INTRODUCTION OF QUALITY MEASURES IN AUDIO-VISUAL IDENTITY VERIFICATION

Meriem Bendris[1,2]*, Delphine Charlet*[1]

[(1)]France Télécom - OrangeLabs,
meriem.bendris@orange-ftgroup.com
delphine.charlet@orange-ftgroup.com

Gérard Chollet[2]

[(2)]CNRS LTCI,
TELECOM-ParisTech,
gerard.chollet@telecom-paristech.fr

ABSTRACT

Audiovisual identity verification exploits both image and audio information to improve the performance of the identification system. Unfortunately, both image and audio systems are sensitive to signal quality. In this paper, we propose a method to combine output classifiers based on both image and audio quality measures. We define classes of signal degradation within which we estimate the fusion weights and normalization parameters. Results of experiments on the BANCA database show that fusion using quality measures improves verification performance by 25% *compared to the baseline fusion method.*

Index Terms— Audiovisual identity verification, robust fusion, quality measures.

1. INTRODUCTION

Audiovisual identity verification integrates two complementary sets of information (audio and image) in order to increase the rate of correct identification. The identity verification systems based solely on audio modality are sensitive to the type of microphones used, the acoustic environment and the recording context. The face verification systems are sensitive to the image quality, lighting, background and appearance. Most of the fusion methods are based on a weighted sum fusion and require a representation of scores on the same scale (achieved by a normalization step). Introducing quality measures on the signal (audio and video) enable adaptation of the confidence given to each mono-modal classifier according to an automatic estimation of their reliability. In [1], the type of fusion rule (sum or product) depends on the quality of the input signal. In [2], the fusion framework makes a binary decision between speech and face classifiers according to quality measures. In [3], a quality-measure based on a weighted sum of the output classifiers is proposed, but without any experimental results. In our work, we do not only introduce a quality-measure based weighting function, we also take into account the dependency of the score normalization parameters on the quality-measure. To our knowledge, this study represents the first time that this dependency is introduced in the fusion framework.

This paper is organized as follows : Sections 2 and 3 present the speaker and face identity-verification systems. Section 4 presents, as a reference system, fusion based on linear combination of scores. Then section 5 and 6 present the proposed method for score fusion based on quality measures. Finally, experiments are reported in section 7.

2. SPEAKER VERIFICATION MODULE

The speaker verification module is based on a classical approach : a Gaussian mixture model with a universal background model (GMM-UBM) [4]. First, the *Mel Frequency Cepstral Coefficients* (MFCC) are extracted using a 32ms window with a 16ms step. In each frame, the extracted feature vector is composed of : the energy, the 13 first MFCC and the first and second MFCC derivatives (total dimension 42). Second, silence detection is performed based on a bi-class GMM (speech and non speech) of the MFCC coefficients. The Universal Background Model (UBM) has 256 components and is trained using the Expectation-Maximization (EM) algorithm. The speaker models are obtained by adapting the UBM to the speaker using a *Maximum A Posteriori* MAP-based method. In what followss, at the score calculation step (when there is an access request), the speaker claims to be the person λ. The MFCC feature vectors extracted from the test sequence X are compared to both λ-GMM and the UBM. The speaker verification model outputs an acoustic score S_s for the test utterance X as follows :

$$S_s(X, \lambda) = \log \frac{p(X|\lambda)}{p(X|\mathit{UBM})} \quad (1)$$

3. FACE VERIFICATION MODULE

The face verification module is also based on a classical approach : *eigenfaces* [5]. First, faces are detected as follows : the OpenCV library face-detector algorithm is used for each frame to propose regions in which it is likely to find a face. To improve the precision of the face detection, eye detection based on the same algorithm is used in the proposed regions. Consequently, the face is normalized so that the eyes are focused and aligned horizontally. An oval mask to remove pixel

background is also applied. The detected faces in the video can then be projected onto the face space, obtained by principal components analysis (PCA) following the principle of *eigenfaces*. For every detected face, the Euclidian distance from the face space (DFFS) between this face and its projection on the face space is computed. Only the best faces (small DFFS) are kept to describe the face appearing in the video sequence. Finally, we use the Mahalanobis distance to calculate the scores S_f.

4. BASELINE FUSION

The fusion module is based on a linear combination of the normalized speaker (S_s) and face (S_f) verification scores [6].

