

# INTRAMODAL AND INTERMODAL FUSION FOR AUDIO-VISUAL BIOMETRIC AUTHENTICATION

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

This paper proposes a multiple-source multiple-sample fusion approach to identity verification. Fusion is performed at two levels: intramodal and intermodal. In intramodal fusion, the scores of multiple samples (e.g. utterances or video shots) obtained from the same modality are linearly combined, where the combination weights are dependent on the difference between the score values and a user-dependent reference score obtained during enrollment. This is followed by intermodal fusion in which the means of intramodal fused scores obtained from different modalities are fused. The final fused score is then used for decision making. This two-level fusion approach was applied to audio-visual biometric authentication, and experimental results based on the XM2VTSDB corpus show that the proposed fusion approach can achieve an error rate reduction of up to 83%.

## 1. INTRODUCTION

Various biometric researches have suggested that no single modality can provide an adequate solution for high security applications. They all point to a common consensus that it is vital to utilize multiple modalities (e.g. visual, infrared, acoustic, chemical sensors, etc.).

In order to cope with the limitations of individual biometrics, researchers have proposed using multiple biometric traits concurrently for verification. Such systems are commonly known as multi-modal verification systems [1]. By using multiple biometric traits, systems gain more immunity to intruder attack. For example, it will be more difficult for an impostor to impersonate another person using both audio and visual information simultaneously. Multi-cue biometrics also helps improve system reliability. For instance, while background noise has a detrimental effect on the performance of voice biometrics, it does not have any influence on face biometrics. On the other hand, while the performance of face recognition systems depends heavily on lighting conditions, lighting does not have any effect on the voice quality. Therefore, audio-visual (AV) biometrics has attracted a great deal of attention in recent years.

Another approach to improving the effectiveness of biometric systems is to combine the scores of multiple input samples based on decision fusion techniques [2, 3, 4]. Although decision fusion is mainly applied to combine the outputs of modality-dependent classifiers, it can also be applied to fuse decisions or scores from a single modality. The idea is to consider multiple samples extracted from a single modality as independent but coming from the same

source. The approach is commonly referred to as multi-sample fusion [5].

This paper extends our recently proposed multi-sample fusion technique [2, 3, 4] to audio-visual biometric authentication systems. The remainder of this paper is organized as follows. Section 2 details the approach to computing the optimal weights for individual scores, based on the score distribution of independent samples and the prior knowledge of the score statistics. Evaluations of this multi-sample fusion technique on speaker verification, face verification and audio-visual (voice plus face) biometric authentication are presented in Sections 3 and 4. Concluding remarks are provided in the Section 5.

## 2. INTRAMODAL MULTI-SAMPLE DECISION FUSION

While decision fusion techniques are originally designed for fusing the scores of multiple modalities, they can easily be adapted to fuse the 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.

We assume that in each verification session,  $T^{(m)}$  normalized scores [6] are obtained from modality  $m$

$$\mathcal{S}^{(m)} = \{s_t^{(m)} \in \mathfrak{R}; t = 1, \dots, T^{(m)}\}, \quad (1)$$

where  $t$  is the frame index. In the *equal-weight* fusion approach [5], the mean score

$$\bar{s}^{(m)} = \frac{1}{T^{(m)}} \sum_{t=1}^{T^{(m)}} s_t^{(m)} \quad (2)$$

is used for decision making.

Instead of assigning an equal weight to all scores, Mak et al. [2] proposed a *zero-sum* intramodal fusion approach in which different weights are assigned to different scores. The approach splits a score sequence into  $K$  sub-sequence:

$$\mathcal{S}^{(m,k)} = \{s_t^{(m,k)} \in \mathfrak{R}; t = 1, \dots, T^{(m)}/K\} \quad k = 1, \dots, K. \quad (3)$$

The frame-level fused scores are then computed as

$$\hat{s}_t^{(m)} = \sum_{k=1}^K \alpha_t^{(m,k)} s_t^{(m,k)}, \quad (4)$$

where  $t = 1, \dots, T^{(m)}/K$ , and  $\alpha_t^{(m,k)} \in [0, 1]$  represents the confidence (reliability) of the score  $s_t^{(m,k)}$ . The fusion weights

$\alpha_t^{(m,k)}$  are made dependent on both the training data (prior information) and recognition data (scores):

