ADAPTIVE MULTI-MODAL PERSON VERIFICATION SYSTEM

Conrad Sanderson and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
{c.sanderson, k.paliwal}@me.gu.edu.au

ABSTRACT

This paper describes an adaptive multi-modal person verification system based on speech and face images. The system adapts to noise present in the speech signal by modifying the parameters of the fusion module. Linear and Support Vector Machine (SVM) based techniques of fusing the similarity measures from speech and face modes are investigated. Experimental results obtained on the Digit Database show that the adaptive system significantly outperforms its non-adaptive counterpart.

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. Recently, multi-modal person verification systems have become popular [1], where similarity measures from different modality experts are fused before the final decision to accept or reject a claimant is made. The attraction of multi-modal systems stems from their ability to have better performance than the individual modality experts. While the robustness of multi-modal systems is better than uni-modal systems, their performance still degrades significantly in presence of noise [2].

In this paper, we present a multi-modal system based on face images and speech signals which adapts itself to the amount of noise present in speech, leading to an improvement in performance for varying noise conditions.

2. DIGIT DATABASE

We have created a database to carry out experiments for person identification/verification using speech and video information. The database is comprised of video and corresponding audio recordings of 37 subjects (16 female and 21 male), divided into 3 sections. Sections 1, 2 and 3 are respectively referred to as the *training*, *validation* and *testing* sections. While wearing different clothes for each section, the subjects were asked to perform the following:

- 20 repetitions of "0 1 2 3 4 5 6 7 8 9" with a small pause between each digit (*digit sequence*),
- recite "he played basketball there while working toward a law degree" (*word sequence*),
- recite "5 0 6 9 2 8 1 3 7 4" (*alternate sequence*), and
- move their head left to right, then up and down, with a pause in the center before each movement (*head rotation*)

The recording was carried out over a period of one week, in a TV studio using a broadcast quality digital camera, a low-noise directional microphone positioned above each subject, 2 overhead lights on either side of the subject (with 2 light diffuser screens) and a blue background lit by 3 overhead lights. Automatic audio gain was disabled as was auto focus. Subjects were asked to sit on a chair which was 3 meters away from the camera. The video was transferred to a PC and edited, consisting of storing each sequence of numbers (or words) individually. To make the size of the video data more manageable, the sequences were converted from DV format (720x576, 25 fps) to a sequence of still images saved as JPEG files. Each frame was downsampled by a factor of 2 and cropped to resolution of 280x260. A high quality setting was used for the creation of JPEG files. Audio sequences were converted from 48 kHz, 16-bit stereo to 32 kHz, 16-bit mono. In total, the database occupies approximately 7 Gigabytes. To obtain a copy of the Digit Database 1.0, please see our web page [1] or contact us.

3. SPEECH MODALITY EXPERT

The speech modality expert is based on the Gaussian Mixture Model (GMM) approach [3]. The speech signal is downsampled to 16 kHz followed by removal of silent or noise parts. The signal is then parametrized into cepstral coeffients derived from Linear Prediction Coding (LPC) parameters [4]. We use a 20 ms Hamming window with a 10 ms frame

[1] http://spl.me.gu.edu.au/digit/

interval, and an analysis order of 12. Deltas are added, thus resulting in 24-dimensional feature vectors.

Client models are generated by pooling training data for a given person and constructing an 8-mixture GMM using a modified k-means algorithm. During a test session, the speech modality expert, using the GMM of the claimed identity, provides a similarity measure obtained by averaging the log-likelihood of the feature vectors of given utterances.

4. FACE MODALITY EXPERT

Colour face images are first converted into greyscale. Then, by using template correlation, the location of the face is found. With correlation constrained to specified areas, eyes and nose are found. Using the distance between the eyes, and the distance between the eye line and the nose, an affine transform was employed to normalize the size. Tilt was not taken into account. Based on the location of the eyes, a 75x95 window was extracted from the normalized image. Brightness was normalized by using the median value of the pixels inside the window as a brightness measure.

By concatenating the rows of a given face image, each face is represented by a 7125-dimensional vector. Principal Component Analysis (PCA) [5] is used to make a 50-dimensional representation [6].

