ROBUST SPEAKER VERIFICATION FROM GSM-TRANSCODED SPEECH BASED ON DECISION FUSION AND FEATURE TRANSFORMATION

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

In speaker verification, a claimant may produce two or more utterances. Typically, the scores of the speech patterns extracted from these utterances are averaged and the resulting mean score is compared with a decision threshold. Rather than simply computing the mean score, we propose to compute the optimal weights for fusing the scores based on the score distribution of the independent utterances and our prior knowledge about the score statistics. More specifically, we use enrollment data to compute the mean scores of client speakers and impostors and consider them to be the prior scores. During verification, we set the fusion weights for individual speech patterns to be a function of the dispersion between the scores of these speech patterns and the prior scores. Experimental results based on the GSM-transcoded speech of 150 speakers from the HTIMIT corpus demonstrate that the proposed fusion algorithm can increase the dispersion between the mean speaker scores and the mean impostor scores. Compared with a baseline approach where equal weights are assigned to all scores, the proposed approach provides a relative error reduction of 19%.

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*. An ideal speaker verification system should not reject registered speakers (*false rejection*) or accept impostors as registered speakers (*false acceptance*).

Recently, there has been increasing interest in recognizing speakers using resynthesized coded speech. For example, speaker verification based on GSM, G.729, and G.723.1 resynthesized speech was studied in [1]. It was shown that the verification performance generally degrades with coders' bit rate. To improve the verification performance of G.729 coded speech, techniques that require knowledge of the coder parameters and coder internal structure were proposed in [1] and [2]. However, the performance of these improved techniques is still poorer than the one that uses features extracted directly from resynthesized speech.

In this work, we investigate the fusion of scores from multiple utterances to improve the performance of speaker verification from GSM-transcoded speech. Instead of averaging the scores of multiple utterances from a claimant, as in [3], we compute the optimal fusion weights based on the score distribution of the utterances and on the prior score statistics determined from enrollment data. As the variation of handset characteristics and the encoding/decoding process will introduce substantial distortion to the speech signals [4], we also apply stochastic feature transformation [5] to the feature vectors extracted from the GSM-transcoded speech before presenting them to the speaker models.

2. DECISION FUSION

Decision fusion can be divided into two levels: abstract level and score level. In the abstract level, the binary decisions made by multiple classifiers are combined, whereas in the score level, the scores of modality-specific classifiers are combined through a set of fusion weights [6]. These weights can be non-adaptive and adaptive. Non-adaptive weights are learned from training data and kept fix during recognition [7]. Adaptive weights, on the other hand, are estimated from the observed data during recognition, e.g. according to the signal-to-noise ratio [8] and degree of voicing [9]. This paper focuses on score-level decision fusion.

Although decision fusion is mainly applied to combine the outputs of modality-dependent classifiers, it can also be applied to fuse 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 claimant. The approach is commonly referred to as multi-

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.

sample fusion [3]. In [3], the scores from multiple samples were averaged, which means that the fusion weights are equal for all scores. Although encouraging results have been obtained, further improvement may be obtained by determining the optimal fusion weights based on the score statistics. In this paper, we refer to this type of fusion as data-dependent decision fusion.

3. 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

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

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

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

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 GMM Λ_ω, $\omega \in \{\omega_c, \omega_b\}$, given observation $\mathbf{o}_t^{(k)}$.

In this work, the expert-in-class architecture [10] was used to combine the normalized score functions probabilistically. Specifically, frame-level fused scores are computed according to

$$s(\mathbf{o}_t^{(1)}, \ldots, \mathbf{o}_t^{(K)}; \Lambda) = \sum_{k=1}^{K} \alpha_t^{(k)} s(\mathbf{o}_t^{(k)}; \Lambda) \quad (3)$$

where $\alpha_t^{(k)} \in [0, 1]$ represents the confidence (reliability) of the observation $\mathbf{o}_t^{(k)}$ and $\sum_k \alpha_t^{(k)} = 1$. Note that for notational convenience, we have assumed that the K utterances contain the same number of feature vectors. If it is not the case, we may repeat the vectors of the shorter utterances to make the number of feature vectors equal. Alternatively, we can fuse the first T patterns of the utterances where $T = \min_j T_j$. Note also that in (3), a larger (resp. smaller) fusion weight means a greater (resp. lesser) influence on the final decision. The fusion weights can be estimated using training data; alternatively, they can be determined purely from the observation data during recognition. Rather than using either training data or recognition data exclusively, we propose a new approach in which the fusion weights depend on both training data (prior information) and recognition data.

