MULTI-SAMPLE DATA-DEPENDENT FUSION OF SORTED SCORE SEQUENCES FOR BIOMETRIC VERIFICATION

Ming-Cheung Cheung, 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 many biometric systems, the scores of multiple samples (e.g. utterances) are averaged and the average score is compared against a decision threshold for decision making. The average score, however, may not be optimal because the distribution of the scores is ignored. To address this limitation, we have recently proposed a fusion model that incorporates the score distribution by making the fusion weights dependent on the dispersion between the frame-based scores and the prior score statistics obtained from training data. As the fusion weights are data-dependent, the positions of scores in the score sequences become detrimental to the final fused scores. In this paper, we propose to enhance the fusion model by sorting the score sequences before fusion takes place. The fusion model was evaluated on a speaker verification task where each claimant utters two utterances in a verification session. Results demonstrate that fusion of sorted scores has the effect of maximizing the dispersion between the client scores and the impostor scores, making the verification process more reliable. Compared with our previous work where no sorting is applied, the new approach reduces the equal error rate by 11%.

1. INTRODUCTION

Although decision fusion is mainly applied to combine the outputs of modality-dependent classifiers (see [1] for a review), it can also be applied to fuse the decisions or scores from 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 application, multi-sample fusion will not impose any burden to users because a single, long sample (e.g. an utterance or video shot) can always be divided into a number of short samples. Typically, the scores from multiple samples are averaged (e.g., [2]). However, this approach is equivalent to the single-sample case because the fusion weights are equal. The benefit of fusing multi-samples arises when the weights for individual scores from multiple samples are different.

To overcome the limitation of the score averaging approach, we have recently proposed a fusion model [3] 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. While it was demonstrated that incorporating the prior score information into the computation of fusion weights helps reduce the error rate, the approach simply fuses two independent streams of scores in a score-by-score basis without considering the best combination of scores for fusion. In this paper, we propose to sort the two streams of scores (one in ascending order and another one in descending order) before fusion takes place so that large scores will always be fused with small scores. The proposed score sorting approach is applied to a speaker verification task involving 150 speakers using 10 different handsets. It was found that multi-sample fusion with the sorting of score sequences can reduce the equal error rate significantly.

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 in Section 2. This is followed by an explanation of the proposed score sorting approach in Section 3, where the benefit of fusing the sorted scores is demonstrated through a Gaussian example. The proposed method is further evaluated in Section 4 via a speaker verification experiment using GSM-transcoded speech. 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

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

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 [4]

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

where $\Lambda = \{\Lambda_{\omega_c}, \Lambda_{\omega_b}\}$ contains the Gaussian mixture models (GMMs) that characterize the client speaker (ω_c) and the back-

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

[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 utterance and determines the score mean of the concatenated utterance. The idea is identical to moving the feature vectors among the utterances here.

ground 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

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

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

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

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:** $\sum_{k=1}^{K} \alpha_t^{(k)} = 1 \quad \forall\, t = 1, \ldots, T$.

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 incorporate 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 score mean $\tilde{\mu}_p$ and score variance $\tilde{\sigma}_p^2$ of the client's and background speakers' speech are computed. We refer to these parameters as prior score and prior variance. Then, during verification, the claimant is asked to utter K utterances and the fusion weights are computed as

$$\alpha_t^{(k)} = \frac{\exp\{(s_t^{(k)} - \tilde{\mu}_p)^2 / 2\tilde{\sigma}_p^2\}}{\sum_{k=1}^{K} \exp\{(s_t^{(k)} - \tilde{\mu}_p)^2 / 2\tilde{\sigma}_p^2\}} \quad (5)$$

where $t = 1, \ldots, T$ and $k = 1, \ldots, K$.

Fig. 1 depicts the dataflow of the verification process and the architecture of the fusion models. While this architecture bears a slight resemblance to that of the hierarchical mixture-of-expert (HME) [5] and the class-in-experts [6] in that all of these models posses a gating network with outputs being dependent on input data, there are also important differences. First, the gating network in our model works on the score space, while that of HME and class-in-experts works on the feature space. Second, there is a major difference in the algorithm for training the gating network. While the HME requires the generalized EM algorithm for training their gating networks in order to capture the importance of individual experts, our model makes use of the highly representative information extracted from the training sessions for the same purpose. Both of these differences make our model more practical because (1) the input dimension of the gating networks can be considerably reduced and (2) iterative optimization of the gating networks is no longer required.