4.1. Score normalization

The scores obtained from the two verification modules are not represented in the same space. Merging scores by linear combination requires a normalization step. In [7], a comparative study is presented and shows that a robust normalization for sum fusion is the tangent hyperbolic normalization *tanh Norm* presented as follows :

$$\tilde{\mathrm{S}} = f_N(S, \mu_c, \sigma_c) = 0.5 + 0.5 \cdot tanh\left(0.01 \cdot \frac{S - \mu_c}{\sigma_c}\right) \quad (2)$$

where μ_c and σ_c represent the average and standard deviation of *Client* scores (from a development set).

4.2. Weighted sum fusion

After normalization of the scores, the fusion is performed according to equation 3. The estimation of the weights (w_f, w_s) is done with DCF (*Detection Cost Function*)[8] optimization using a development set.

$$\mathrm{S} = w_s \tilde{S}_s + w_f \tilde{S}_f \ \text{ with } \ w_f + w_s = 1 \quad (3)$$

Finally, the *Weighted sum fusion* has to estimate two types of parameters : normalization parameters (μ, σ) of the function f_N (for face and speech) and weights (w_f, w_s) of the fusion.

5. QUALITY MEASURES

It is often very difficult to define measures of signal quality as they remain very subjective. In the case of our study, the measures must represent confidence measures of the two identity verification modules *Speaker/Face*.

5.1. Audio quality

To measure the quality of audio sequences, the classical information used is the *Signal to Noise ratio*. In [9], the effects of audio degradation on identity verification, as seen through the SNR is demonstrated. The SNR measures the audio strength compared to background noise. A low SNR means a noisy signal while a high ratio indicates a clear audio. The estimation of SNR (expressed in dB) is calculated as follows :

$$q_s = \mathrm{SNR} = 10 \log_{10}\left(\frac{E_{speech}}{E_{noise}}\right) \quad (4)$$

E_{noise} and E_{speech} represent the average energy over all frames of the audio sequence detected as noise and speech respectively.

5.2. Image quality

Measures of image quality depend on the type of the database. Many measures are proposed in the literature, such as deviation from the frontal face *"frontal quality"* and illumination [1]. In our study, we are interested in image quality in terms of sharpness, an important propriety of images taken by Webcams (grainy image, vague..). Entropy, which is a measure of disorder, describe the image sharpness. A low entropy indicates a clear image. For each image, we compute the entropy of grayscale as follows :

$$q_f = \mathrm{Entropy} = -\sum_{i=1}^{256} P_i * log(P_i) \quad (5)$$

where P_i is the probability that a random pixel chosen from the image will have intensity *i* (256 gradients are used to encode the images).

6. QUALITY-BASED FUSION

The speaker and face verification systems are sensitive to the environment. The performance declines when the audio and image are noisy or have poor quality. It can be useful to adapt the confidence associated to each modality according to the conditions of recording. Moreover, the score distributions in each modality also depend on the quality of the input signal (see experimental results in figure 3) : thus, the score normalization parameters should also depend on the quality of the input signal. With these considerations, equation 3 becomes the following :

$$\begin{aligned} \mathrm{S} = {} & w_s(q_s, q_f) f_N(S_s, \mu_s(q_s), \sigma_s(q_s)) + \\ & w_f(q_s, q_f) f_N(S_f, \mu_f(q_f), \sigma_f(q_f)) \end{aligned} \quad (6)$$

To learn the fusion parameters with a dependence on quality measures, we represent the fusion parameters with a piecewise step function. First, we define intervals of signal degradation (M classes according to image and audio qualities) where we estimate the fusion parameters (weights and normalization parameters in development set). Then, a function

$C(q_s, q_f) = C_i$ depending on image and audio qualities is learned, in the development set, in order to automatically predict the classes C_i. The final score is calculated as follows :

$$\begin{aligned} \mathrm{S} = w_s(C_i) f_N(S_s, \mu_s(C_i), \sigma_s(C_i)) + \\ w_f(C_i) f_N(S_f, \mu_f(C_i), \sigma_f(C_i) \end{aligned} \quad (7)$$

The function $Ci = C(q_s, q_f)$ can be learned using standard classification methods such as a support vector machine (SVM), logistic regression or K-means.