During enrollment, the mean score of each client speaker ($\tilde{\mu}_c$) and of the background speakers ($\tilde{\mu}_b$) are determined. Then, the overall mean score

$$\tilde{\mu}_p = \frac{K_c \tilde{\mu}_c + K_b \tilde{\mu}_b}{K_c + K_b}, \quad (4)$$

where K_c and K_b are respectively the numbers of speaker's utterances and background speakers' utterances, will be used as a prior score for that client. A prior variance

$$\tilde{\sigma}_p^2 = \frac{1}{K_c + K_b} \sum_{k=1}^{K_c+K_b} \left[\tilde{s}(O^{(k)}; \Lambda) - \tilde{\mu}_p\right]^2 \quad (5)$$

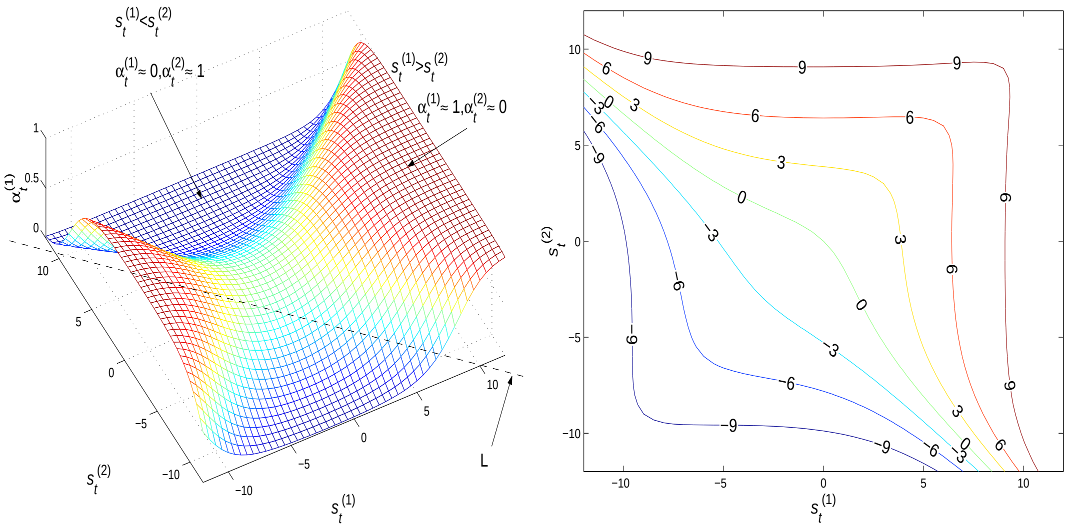
will also be computed, where $\tilde{s}(O^{(k)}; \Lambda)$ denotes the mean score of the k-th utterance. Then, during verification, the claimant is asked to utter K utterances, and the fusion weights are computed according to

$$\alpha_t^{(k)} = \frac{\exp\{(s_t^{(k)} - \tilde{\mu}_p)^2 / 2\tilde{\sigma}_p^2\}}{\sum_{l=1}^{K} \exp\{(s_t^{(l)} - \tilde{\mu}_p)^2 / 2\tilde{\sigma}_p^2\}} \quad k = 1, \ldots, K \quad (6)$$

where for notation convenient, we have defined $s_t^{(k)} \equiv s(\mathbf{o}_t^{(k)}; \Lambda)$.