3. FUSION OF SORTED SCORES

As the fusion model described in Section 2 depends on the pattern-based scores of individual utterances, the positions of scores in the score sequences may also affect the final fused scores. We propose to sort the scores in the score sequences before fusion such that small scores will always be fused with large scores. This is achieved by sorting half of the score sequences in ascending order and the other half in descending order. Note that the fusion of sorted score applies only to even numbers of utterances.

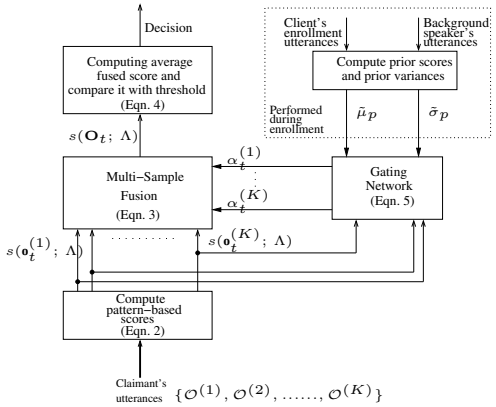

Fig. 1. Architecture of the multi-sample fusion model.

3.1. Theoretical Analysis

Here, we provide a theoretical analysis to explain why the fusion of sorted scores is better than the fusion of unsorted scores. Let us consider a hypothetical situation in which the distributions of the speaker scores are Gaussian. Let us also assume that a claimant will utter two utterances in a verification session. We denote the scores of the two utterances as $s_t^{(1)}$ and $s_t^{(2)}$ for $t = 1, \ldots, T$. Fig. 2 depicts the probability of occurrences of $(s_t^{(1)}, s_t^{(2)})$ when the means of $s_t^{(1)}$ and $s_t^{(2)}$ are 0.8 and 1.2 respectively and their variances are identical. The straight line L in the lower part of Fig. 2 separates the regions for which $s_t^{(2)} \gtrless -s_t^{(1)} + 2\tilde{\mu}_p$ where $\tilde{\mu}_p = 0$ is the prior score. According to (5), the fusion function (3) will emphasize the larger scores of the score pairs $(s_t^{(1)}, s_t^{(2)})$ if the pairs fall in the region above Line L, i.e. $s_t^{(2)} - \tilde{\mu}_p > \tilde{\mu}_p - s_t^{(1)}$. This is because in that region, there are two possible combinations: (a) both scores are larger than the prior score and (b) only one of the scores is larger than the prior score and the other one is smaller. In the former case, it is obvious that (5) will assign a larger weight to the larger score; in the latter case, as the larger score is always further away from $\tilde{\mu}_p$ than the smaller score does, a larger weight will still be assigned to the larger score. Both of these situations will make the fusion function (3) to emphasize the larger score. On the other hand, the fusion function (3) will emphasize the smaller scores when the score pairs fall in the region below the dashed line L. This is because the smaller score is now further away from the prior score than the larger score does. For the score distribution shown in Fig. 2, the mean of the data-dependent fused scores computed using (3)-(5) will be larger than that computed using equal-weight fusion. On the other hand, if the majority of $(s_t^{(1)}, s_t^{(2)})$ pairs fall in the lower region of Line L, the opposite situation will occur.

In the above analysis, the increase or decrease of the mean fused scores is only probabilistic because there is no guarantee that the scores of the two utterances $(s_t^{(1)}, s_t^{(2)})$ $\forall t$ will fall in either the upper-right region or the lower-left region of the score space

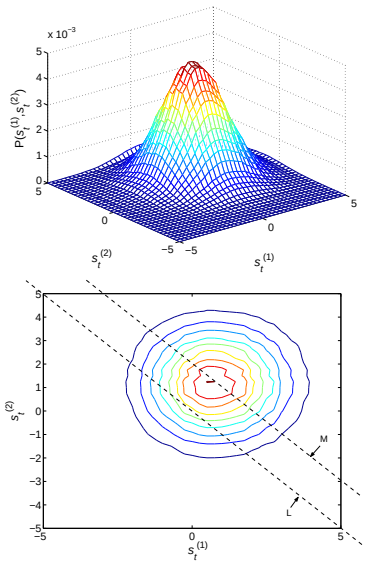

Fig. 2. Probability of $(s_t^{(1)}, s_t^{(2)})$ pairs. The means of the two Gaussian distributions are 0.8 and 1.2. We assume that the prior score $\tilde{\mu}_p$ is 0.

