently, the upper part of Fig. 3 shows that the number of large client-speaker scores is larger in data-dependent fusion, and the lower part of Fig. 3 shows that there are more small impostor scores in data-dependent fusion than in equal weight fusion. The dispersion between the mean client score and the mean impostor score increases from 2.63 $(= -0.53 - (-3.16))$ to 4.04 $(= 0.25 - (-3.79))$ without prior score adaptation and to 4.61 $(= 0.22 - (-4.39))$ with prior score adaptation. As verification decisions are based on the mean scores, the wider the dispersion between the mean client scores and the mean impostor scores, the lower the error rate. This shows that data-dependent fusion with prior score adaptation outperforms the other two fusion approaches.

2.3. Stochastic feature transformation

In feature transformation [5], a telephone channel can be represented by a stochastic cepstral bias $\mathbf{b} = [b_1, \ldots, b_D]^T$, and the recovered vectors are given by

$$\hat{\mathbf{o}}_t = f_\nu(\mathbf{o}_t) = \mathbf{o}_t + \mathbf{b} \tag{8}$$

where $\mathbf{o}_t$'s are D-dimensional distorted vectors and f_ν denotes the transformation function. Intuitively, the bias $\mathbf{b}$ compensates the convolutive distortion caused by the channel. Given a clean GMM speech model $\Lambda = \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^M$, where $\Sigma_j = \text{diag}\{\sigma_{j1}^2, \ldots, \sigma_{jD}^2\}$, derived from the clean speech of several speakers (ten speakers in this work) and distorted speech $\mathbf{o}_t$, $t = 1, \ldots, T$, the maximum likelihood estimates of $\mathbf{b}$ can be obtained by the EM algorithm. Specifically, in each M-step, we compute the new estimate of $\mathbf{b}$ by

$$b_i' = \frac{\sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{o}_t))(\sigma_{ji})^{-2}(\mu_{ji} - \mathbf{o}_t)}{\sum_{t=1}^T \sum_{j=1}^M h_j(f_\nu(\mathbf{o}_t))(\sigma_{ji})^{-2}} \tag{9}$$

where $i = 1, \ldots, D$, $f_\nu(\mathbf{o}_t) = \mathbf{o}_t + \mathbf{b}$ and $h_j(f_\nu(\mathbf{o}_t))$ is the posterior probability of using the j-th mixture, which has been computed in the E-step (see [5] for details).

In this work, the feature transformation was combined with a handset selector [6] for robust speaker verification. Specifically, before verification takes place, we compute one set of transformation parameters for each type of handsets that claimants are likely to use. Then, during a verification session, we identify the most likely handset that is used by the claimant and select the best set of transformation parameters accordingly.

3. Experiments

We used a GSM speech coder to transcode the HTIMIT corpus [7] and applied the resulting transcoded speech in a speaker verification experiment similar to [5] and [4]. 12 MFCCs were extracted from the 28ms speech frames of uncoded and GSM-transcoded utterances at a frame rate of 71 Hz.

During enrollment, we used the SA and SX utterances from handset "senh" of the uncoded HTIMIT to create a 32-center GMM for each speaker. A 64-center universal background GMM [8] was also created based on the speech of 100 client speakers recorded from handset "senh". The background model will be shared among all client speakers in subsequent verification sessions. Finally, two sets of pseudo-impostor scores were collected. For the first set, we fed the SA and SX utterances of all the speakers in the speaker set to the background model and each of the speaker models to obtain the pseudo-impostor scores corresponding to the enrollment data. For the second set, we fed the utterances of all the speakers in the pseudo-impostor sets to the background model and each of the speaker models to obtain the pseudo-impostor scores corresponding to "unseen" impostor data. These pseudo-impostor scores were averaged for each utterance (see (6)). The resulting utterance-based scores were used to create a 2-center, 1-D GMM pseudo-impostor score model (Ω_{score} in Section 2.2).

For verification, we used the GSM-transcoded speech from handset "cb1". As a result, there were handset and coder mismatches between the speaker models and the verification utterances. We used stochastic feature transformation with handset identification [5][6] to compensate the mismatches. We assume that a claimant will be asked to utter two sentences during a verification session. Therefore, for each client speaker and each impostor, we applied the proposed fusion algorithm to fuse two independent streams of scores obtained using his/her SI sentences. Note that we need to make the two utterances to have an identical number of feature vectors (length) before fusion takes place. This is achieved by computing the average length of the two utterances and then appending the extra patterns in the longer utterance to the end of the shorter utterance.

In the experiments, we set the parameters a and b in (4) to 10 and 5 respectively. These values were found empirically to obtain reasonably good results.