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

where  $t = 1, \dots, T^{(m)}/K$  and  $k = 1, \dots, K$ . By using enrollment data, the user-dependent prior score  $\tilde{\mu}_p^{(m)}$  and prior variance  $(\tilde{\sigma}_p^{(m)})^2$  are computed as follows:

$$\tilde{\mu}_p^{(m)} = \frac{K_c \tilde{\mu}_c^{(m)} + K_b \tilde{\mu}_b^{(m)}}{K_c + K_b} \quad (6)$$

and

$$(\tilde{\sigma}_p^{(m)})^2 = \frac{1}{K_c + K_b} \sum_{k=1}^{K_c + K_b} [\bar{s}^{(m,k)} - \tilde{\mu}_p^{(m)}]^2, \quad (7)$$

where  $K_c$  and  $K_b$  are respectively the numbers of client's enrollment utterances and pseudo-impostors' utterances,  $\tilde{\mu}_c^{(m)}$  and  $\tilde{\mu}_b^{(m)}$  are respectively the score means of client's and pseudo-impostors' utterances and  $\bar{s}^{(m,k)}$  denotes the mean score of the  $k$ -th enrollment utterance. Finally, the mean fused score

$$\hat{s}^{(m)} = \frac{K}{T^{(m)}} \sum_{t=1}^{T^{(m)}/K} \hat{s}_t^{(m)} \quad (8)$$

is used for decision making.

### 3. AUDIO-VISUAL BIOMETRIC AUTHENTICATION

This section explains how the multi-sample fusion techniques described in Section 2 can be applied to audio-visual biometric authentication.

#### 3.1. Audio-Visual Feature Extraction

##### 3.1.1. Audio-Visual Data Sets.

We used the XM2VTSDB corpus [7, 8] in our evaluations. XM2VTSDB is an audio-visual corpus for biometric research. The corpus consists of the audio and video recordings of 295 subjects taken over a period of four months. We adopted Configuration II as specified in [8] in the evaluation. More precisely, the database was divided into 200 clients, 70 impostors (part of the 95 impostors in DVD003b) for testing and 25 pseudo-impostors (the remaining impostors in DVD003b) for finding decision thresholds or other system parameters. For each client, the first two sessions were used for training, and the last session was used for testing. Each client was impersonated by 70 impostors using the audio and video data of the four sessions.

##### 3.1.2. Pre-processing of Audio Files.

As the original audio files were captured in a quiet, controlled environment using a high quality microphone, the equal error rate using the audio data alone is very low (about 0.7%). As a result, there is not much point in performing audio-visual fusion. Therefore, we introduce coder distortion and factory noise to the sound files in an attempt to simulate a more realistic acoustic environment.

The audio files in the corpus were down-sampled from 32kHz to 8kHz. Factory noise ("factory1.wav" of the NOISE92 database [9]) was added to the down-sampled files at a signal-to-noise ratio of 4dB. The noisy PCM files were transcoded by a GSM codec [10]. Twelve MFCCs and their time derivative (delta MFCCs) were extracted from the noisy, transcoded files using a 28ms Hamming window at a rate of 71Hz.

We used the training sessions of 200 client speakers in the speaker set to create a 128-center background model. The background model was then adapted to speaker models using MAP adaptation [6]. As defined in Configuration II of XM2VTSDB, two sessions (i.e. four utterances) per speaker were used for model training. Cepstral mean subtraction (CMS) was performed on all MFCCs before they were used for training, testing and evaluation.

##### 3.1.3. Pre-processing of Video Files.

Similar to audio files, the quality of video files in the corpus is also very good, making audio-visual fusion unnecessary (as face verification on the original video data already approaches 0% EER). As a result, we introduced distortion to the video sequences using PhotoShop Version 7.0 as follows. First, we converted each of the AVI files in the corpus into a sequence of high quality JPEG files with  $720 \times 576$  pixels. Second, we reduced the frame rate to one frame per second, and for each frame, down-sampled the JPEG images to  $176 \times 144$  pixels. Third, we blurred the images by setting the "Gaussian Blur" of PhotoShop to 1.0. Finally, we added Gaussian noise to the image by setting the "Gaussian Noise" of PhotoShop to 1.5. The noise-added image sequences were input to the Identix's Face Verification SDK [11] to locate the head and compute the scores. The scores have a range between 0 and 10. The higher the score, the more likely the claimant is genuine.

#### 3.2. Audio-Visual Multi-Sample Fusion

We assumed that in a verification session, we can obtain one utterance and one video shot from the claimant. We divided the utterance and the video shot into two equal-length sub-utterances and two equal-length sub-video shots, i.e.,  $K = 2$  and  $m \in \{A, V\}$  in Eq. 3, where  $A$  represents the audio channel and  $V$  the video channel. Feeding these sub-utterances and sub-video shots to the speaker verification system and the face verification system (FaceIT) [11] gives two streams of audio scores and two streams of visual scores. We applied intramodal fusion to the two audio score streams and also to the two visual score streams independently to obtain the mean of the fused audio scores,  $\hat{s}^{(A)}$ , and the mean of the fused visual scores,  $\hat{s}^{(V)}$ .

