Fifth International Symposium on Signal Processing and its Applications, ISSPA '99, Brisbane, Australia, 22-25 August, 1999
Organised by the Signal Processing Research Centre, QUT, Brisbane, Australia

MULTI-MODAL PERSON VERIFICATION SYSTEM BASED ON FACE PROFILES AND SPEECH

Conrad Sanderson and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
e-mail: C.Sanderson@me.gu.edu.au, K.Paliwal@me.gu.edu.au

ABSTRACT

This paper describes a person verification system based on facial profile views and features extracted from speech. The system is comprised of two non-homogeneous classifiers whose outputs are fused after a normalization step. Experiments are reported which show that integration of the face profile and speech information results in superior performance to that of its subsystems. Additionally, the performance of the combined system in noisy conditions is shown to be more robust than the speech-based subsystem alone.

1. INTRODUCTION

A person verification system attempts to verify the claimed identity of an individual. This can be useful in situations where security considerations preclude obtaining access by simpler means such as a key. Many person verification systems are described in the literature, relying on features derived from speech [1]. However, these systems can easily fail in the presence of background noise. In this paper a multi-modal person verification system is presented which relies on the shape of the profile of a person's head as well as the speech uttered by that person. The system is made up of a Profile Verification System (PVS), a Speaker Verification System (SVS) and a Fusing and Classification Module (FCM). The voice and visual cues are combined by the FCM allowing the resulting system to have superior performance, as shown in the experimental section, than either of its subsystems alone. The performance and robustness of the SVS and the combined system are compared in noisy conditions, to simulate real life conditions.

The paper is organized as follows: Section 2 describes the system architecture, Section 3 shows the setup for experiments, and Section 4 presents the results.

2. SYSTEM ARCHITECTURE

As stated before, the system is made up of 3 modules:

- Speaker Verification System
- Profile Verification System
- Fusing and Classification Module

The SVS used is based on the Gaussian Mixture Model (GMM) approach [1]. The speech signal, sampled at 16 kHz and quantized over 16 bits, is analyzed every 10 msec using a 20 msec Hamming window. For each window (also referred to as a *frame*), the energy is measured, and if it is above a set threshold (corresponding to voiced sounds), 12th order cepstral parameters are derived from Linear Prediction Coding (LPC) parameters [2]. Each set of extracted parameters can be treated as a 12-dimensional vector. During the training phase of the system, a 12-dimensional, 4-mixture GMM is computed for each speaker using parameters extracted from the speech signal.

For testing of the SVS, the same process of feature extraction is performed. Using a GMM, belonging to the person whose identify is being claimed, a similarity measure is computed by averaging the log-likelihood of individual frames. If the average log-likelihood is above a certain threshold, then the identity of the speaker is verified.

The PVS used is very similar to the one described in [3]. Given a head shot of a person who is facing sideways (see Figure 1), the head is extracted from the background, and then the profile is extracted from the head. The profile is refined by searching for the nose and then depending on the hair style and amount of facial hair present, an unoccluded portion of the profile is used. Using this refined profile, a distance map [4] (see Figure 2) is calculated and stored with the profile.

For testing of the PVS, the profile is extracted as previously. To compare one profile against another, it is necessary to account for possible tilt, translation and scale of the profile. Initially the profile is superimposed over the distance map belonging to the profile of the person whose identity is being claimed, with the noses aligned and scales roughly adjusted. Distance is computed by summing up all

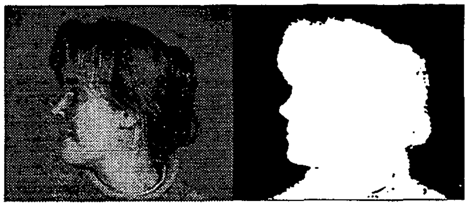

Figure 1: Example of a profile shot (*mu 1*) extracted from the M2VTS database (left), and head segmentation (right).