7. EXPERIMENTS AND RESULTS

7.1. The BANCA database

The BANCA[10] audiovisual database was collected from 52 subjects speaking in English (26 men and 26 women). The population was separated into two groups G1/G2 of 26 individuals (13 men and 13 women). 12 sessions were recorded in 3 different conditions : *controlled* indicates recording in a controlled environment with a neutral background and a good camera, *degraded* indicates recording by a webcam in a somewhat noisy environment (offices) and the condition *adverse* indicates recording in a dining hall with a lot of background noise but good camera. According to the P protocol (*pooled*), the database consists of 78 access clients and 104 access impostors for each condition. We choose to use each group as the development set of the other. For the speaker verification module, UBM is learned from a very large database from *NIST*[8] evaluation (English speech). Then using a MAP-based algorithm , the UBM is adapted to the speech variability in Banca with a recording of 10mn30s of different persons provided in the database. The results will be evaluated with the Equal Error Rate value *EER*.

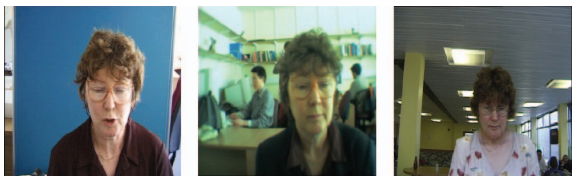
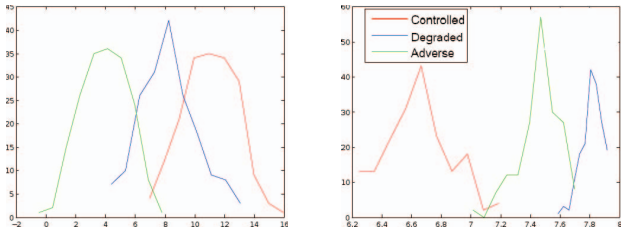


Fig. 1. BANCA example of the images collected in the 3 conditions *Controlled*, *Degraded* and *Adverse*

7.2. Quality measure

Figure 2 shows the histogram of audio SNR and image entropy values for each condition in Banca (*Controlled, Degraded and Adverse*). The conditions are quite separable and sorted by order of recording quality. Note in figure 2.b that image entropy indicates the degradation levels, showing that the *controlled* condition is the best, followed by the *Adverse* and then *degraded* condition recorded by a webcam. Ultimately, figure 2 illustrates that audio SNR and image entropy are good indicators of image/audio quality in Banca.

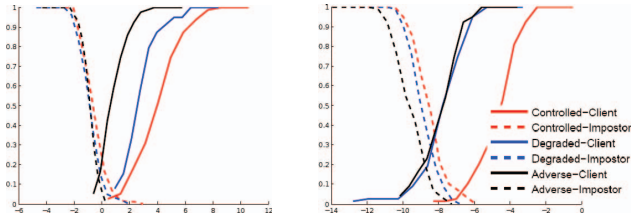


(a) (b)

Fig. 2. Histograms of the (a) SNR and (b) image Entropy for each condition in group 1 of BANCA.

7.3. Score distributions

Figure 3 represents the cumulative histograms of the *Client/Impostor* scores for each condition in Banca. For the client access, the experimental results show that the output of verification systems is very dependent on the quality of the signal concerned. So, the normalization parameters vary depending on these conditions and it is necessary to adapt them to the signal quality.



(a) (b)

Fig. 3. Cumulative histograms of (a) speaker and (b) face scores fo each condition in group 2 of BANCA.

7.4. Experimental results on Fusion

7.4.1. Protocol

The quality-based fusion is carried out following equation 7. The classes of signal degradation are already labeled in Banca database (M=3) according to the conditions :*controlled, degraded and adverse*. Using the SVM[1] algorithm, the function C is learned on a development set to predict the conditions in Banca through the measures (q_s, q_f). Correct classification rate is on average 82%. The estimation of the fusion parameters (normalization and weights) is already done for each condition in the development set. To demonstrate the contribution of quality-dependent normalization parameters, an experiment was also conducted in witch only the dependence of the weights on the image and audio qualities is taken into account :

$$\begin{aligned} \mathrm{S} = w_s(C_i) f_N(S_s, \tilde{\mu_s}, \tilde{\sigma_s}) + \\ w_f(C_i) f_N(S_f, \tilde{\mu_f}, \tilde{\sigma_f}) \end{aligned} \quad (8)$$

[1] http ://svmlight.joachims.org/.