Fig. 1(a) illustrates the fusion weights $\alpha_t^{(1)}$ as a function of $s_t^{(1)}$ and $s_t^{(2)}$ where $s_t^{(k)} \in [-12, 12]$, $K = 2$, $\tilde{\mu}_p = -2$ and $\tilde{\sigma}_p = 3.5$. A closer look at Fig. 1(a) reveals that scores falling on the upper right hand region of the dashed line L will be increased by the fusion function (3). This is because in that region, for $s_t^{(1)} > s_t^{(2)}$, $\alpha_t^{(1)} \approx 1$ and $\alpha_t^{(2)} \approx 0$; moreover, for $s_t^{(1)} < s_t^{(2)}$, $\alpha_t^{(1)} \approx 0$ and $\alpha_t^{(2)} \approx 1$. Both of these conditions favor the larger score. On the other hand, the fusion algorithm will put more emphasis on the small scores if they fall on the lower left hand region of the dashed line. The effect of the fusion weights on the scores is depicted in Fig. 1(b). Evidently, the fusion weights will favor large scores if they fall on the upper right hand region, whereas the fused scores will be close to the small scores if they fall on the lower-left hand region.

The rationale behind this fusion approach is the observation that most of the client-speaker scores are larger than the prior score while most of the impostor scores are smaller than the prior score. As a result, if the claimant is a client speaker, the fusion algorithm will favor large scores; on the other hand, the algorithm will favor small scores if the claimant is an impostor. This has the effect of reducing the overlapping area of the score distribution of the client speakers and the impostors, thus reducing the error rate. To demonstrate this phenomenon, we arbitrarily select a client speaker (mdac0) from HTIMIT and plot the distributions of the fused speaker scores and fused impostor scores in Fig. 2, using equal weight fusion ($\alpha_t^{(1)} = \alpha_t^{(2)} = 0.5 \; \forall t$) and data-dependent fusion (6). Evidently, the upper part of Fig. 2 shows that the number of large client-speaker scores is larger in data-dependent fusion, and the lower part



(a) (b)

Fig. 1. (a) Fusion weights $\alpha_t^{(1)}$ as a function of scores $s_t^{(1)}$ and $s_t^{(2)}$. (b) Contour plot of fused scores based on the fusion formula (3) and the fusion weights in (a).

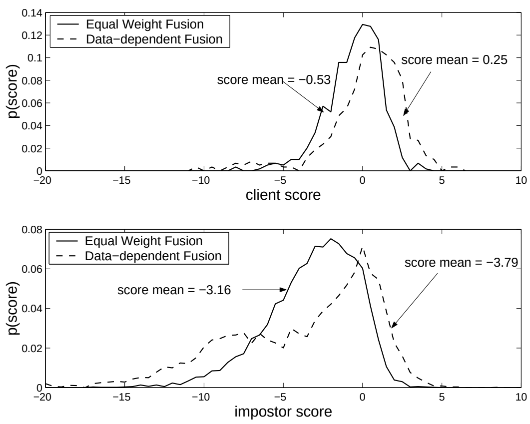


Fig. 2. Distribution of pattern-by-pattern speaker scores (upper figure) and impostor scores (lower figure) based on equal weight fusion (score averaging) and data-dependent fusion. The means of speaker scores and impostor scores obtained by both fusion approaches are also shown.

of Fig. 2 shows that there are more small impostor scores in data-dependent fusion than in equal weight fusion. As a result, the dispersion between the mean client score and the mean impostor score was increased from 2.63 $(= -0.53 - (-3.16))$ to 4.04 $(= 0.25 - (-3.79))$. As verification decision is based on the mean scores, the wider the dispersion between the mean client scores and the mean impostor scores, the lower the error rate.

4. 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

$$\hat{\mathbf{o}}_t = f_\nu(\mathbf{o}_t) = \mathbf{o}_t + \mathbf{b} \qquad (7)$$

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 $\Sigma_j = \mathrm{diag}\{\sigma_{j1}^2, \ldots, \sigma_{jD}^2\}$, 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

$$b_i' = \frac{\sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{o}_t))(\sigma_{ji})^{-2}(\mu_{ji} - \mathbf{o}_t)}{\sum_{t=1}^{T}\sum_{j=1}^{M} h_j(f_\nu(\mathbf{o}_t))(\sigma_{ji})^{-2}} \qquad (8)$$

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 [11] 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.