The client-dependent fusion parameters, including the prior scores and prior variances  $(\tilde{\mu}_p^{(m)}, (\tilde{\sigma}_p^{(m)})^2; m \in \{A, V\})$ , were obtained by feeding the utterances and video shots of 25 pseudo-impostors to the client and background models. The mean of intramodal fused scores (Eq. 8) was then compared with a client-independent decision threshold for decision making. There were a total of 400 client trials (200 clients  $\times$  2 utterances per client) and 120,000 impostor attempts (200 clients  $\times$  75 impostors per client  $\times$  8 utterances per impostor).

Given the mean fused audio score  $\hat{s}^{(A)}$  and visual score  $\hat{s}^{(V)}$ , the audio-visual score  $\hat{s}$  was obtained by linearly combining the two scores:

$$\hat{s} = \beta \hat{s}^{(A)} + (1 - \beta) \hat{s}^{(V)}, \quad (9)$$

where  $\beta$  is a combination weight, which can be computed using training data or made dependent on the quality of audio or visual data.

A system that uses a single client-independent decision threshold must ensure that all client and impostor scores have values comparable to the threshold. This requirement can be fulfilled by normalizing the scores so that they fall into a predefined range. One possible approach (called Z-norm [12]) is to shift the mean and scale the variance of the impostor scores to zero and unity, respectively. More specifically, the claimant's scores  $s_t^{(m,k)}$  (see Eq. 3) are normalized by

$$s_{t,norm}^{(m,k)} = \frac{s_t^{(m,k)} - \mu_b^{(m)}}{\sigma_b^{(m)}} \quad m \in \{A, V\},$$

where  $\mu_b^{(m)}$  and  $\sigma_b^{(m)}$  are respectively the mean and standard deviation of client-dependent impostor scores. These impostor scores can be obtained during training by testing a client model against impostor observations. In this work, the impostor observations were obtained from 25 pseudo-impostors defined in Configuration II of the XM2VTSDB corpus.

#### 4. RESULTS AND DISCUSSIONS

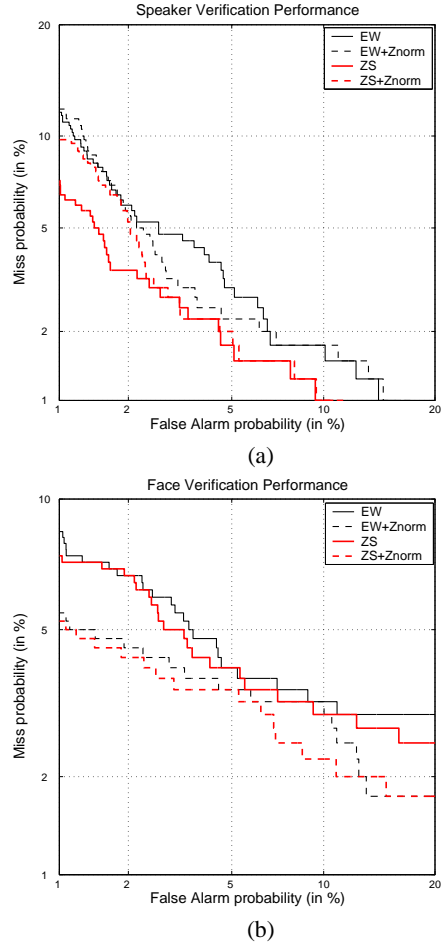
Table 1 shows the results of speaker verification and face verification using different types of intramodal multi-sample fusion techniques described in Section 2, and Fig. 1 plots the corresponding DET curves. We can observe that zero-sum fusion generally performs better than equal-weight fusion. We can also see that Z-norm helps lower the EER of equal-weight fusion but its effect on zero-sum fusion is not significant: Z-norm reduces the error rate of face verification but increases slightly the error rate of speaker verification.

Table 2 summarizes the results of audio-visual multi-sample fusion with  $\beta$  in Eq. 9 set to 0.6, and Fig. 2 plots the corresponding DET curves. As the ranges of audio and visual scores are different, Z-norm was applied to these scores independently so that the normalized audio- and visual scores can be fused. The results show that zero-sum fusion always performs better than equal-weight fusion.

In order to investigate the sensitivity of the fusion parameter  $\beta$  in Eq. 9, we varied the value of  $\beta$  and obtained the corresponding EERs. The results are shown in Table 3. Evidently, EERs depend on  $\beta$ . This calls for further investigation on automatic techniques for determining fusion parameters such as support vector machines, Bayesian classifiers, neural networks, etc. Some examples of this area can be found in [13, 14, 15, 16].