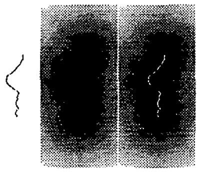

Figure 2: Profile extracted from Figure 1 (left), its distance map (center), the profile superimposed on the distance map (right).

distance values found where the profile's pixels are present within the distance map. The downhill simplex algorithm [5] is employed to minimize this distance by automatically adjusting parameters for an affine transform of the profile, ie. scale, translation and rotation (within preset limits). The residual distance between the compensated profile and the distance map can be used to decide whether the profile belongs to the person whose identity is being claimed. If the distance is below a certain threshold, the person is deemed to be verified. The process of comparing profiles is referred to as *matching*.

The FCM uses raw scores from the subsystems rather than relying on them for classification - this method is often referred to as *soft fusion*. FCM's first job is to reverse the sign of the value coming from the SVS in order to make it compatible with the PVS. To prevent the PVS from dominating, the value coming from it is limited to a preset maximum. The FCM then normalizes the values from each of the subsystems by making them zero mean and unity variance, and then placing them in the [0,1] interval. The mean and variance values used during this process must be estimated by first running the subsystems on training data and analyzing their probability density functions (PDFs).

Finally the normalized values can be combined:

$$f = w * p_n + (1 - w) * s_n$$

where w is a weight factor between 0 and 1, p_n = normalized distance value from the PVS, s_n = normalized negative log-likelihood value from the SVS. If f is below a predefined threshold, then the person requesting access is accepted.

3. EVALUATION OF PERFORMANCE

3.1. Multi-modal Database

The M2VTS database [6] has been used for evaluating the combined system. It is comprised of 37 people counting from zero to nine (mostly in French) and facing the camera. The database is made up of 5 sections, each with video sequences for each person. From section to section, the video sequences often differ in hair styles, clothes, lighting conditions and zoom factors. For each video sequence, a synchronized speech signal sampled at 44 kHz with 16 bit resolution is available. There are additional video sequences where each person rotates their head from one side to the other. If the person is wearing glasses, another head moving sequence is available without them.

Profile shots were obtained by manually finding the frames in head rotating sequences where the person is facing left and not wearing glasses. Each frame has a resolution of 350x286 pixels. Figure 1 presents an example frame.

3.2. Experiment Setup

For each person, speech files and video sequences from the first four sections are used for experiments. Sections 1 to 3 are used for training, while section 4 is used for testing. Profiles extracted from the first three sections are used to select the best representative profile during the training session. The database allows for 37 correct verification trials and 37*36 impostor trials.

3.3. Training Setup

For the SVS, the speech files are downsampled to 16 kHz at 16 bit resolution. The training session is the same as described in Section 2.

There are three matching operations for the training of the PVS. For each person, profile from section 1 (P1) is matched with P3, P2 with P1 and P3 with P2. The profile that appears in the 2 best matchings is selected as the reference profile.

Figures 3 and 4 show the PDFs of the SVS and PVS scores. In order to fuse these scores in the FCM, we need the mean (μ) and standard deviation (σ^2) values of these PDFs. These are estimated with the following procedure: both of the subsystems are trained and tested on the training sections of the database. Outliers must first be removed since they reduce the reliability of estimation of μ and σ^2. For the SVS, an adequate method of outlier removal is by finding the median (m) and the deviation from the median ${\sigma_m}^2$ (same as standard deviation, except substituting the median for the mean). Any value which is outside of the interval defined by $m \pm 2 * {\sigma_m}^2$ is ignored. For the PVS,

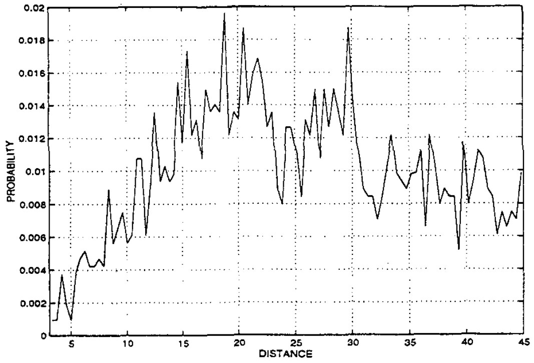

Figure 3: PDF of the PVS score.