5. EXPERIMENTS AND RESULTS

We applied the proposed fusion algorithm to fuse two independent streams of scores. 12th-order MFCCs were extracted from independent utterances at a frame rate of 14 ms. We used a GSM speech coder to transcode the HTIMIT corpus [12] and applied the resulting transcoded speech in a speaker verification experiment similar to [5] and [4]. For each speaker, we used the SA and SX utterances from handset "senh" of the uncoded HTIMIT to create a 32-center speaker model. A 64-center universal background model was also created based on the speech of 100 client speakers. The background model will be shared among all client speakers in subsequent verification sessions. 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 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. Since different utterances contain different numbers of feature vectors, 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. To compare with the score averaging approach proposed in [3], we also fused the speech segments using equal fusion weights, i.e., $\alpha_t^{(1)} = \alpha_t^{(2)} = 0.5$.

Fig. 3 depicts the speaker detection performance of 100 speakers and 50 impostors for the equal weight fusion (score averaging) approach and the proposed fusion approach. Fig. 3 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 is 4.14%. When compared to equal weight fusion (which achieves an EER of 5.11%), a relative error reduction of 19% 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 (6) and the scores of the distorted features. This result demonstrates that it is very important to use feature transformation in data-dependent fusion.

6. CONCLUSIONS

We have presented a decision fusion algorithm that makes use of prior score statistics and the distribution of the recognition data. The fusion algorithm was also combined with feature transformation for speaker verification using GSM-transcoded speech. Results based on 150 speakers show that combining stochastic transformation with the proposed fusion algorithm can reduce error rate significantly. We are currently extending the algorithm to multi-modality fusion with multi-feature transformation.

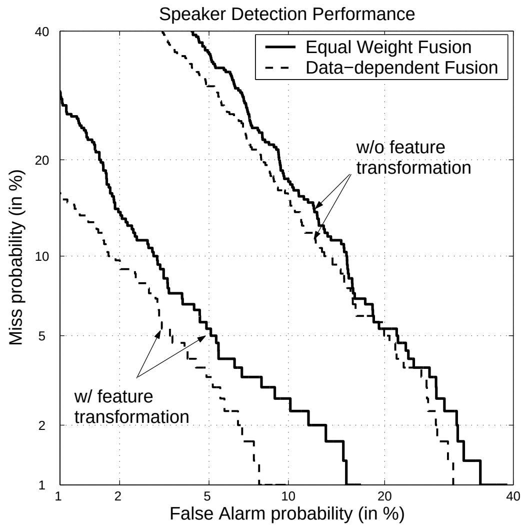


Fig. 3. Speaker detection performance of equal weight fusion (score averaging) and data-dependent fusion.

7. REFERENCES

[1] T. F. Quatieri, E. Singer, R. B. Dunn, D. A. Reynolds, and J. P. Campbell, "Speaker and language recognition using speech codec parameters," in *Proc. Eurospeech'99*, 1999, vol. 2, pp. 787–790.

[2] T. F. Quatieri, R. B. Dunn, D. A. Reynolds, J. P. Campbell, and E. Singer, "Speaker recognition using G.729 codec parameters," in *Proc. ICASSP'2000*, 2000, pp. 89–92.

[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 *Proc. PCM'02*, 2002.

[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.

[6] 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.

[7] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST'99 1-speaker submisions," *Digital Signal Processing*, vol. 10, pp. 237–248, 2000.

[8] U. Meier, W. Hurst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speech reading," in *Proc. ICASSP'96*, 1996, pp. 833–836.

[9] C. Neti et al., "Audio-visual speech recognition," in *Final Workshop 2000 Report*, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, 2000.

[10] S.Y. Kung, J. Taur, and S.H. Lin, "Synergistic modeling and applications of hierachical fuzzy neural networks," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1550–1574, 1999.

[11] 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.

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