Like the speech modality expert, client models for the face modality expert are generated by pooling training data for a given person and constructing a single mixture GMM. During a test session, the face expert, using the GMM of the claimed identity, provides a similarity measure by averaging the log-likelihood of feature vectors of given face images.

5. EXPERT FUSION MODULE

5.1. Likelihood normalization

The log-likelihood values from the above experts have different ranges and hence cannot be fused directly. They are mapped to a common interval, [0,1], by the following procedure: The median (μ_m) and the variance from median (σ_m^2) of the likelihood values of correct claimants are found by testing each expert on the validation section of the database. Assuming the values for impostors and correct claimants follow Gaussian distributions $\mathcal{N}(\mu_m - 4 * \sigma_m, \sigma_m^2)$ and $\mathcal{N}(\mu_m, \sigma_m^2)$ respectively, 95% of the values lie in the $[\mu_m - 6*\sigma_m\ ,\ \mu_m - 2*\sigma_m]$ and $[\mu_m - 2*\sigma_m\ ,\ \mu_m + 2*\sigma_m]$ intervals, respectively. Subtraction of $\mu_m - 2*\sigma_m$ from all values, then division by $2*\sigma_m$, results in approximate mapping to [-2,0] and [0,2], respectively. The [-2,2] interval corresponds to the approximately linearly changing portion of the sigmoid function, $f(x) = \frac{1}{1+exp(-x)}$, used to finally map the values to the [0,1] interval.

5.2. Linear Fusion

Let f and s be the normalized log-likelihood values from the face and speech modality experts, respectively. These likelihoods can be fused into a single value using a simple linear relation:

$$x(f, s, w) = wf + (1 - w)s \tag{1}$$

where $\{w : [0, 1)\}$ is the weight factor assigned to the face modality expert. Given a decision threshold, t, the claimant is rejected if $x(f, s, w) < t$. Otherwise, the claimant is accepted.

Treating the normalized log-likelihood values as points in 2-D space, equation (1) translates to a linear decision boundary (see Figure 1) described by:

$$\frac{-w}{1-w} f + \frac{t}{1-w} - s = 0 \tag{2}$$

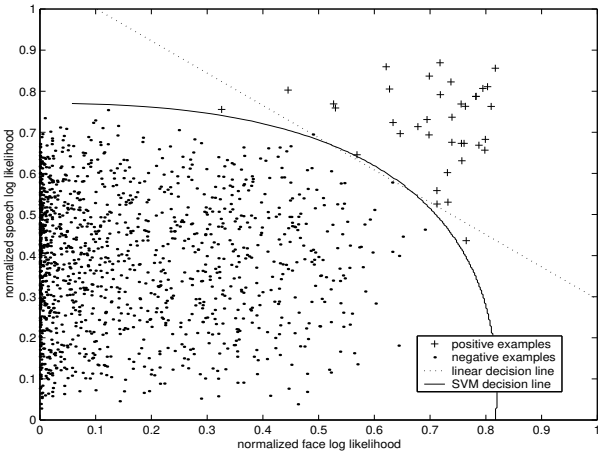

Figure 1: Distribution of face and speech normalized log-likelihood values on validation data for PSNR of 24 dB. Linear and SVM decision lines are shown.

5.3. SVM Fusion

The Support Vector Machine (SVM) is based on the principle of structural risk minimization [7] as opposed to empirical risk minimization used in classical learning approaches. Given a data set with n-dimensional points belonging to two different classes $+1$ and -1, a function is found that maps the points from their data space to their label space. Since a thorough description of SVM is beyond the scope of this paper, the reader is encouraged to see [8]. We have used the *SVM-light* toolkit [9] in this work. In the training process, examples of correct claimants and impostors were labelled as classes $+1$ and -1 respectively. The polynomial kernel with the default settings was used. During testing, claimants where the SVM result was above 0 were

accepted, otherwise they were rejected. An example of the decision line made by SVM is shown in Figure 1.

6. ADAPTATION

In a traditional multi-modal verification system, the similarity measures from modality experts are fused to obtain best possible performance when the training and testing conditions are matched. If one expert is more susceptible to noise, an intuitive improvement is to emphasize the expert less affected by noise during fusion. However, it has been shown that this can degrade the performance of the system in conditions where there is different amount of noise present than anticipated [2] since the latter expert has worse performance.