4. Results and discussion

Fig. 4 depicts the speaker detection performance based on 100 speakers and 50 impostors for equal weight fusion and data-dependent fusion with and without prior score adaptation. Fig. 4 clearly shows that with feature transformation, data-dependent fusion is able to reduce the error rates significantly. In particular, with feature transformation, the equal error rate (EER) achieved by data-dependent fusion with prior score adaptation is 4.01%. When compared to equal weight fusion (which

	PI-20 GSM-HTIMIT	PI-100 GSM-HTIMIT	SS-100 HTIMIT
DF w/ PS adaptation (a=10,b=5)	3.01%	2.92%	4.01%
DF w/o PS adaptation [1]	3.61%	3.52%	4.14%
Equal weight fusion [3]	5.11%	5.11%	5.11%
No fusion (Single utterance)	6.31%	6.31%	6.31%

Table 1: The equal error rates obtained from different fusion approaches. Each figure is based on the average of 100 speakers, each impersonated by 50 impostors. PI-20 GSM-HTIMIT and PI-100 GSM-HTIMIT represent the cases where Ω_{score} in (6) is obtained from a pseudo-impostor set containing 20 and 100 speakers respectively, with GSM-transcoded speech from handset "cb1" being used for training. SS-100 HTIMIT represents the case where Ω_{score} is obtained from a speaker set containing 100 speakers, with HTIMIT speech from handset "senh" being used for training. DF stands for the data-dependent fusion in [1].

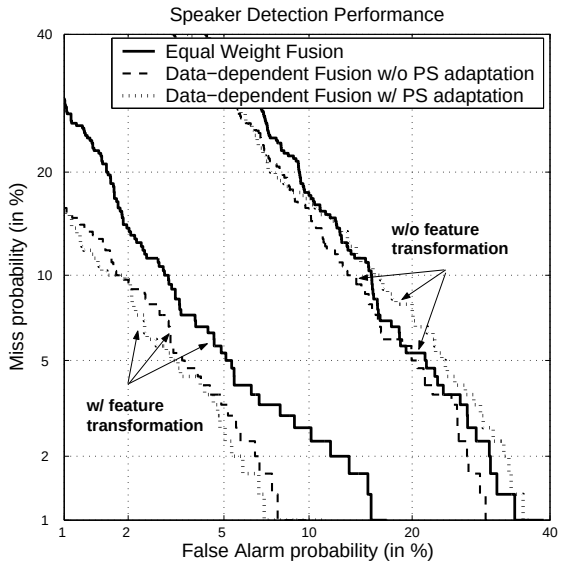

Figure 4: Speaker detection performance for equal weight fusion (score averaging) and data-dependent fusion. PS stands for prior score.

achieves an EER of 5.11%), a relative error reduction of 22% was obtained. However, without feature transformation, the performance of data-dependent fusion is not significantly better than that of the equal weight fusion. This is caused by the mismatch between the prior scores $\tilde{\mu}_p$'s in (3) and the scores of the distorted features. This result demonstrates that it is very important to use feature transformation in data-dependent fusion. In order to create a more stable and reliable pseudo-impostor score model, we use another set of speakers, namely the pseudo-impostor set, to train the pseudo-impostor score model. Also, we used the GSM-transcoded speech instead of HTIMIT speech to train the model in an attempt to create a verification environment as close to the real one as possible. The improvement after adapting the prior score is 17% (from 3.52% to 2.92%), as demonstrated in the second column of Table 1. We can see from Table 1 that even for pseudo-impostor set containing as little as 20 speakers, a lower equal error rate (3.01%) can still be obtained by using the GSM-transcoded speech to train the pseudo-impostor score model.

5. Conclusions

We have presented a decision fusion algorithm that makes use of prior score statistics and the distribution of the recognition data. The statistics of pseudo-impostor scores are used to adapt the prior scores in the fusion algorithm. The fusion algorithm was combined with feature transformation for speaker verification using GSM-transcoded speech. Results based on 100 speakers and 50 impostors show that combining stochastic transformation with the proposed fusion algorithm can reduce error rate significantly. Also, our proposed adaptive fusion algorithm outperforms fixed-weight fusion by 43%.

6. References

[1] 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 Int. Conf. on Acoustic, Speech and Signal Processing*, 2003.

[2] C. Sanderson and K. K. Paliwal, "Joint cohort normalization in a multi-feature speaker verification system," in *The 10th IEEE International Conference on Fuzzy Systems, 2001*, 2001, vol. 1, pp. 232 –235.

[3] 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.

[4] Eric W.M. Yu, M. W. Mak, and S.Y. Kung, "Speaker verification from coded telephone speech using stochastic feature transformation and handset identification," in *The 3rd IEEE Pacific-Rim Conference on Multimedia 2002*, 2002, pp. 598–606.

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

[6] C.L. Tsang, M. W. Mak, and S.Y. Kung, "Divergence-based out-of-class rejection for telephone handset identification," in *Proc. ICSLP'02*, 2002, pp. 2329–2332.

[7] D. A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in *ICASSP'97*, 1997, vol. 2, pp. 1535–1538.

[8] 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.