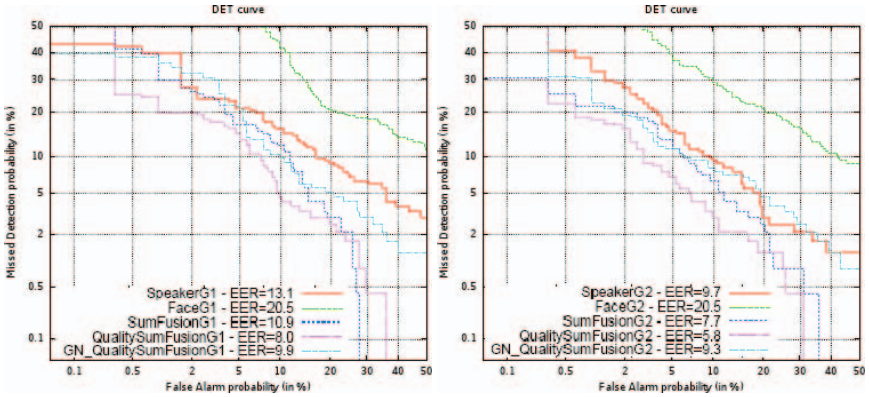


Fig. 4. The performance of the fusion systems

The normalization parameter $\tilde{\mu}$ et $\tilde{\sigma}$ are estimated using the total development set. Finally, it is important to note the problem of scarcity of Banca data in estimating the parameters for each condition (78 client accesses, 104 impostor accesses).

7.4.2. Results

Figure 4 summarizes the performance of the fusion methods. The baseline fusion (*SumFusion*) shows a large improvement over the single modality (Face or Speaker). The system introducing quality-dependent weighting without quality-dependent score normalization (*GN_QualitySumFusion*) does not yield consistent results (improvement on group 1, degradation on group 2), whereas considering the whole set of dependencies (weight and score normalization parameters in *QualitySumFusion*) leads to a relative improvement of 25% of the *EER* on both groups, compared to the fixed fusion scheme.

8. CONCLUSION AND FUTURE WORK

In this paper, we have presented a method to integrate automatic quality measures of the signal in a fusion system for audio-visual identity verification. The proposed method uses quality-based weighting for the classifier output and also takes into account the dependency of the score distribution on the signal quality by introducing quality-dependent score normalization parameters. The dependency of fusion parameters on audio and visual quality measures is modeled through a piecewise function that corresponds to three different conditions of recording, which are unknown during the test. A significant improvement (25% relative reduction of error rate) is observed on Banca with the proposed method, compared to a quality-independent fusion scheme. In future work, we intend to investigate other quality measures (frontal quality for instance) as well as others kinds of functions to model dependencies on audio and image quality.

9. REFERENCES

[1] Norman Poh, Josef Kittler, and Omolara Fatukasi. Quality controlled multimodal fusion of biometric experts. *in 12th Iberoamerican Congress on Pattern Recognition CIARP, pp. 881-890*, 2007.

[2] Krzysztof Kryszczuk, Jonas Richiardi, Plamen Prodanov, and Andrzej Drygajlo. Error handling in multimodal biometric systems using reliability measures. *In 13th European Signal Processing,EUSIPCO*, 2005.

[3] Norman Poh and Samy Bengio. Improving fusion with margin-derived confidence in biometric authentication tasks. *AVBPA, LNCS 3546*, 2005.

[4] Quatieri Thomas F. Reynolds, Douglas A. and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing Volume 10, Issues 1-3, Pages 19-41, January 2000.*

[5] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1) :71–86, 1991.

[6] H. Bredin and G. Chollet. Making talking-face authentication robust to deliberate imposture. *ICASSP 2008.*

[7] Hongzhou Zhang Lin Wang Chengbo Wang, Yongping Li. Multi-modal biometric verification based on far-score normalization. *International Journal of Computer Science and Network Security (IJCSNS)*, April 2008.

[8] A. Martin and M. Przybocki. The nist 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 2000.

[9] N.A. Fox, R. Gross, J.F. Cohn, and R.B. Reilly. Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts. *IEEE Transactions on Multimedia*, 9(4) :701–714, 2007.

[10] V.Popovici et al. The banca database and evaluation protocol. *AVBPA, LNCS 2688*, 2003.