We propose an adaptive system where the parameters of the fusion module are made dependent on the Peak Signal to Noise Ratio (PSNR) of the speech signal. A set of parameters, for varying PSNR levels, is estimated a priori during the training stage. During system usage, the PSNR of the given speech signal is estimated and parameters most closely corresponding to that PSNR are used by the fusion module.

An estimate of the peak signal to noise ratio (PSNR) is obtained by using the following procedure: Divide the signal into 20ms frames with an overlap of 10ms. For each frame calculate the power. Select about 25 frames with the lowest power and take their mean power as the noise power. Select 100 frames with the highest power and take their mean power as signal power. Ratio of signal and noise powers in dB provides an estimate of PSNR.

7. PERFORMANCE EVALUATION

7.1. Performance Criteria

The basic error measures of a verification system are false acceptance rate, F_A (in %), and false rejection rate F_R (in %). By varying the parameters of the technique used in expert fusion, one can obtain a F_R value for a given F_A. To evaluate the performance of the system, we have chosen an operating point of $F_A < 0.1\%$, which simulates real life applications. In this work there were 37 tests for correct claimants and 36*37 tests for impostors.

7.2. Speech Data Preparation

Due to the nature of the audio recording in the Digit Database, the loudness of speech varies between subjects, while the noise level stays constant. All speech files were first normalized to have a PSNR of 24 dB by adding white gaussian noise. Versions with a PSNR ranging from 22 dB to 10 db were generated similarly.

7.3. Training

The speech expert was trained on normalized digit sequences with a PSNR of 24 dB from the training section. The face expert was trained on 1000 images per person from the training section.

The expert fusion module was trained on the validation section. For a given PSNR, ranging from 24 to 10 dB, a set of parameters was found that optimized the performance for a given technique.

7.4. Results

Experiments were performed where the system was tested on the validation and testing sections, in adaptive and non-adaptive setups and varying the technique used. In adaptive operation, the parameters used by the fusion technique were updated depending on the PSNR of the speech file, while in non-adaptive operation the parameters were fixed to the ones found for speech data with PSNR of 24 dB. Results are presented in Tables 1 and 2 and Figures 2 and 3. Since the system is trained for $F_A < 0.1\%$, the corresponding Figures use $[{F_A}^2 + (\frac{F_R}{10})^2]^{\frac{1}{2}}$ to emphasise the F_A result.

The adaptive systems for both techniques outperform the non-adaptive counterparts in almost all cases. Interestingly, the performance for the two techniques in adaptive systems is quite similar. As the PSNR decreases, the performance of the adaptive systems remains relatively constant, while it rapidly deteriorates for the non-adaptive cases, especially for SVM.

The decision lines made by SVM are more data dependent than the linear case. In the presence of noise, the distribution of similarity measures moves significantly, hence SVM's greater sophistication works against generalization over varying PSNR levels. In contrast, the linear technique's simplicity translates to better generalization over varying PSNR levels.

8. CONCLUSION

We have described an adaptive multi-modal person verification system based on speech and face images. By using an estimate of the Peak Signal to Noise Ratio, the system adapts to noise present in the speech signal by selecting the parameters of the fusion technique best matched to given noise conditions. Fusion of the similarity measures from modality experts was accomplished using linear and Support Vector Machine (SVM) techniques with both techniques exhibiting similar performance. The adaptive system significantly outperformed its non-adaptive counterpart, especially at low PSNR levels. In non-adaptive cases, the linear technique was found to outperform the SVM.

$PSNR\,(dB)$	Adaptive		Non-Adaptive	
	F_A	F_R	F_A	F_R
24	1.35	8.11	1.35	8.11
22	0.75	8.11	4.280	0
20	0.45	8.11	9.91	0
18	0.30	8.11	17.12	0
16	0.15	21.62	24.55	0
14	0.15	21.62	30.86	5.41
12	0.15	24.32	36.49	8.11
10	0.15	24.32	39.72	5.41

Table 1: SVM performance on test data.