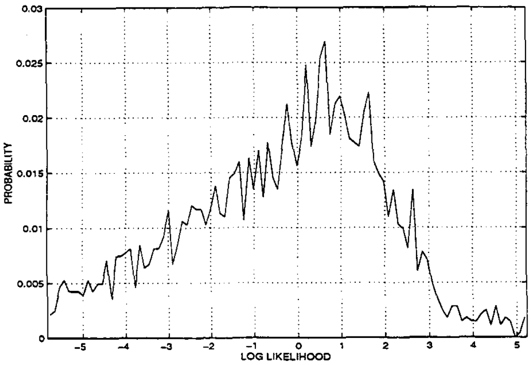

Figure 4: PDF of the SVS score.

ignoring values greater than a predefined maximum proved to be sufficient method for removing outliers.

After outlier removal the values from the SVS are changed in polarity in order to make them compatible with the PVS, as this is required by the FCM. The μ and σ^2 for the PDF of the SVS were set to the median and deviation from the median, respectively, as they were found to improve the performance of the system.

4. RESULTS

Four experiments were performed. For a given decision threshold, False Acceptance (FA) and False Rejection (FR) rates were calculated. For each experiment, a Receiver Operating Characterstics (ROC) curve was generated by varying the decision threshold continuously. Figure 5 shows the ROC curve with $w = 1$.

A good way to evaluate the performance of a verification system is by computing the equal error rate (EER), where $FA = FR$, the success rate (SR), where $1 - FA - FR$

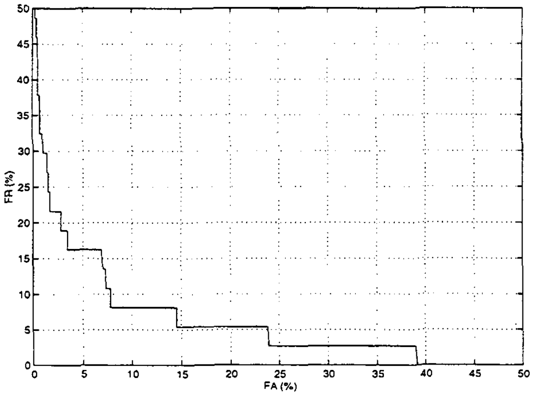

Figure 5: ROC curve of the PVS subsystem, ie. $w = 1$.

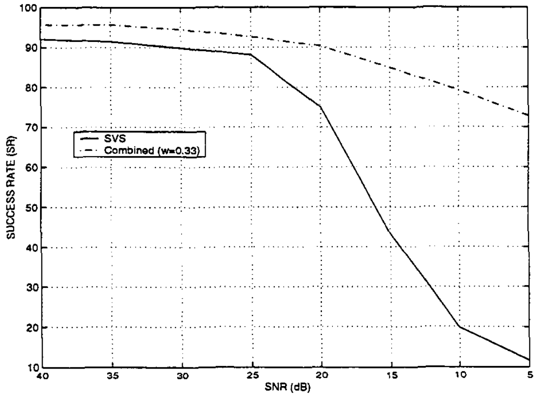

Figure 6: Success Rate of the SVS compared to the combined system ($w = 0.33$) with decreasing SNR.

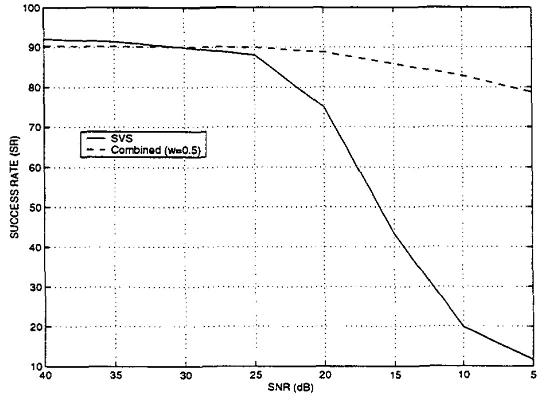

Figure 7: Success Rate of the SVS compared to the combined system ($w = 0.5$) with decreasing SNR.

reaches a maximum, and the FR for an FA of 1%.