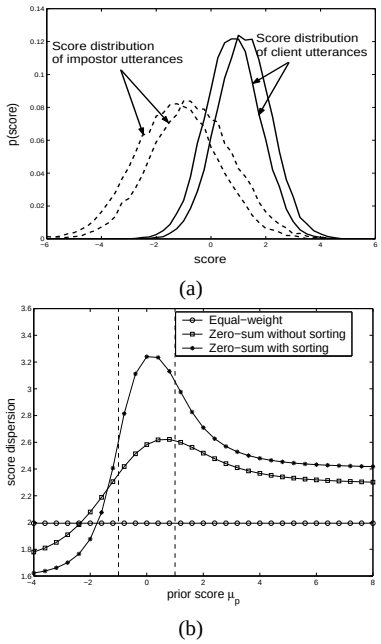

Fig. 3. (a)Distributions of client scores and impostor scores as a result of four utterances: two from a client speaker and another two from an impostor. The mean of the client scores is 1.0 and the mean of impostor scores is -1.0. (b)Dispersion between the means of the fused client scores and the fused impostor scores based on equal-weight fusion and zero-sum fusion.

together. There are many cases in which some of the $(s_t^{(1)}, s_t^{(2)})$ pairs fall in the region above Line L in Fig. 2 and others in the region below it, even though the two utterances are obtained from the same speaker. This situation is undesirable because it introduces uncertainty to the increase or decrease of the mean fused scores. This uncertainty, however, can be removed by sorting the two score sequences before fusion takes place, because the scores to be fused will always lie on a straight line. For example, if we denote the mean scores of both utterances from the client speaker as μ and assume that $s_t^{(1)}$ is sorted in ascending order and $s_t^{(2)}$ in descending order, we can obtain the relationship: $\mu - s_t^{(1)} \approx s_t^{(2)} - \mu$ $\forall t$. This is the straight line M $(s_t^{(2)} = -s_t^{(1)} + 2)$ shown in Fig. 2 when $\mu = 1$. Evidently, Line M lies in the region where large scores are emphasized. As a result, an increase in the mean fused score can be guaranteed.

3.2. Gaussian Example

Here, we provide a Gaussian example to demonstrate the merit of data-dependent fusion and the fusion of sorted scores. Fig. 3 illustrates an example where the distributions of the client speaker scores and the impostor scores are assumed to be Gaussian. It is also assumed that both the client and the impostor utter two utterances. The client speaker's mean scores for the first and second utterances are equal to 1.2 and 0.8 respectively. Likewise, the impostor's mean scores for the two utterances are equal to -1.3 and -0.7. Fig. 3(a) depicts the score distributions of the four utterances before fusion, and Fig. 3(b) plots the dispersion between the mean of the fused client scores $s(\mathcal{U}; \Lambda | \mathcal{U} \in \text{client})$ and the mean of the fused impostor scores $s(\mathcal{U}; \Lambda | \mathcal{U} \in \text{impostor})$ against the prior score $\tilde{\mu}_p$ using different fusion approaches. Obviously, equal weight fusion will produce a mean speaker score of 1.0 and a mean impostor score of -1.0, resulting in a score dispersion of 2.0. These two mean scores (-1.0 and 1.0) are indicated by the two vertical lines in Fig. 3(b). We can see from Fig. 3(b) that when the prior score $\tilde{\mu}_p$ is set to a value between these two means (i.e. between the vertical lines), the scores dispersion can be larger than 2.0.

As the verification decision depends on the difference between the mean fused score $s(\mathcal{U}; \Lambda)$ and the decision threshold, a large dispersion between the mean of the fused client scores and the mean of the fused impostor scores suggests that a more reliable decision can be made. We can see from Fig. 3(b) that there is a range of prior score $\tilde{\mu}_p$ for which the score dispersion obtained by zero-sum fusion can be larger than that of equal-weight fusion. There is also an optimum prior score at which the score dispersion is maximum. This explains why our previous proposal [7] on the adaption of prior scores based on the likelihood that the claimant is an impostor can reduce the error rate by as much as 17% when compared to the case without prior score adaptation.