$PSNR\,(dB)$	Adaptive		Non-Adaptive	
	F_A	F_R	F_A	F_R
24	0.53	8.11	0.53	8.11
22	0.23	10.81	0.98	2.70
20	0.08	10.81	2.48	2.70
18	0.30	13.51	4.43	2.70
16	0.15	24.32	6.46	2.70
14	0.15	24.32	7.81	5.41
12	0.15	27.03	8.78	8.11
10	0.15	24.32	10.06	10.81

Table 2: Linear technique performance on test data.

9. ACKNOWLEDGMENTS

Many thanks to Claire Fletcher, Brett Wildermoth, Dave Strobel, Gavin De Zilva, George Slavov and all the volunteers for their help in making of the database, as well as Dr Jingdong Chen, Farshid Golchin and Seppo Saario for their suggestions.

10. REFERENCES

[1] S. Ben-Yacoub, Y. Abdeljaoued, E. Mayoraz, "Fusion of Face and Speech Data for Person Identity Verification", *Proceedings of the IEEE Transactions on Neural Networks*, Vol. 10, No. 5, Sept. 1999, pages 1065 - 1074.

[2] C. Sanderson, K. K. Paliwal, "Multi-Modal Person Verification System Based on Face Profiles and Speech", *Proc. Fifth International Symposium on Signal Processing and its Applications (ISSPA'99)*, Brisbane, Australia, Aug. 1999, pages 947 - 950. *http://spl.me.gu.edu.au/papers/cs/isspa99.ps.gz*

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

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

[5] K. K. Paliwal, "Dimensionality Reduction of the Enhanced Feature Set for the HMM-Based Speech Recognizer", *Digital Signal Processing 2*, 1992, pages 157 - 173.

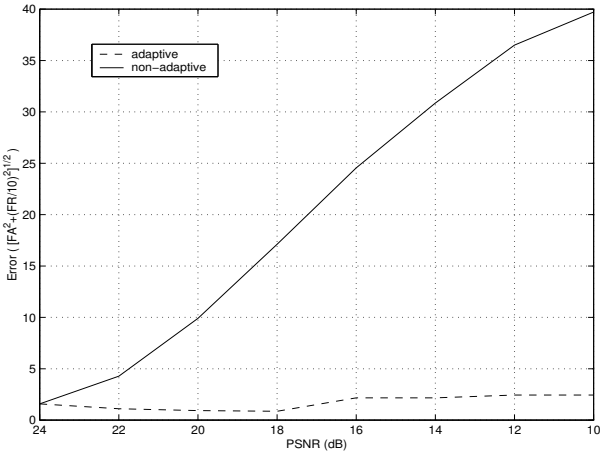

Figure 2: SVM performance on test data.

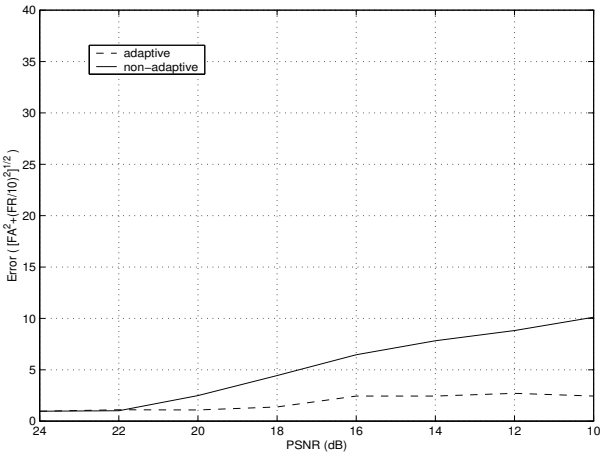

Figure 3: Linear classfier performance on test data.

[6] Matthew Turk, Alex Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, 1991, pages 71 - 86.

[7] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[8] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, Vol. 2, No 2, 1998, pages 121 - 167. *http://svm.research.bell-labs.com/papers/tutorial_web_page.ps.gz*

[9] T. Joachims, "Making large-Scale SVM Learning Practical", *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges and A. Smola (ed.), MIT-Press, 1999. *http://www-ai.cs.uni-dortmund.de/DOKUMENTE/joachims_99a.pdf*