#### 5. CONCLUSION

This paper has presented an audio-visual biometric authentication system. A novel two-level fusion technique that fuses the scores obtained from speaker and face models was detailed. The proposed technique is general and is applicable to multi-modal biometric systems. This is evident by promising experimental results on the XM2VTSDB audio-visual database. It was found that an error rate reduction of up to 83% can be achieved when the proposed fusion technique is applied to fuse the scores derived from speaker models and face models.



**Fig. 1.** DET plots of different multi-sample fusion techniques in (a) speaker verification and (b) face verification. *EW* stands for equal-weight fusion and *ZS* stands for zero-sum fusion. *EW+Znorm* means that equal-weight fusion was performed on Z-norm scores. Similar definition applied to *ZS+Znorm*. For clarity, the labels in the legend are arranged in descending order of EERs.

#### 6. ACKNOWLEDGEMENT

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

#### 7. REFERENCES

- [1] J. Kittler, G. Matas, K. Jonsson, and M. Sánchez, "Combining evidence in personal identity verification systems," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 845–852, Sept. 1997.
- [2] 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.
- [3] 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, pp. 1681–1684.
- [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.

Fusion Method	Voice	Rel. Red.	Face	Rel. Red.
Equal-weight	3.99%	N.A.	4.45%	N.A.
Equal-weight+Znorm	3.04%	23.81%	3.55%	20.22%
Zero-sum	2.72%	31.83%	3.91%	12.13%
Zero-sum+Znorm	2.77%	30.58%	3.27%	26.52%

**Table 1.** Equal error rates (EERs) and their relative reduction (Rel. Red.) with respect to equal-weight fusion achieved by the speaker and face verification systems using intramodal multi-sample fusion. Note that fusion takes place only within the audio and visual scores, not between them. *Equal-weight+Znorm* (*Zero-sum+Znorm*) means that equal-weight fusion (zero-sum fusion) was performed on Z-norm scores.

Voice Score Fusion Type	Face Score Fusion Type	EER (%)	Relative Reduction (w.r.t. voice only)
EW+Znorm	EW+Znorm	0.70%	76.97%
EW+Znorm	ZS+Znorm	0.68%	77.63%
ZS+Znorm	EW+Znorm	0.61%	77.98%
ZS+Znorm	ZS+Znorm	0.47%	83.03%

**Table 2.** EERs and relative error reduction with respect to the EER of speaker verification (Table 1) obtained by linearly combining the means of intramodal fused scores. The combination weight  $\beta$  in Eq. 9 was set to 0.6. *EW* and *ZS* stand for equal-weight and zero-sum respectively.

$\beta$	0.4	0.5	0.6	0.7
EER	1.00%	0.80%	0.47%	0.59%

**Table 3.** EERs of multi-sample audio-visual fusion for different values of combination weights  $\beta$ . Zero-sum fusion of Z-norm scores (*ZS+Znorm*) was used for fusing the audio- and visual scores.

[5] 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.

[6] 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.

[7] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, Washington D.C., 1999.

[8] J. Luettin and G. Maitre, "Evaluation protocol for the extended M2VTS database," Tech. Rep., IDIAP, Martigny, Valais, Switzerland, Oct. 1998.

[9] [http://spib.rice.edu/spib/select\\_noise.html](http://spib.rice.edu/spib/select_noise.html), .

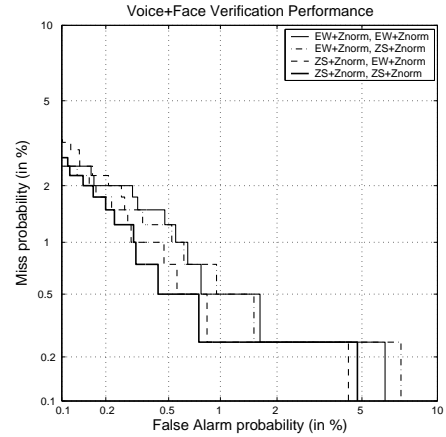
[10] European Telecommunication Standards Institute, *European digital telecommunications system (Phase 2); Full rate speech; Part 2: Transcoding (GSM 06.10 version 4.1.1)*, 1998.

[11] *SDK Developer's Guide FaceIT Verification*, Identix Incorporated, 2003.

[12] 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.

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

[14] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognition*, vol. 36, pp. 293–302, 2003.



**Fig. 2.** DET plots of different multi-sample fusion techniques for audio-visual person authentication. *EW* stands for equal-weight fusion and *ZS* stands for zero-sum fusion. *EW+Znorm* means that equal-weight fusion was performed on Z-norm scores. A similar definition applies to *ZS+Znorm*. For clarity, the labels in the legend are arranged in descending order of EERs.

[15] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Processing*, vol. 11, pp. 169–186, 2001.

[16] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 1065–1074, 1999.