To conclude, our fusion algorithm will either increase or decrease the fused score mean depending on the value of the prior score and the score mean before fusion. From Fig. 3(b), we can observe that when the prior scores are set between the mean of client scores and the mean of impostor scores (i.e. between the two vertical lines), theoretically the mean of fused client scores increases and the mean of fused impostor scores decreases. This has the effect of increasing the difference between the mean of fused client scores and the mean of fused impostor scores. As the mean of the fused scores is used to make the final decision, increasing the score dispersion can improve the reliability of the decision. Fig. 3(b) also shows that for a wide range of prior score $\tilde{\mu}_p$, the fusion of sorted scores produces a larger score dispersion as compared to the fusion of unsorted scores. This will lead to a reduction in error rate, as will be demonstrated in Section 4.

Fusion Method	Equal Error Rate (%)										
	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	senh	**average**
Equal weight fusion	5.11	4.33	19.15	12.89	4.42	8.31	9.96	6.29	7.57	2.99	**8.10**
ZS w/o sorting	4.01	3.27	15.92	10.55	3.04	6.51	8.67	4.75	7.51	2.32	**6.67**
ZS w/ sorting	3.60	2.86	15.30	9.91	3.49	4.65	6.81	4.02	6.59	1.99	**5.92**

Table 1. The equal error rates achieved by different fusion approaches, using utterances from 10 different handsets for verification. Each figure is based on the average of 100 speakers, each impersonated by 50 impostors. ZS stands for zero-sum fusion.

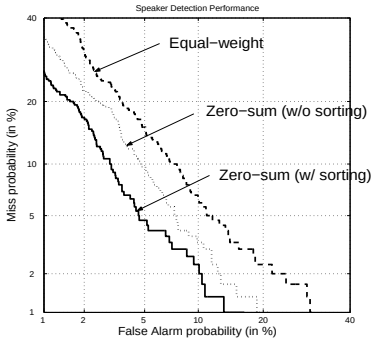

Fig. 4. DET curves for equal-weight fusion and zero-sum fusion with and without score sorting. The curves were obtained by using the utterances of handset "el2" as verification speech.

4. EXPERIMENTS AND RESULTS

We used a GSM speech coder to transcode the HTIMIT corpus [8] and applied the resulting transcoded speech in a speaker verification experiment similar to [9] and [10]. Sequences of 12th order MFCCs were extracted from 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 [4] 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.

For verification, we used the GSM-transcoded speech from all ten handsets in HTIMIT. 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 [9] 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. As the fusion algorithm requires the two utterances to have an identical number of feature vectors (length), we computed the average length of the two utterances and then appended the extra patterns in the longer utterance to the end of the shorter utterance. After that, we sorted the score sequences in opposite order and fused the sorted scores according to (3) and (5).

Fig. 4 depicts the detection error tradeoff curves based on 100 client speakers and 50 impostors using utterances from handset "el2" for verification. Fig. 4 clearly shows that data-dependent fusion is able to reduce the error rates significantly, and sorting the scores before fusion can reduce the error rate further.

Table 1 shows the speaker detection performance of 100 speakers and 50 impostors for the equal-weight fusion approach and the proposed fusion approach with and without sorting the score sequences. Table 1 clearly shows that our proposed fusion approach outperforms the equal-weight fusion. In particular, after the score sequences have been sorted, the equal error rates are further reduced from 6.67% to 5.92%, which represents an 11% error rate reduction.

5. CONCLUSIONS

We have presented a novel fusion model that makes use of prior score statistics and the distribution of the recognition data. By using a Gaussian example, we have shown that a simple but useful score sorting method can significantly increase the dispersion between speaker scores and impostor scores. Results of speaker verification using GSM-transcoded speech and feature transformation also agree with the Gaussian example, and an 11% error reduction as compared to the fusion of unsorted score sequences has been achieved. Compared to the equal-weight fusion approach, data-dependent fusion with score sorting reduce the error rate by 26%.

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 ICASSP03*, 2003, pp. II745–II748.

[4] 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.

[5] M. I. Jordan and R. A. Jacobs, "Hierarchical mixture of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[6] 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.

[7] M. C. Cheung, M. W. Mak, and S. Y. Kung, "Adaptive decision fusion for multi-sample speaker verification over GSM networks," in *Eurospeech'03*, 2003.

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

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

[10] 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.