In the first experiment, w was varied from 0 to 1. The results are shown in Table 1. For $w = 0$, only the SVS was used, while for $w = 1$ only the PVS was used, hence it can be seen that the SVS has better performance than the PVS. For $w = 0.33$, the combined system outperforms both of the two subsystems.

In the second experiment, with $w = 0$, the speech was progressively corrupted by lowering the Signal to Noise Ratio (SNR) from 40dB to 5dB. The results are shown in Table 2 and Figure 6. The third experiment is a repeat of the 2nd experiment, but with $w = 0.33$. The results are shown in Table 3 and Figure 6. The fourth experiment is also a repeat of the 2nd experiment, this time with $w = 0.5$. Results are shown in Table 4 and Figure 7.

As it can be seen, when $w = 0.33$, the combined system outperforms the SVS for all SNRs. For $w = 0.5$, the SVS initially outperforms the combined system, however its performance drops rapidly with decreasing SNR. This is in contrast to the combined system, where the performance curve has a much more graceful dropoff. The SR at 10dB and lower of the combined system with $w = 0.5$ is better than with $w = 0.33$, hence there is a trade-off between lower performance at high SNRs versus more robust performance at low SNRs.

w	SR	$FR_{FA=1\%}$	EER
1.0	84.08	29.73	8.11
0.66	88.74	19.92	8.15
0.5	90.47	16.22	5.41
0.33	95.50	8.11	2.70
0.0	92.49	16.22	5.52

Table 1: Performance of the combined system, for varying weight factors.

$SNR\,(dB)$	SR	$FR_{FA=1\%}$	EER
40	92.04	18.92	5.40
35	91.37	21.62	5.37
30	89.87	21.62	5.52
25	88.06	37.84	8.15
20	75	64.87	13.55
15	43.32	91.89	29.69
10	19.82	100	45.38
5	11.64	100	50.75

Table 2: Performance of the SVS, quoted in %, with decreasing SNR (see also Figure 6).

$SNR\,(dB)$	SR	$FR_{FA=1\%}$	EER
40	95.57	8.11	2.74
35	95.57	8.11	2.74
30	94.44	8.11	2.78
25	92.57	13.51	5.41
20	90.32	16.22	5.44
15	84.91	32.43	8.63
10	79.20	67.57	13.51
5	72.82	75.68	16.22

Table 3: Performance of the combined system with $w = 0.33$, quoted in %, with decreasing SNR (see also Figure 6).

$SNR\,(dB)$	SR	$FR_{FA=1\%}$	EER
40	90.31	16.22	5.40
35	90.24	16.22	5.44
30	90.09	16.22	5.37
25	89.94	18.92	5.71
20	88.81	18.92	8.11
15	85.89	24.32	8.15
10	82.81	43.24	10.81
5	78.75	59.46	10.81

Table 4: Performance of the combined system with $w = 0.5$, quoted in %, with decreasing SNR (see also Figure 7).

5. CONCLUSION

The results presented support the use of multi-mode, based on profile views and speech, person verification systems. It was demonstrated that a combined system outperforms a speaker verification system, and is much more robust in noisy conditions.

6. REFERENCES

[1] Douglas A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication* 17, 1995, pages 91 - 108.

[2] K. K. Paliwal, "Speech processing techniques", *Advances in Speech, Hearing and Language Processing*, Vol. 1, 1990, pages 1 - 78.

[3] Stephane Pigeon, Luc Vandendorpe, "Profile Authentication Using a Chamfer Matching Algorithm", *Audio- and Video-based Biometric Person Authentication - proceedings of AVBPA'97*, Crans-Montanta, Switzerland, March 12-14, Josef Bigun et al. (ed), Springer, 1997, pages 185 - 192.

[4] Gunilla Borgefors, "Hierarchical Chamfer Matching: A Parametric Edge Matching Algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 6, Nov. 1988, pages 849 - 865.

[5] William H. Press et al., *Numerical Recipes in C*, 2nd ed., Cambridge, New York, Cambridge University Press, 1992, pages 408 - 412.

[6] M2VTS Database: *http://www.tele.ucl.ac.be/M2VTS/*