NOISE COMPENSATION IN A MULTI-MODAL 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

In this paper we propose an adaptive multi-modal verification system comprised of a modified Minimum Cost Bayesian Classifier (MCBC) and a method to find the reliability of the speech expert for various noisy conditions. The modified MCBC takes into account the reliability of each modality expert, allowing the de-emphasis of the contribution of opinions from the expert affected by noise. Reliability of the speech expert is found without directly modeling the noisy speech or finding the reliability a priori for various conditions of the speech signal. Experiments on the Digit Database show the Total Error (TE) to be reduced by 78% when compared to a non-adaptive system.

1. INTRODUCTION

Access control systems are becoming an increasingly important part of our life. As an example, Automatic Teller Machines (ATMs) employ a simple identity verification where the user is asked to enter their Personal Identification Number (PIN), known only to the user, after inserting their ATM card. If the PIN matches the one prescribed to the card, the user is allowed access to their bank account. Similar verification systems are widely employed to restrict access to rooms and buildings.

The verification system such as the one used in the ATM only verifies the validity of the combination of a certain possession (in this case, the ATM card) and certain knowledge (the PIN). The ATM card can be lost or stolen, and the PIN can be compromised (eg. somebody looks over your shoulder while you're entering the PIN). Hence new verification methods have emerged, where the PIN has either been replaced by, or used in addition to, biometrics such as the person's speech, face image or fingerprints. The use of biometrics is attractive since they cannot be lost or forgotten and vary significantly between people.

Recently, person verification systems have evolved from using single-mode data (eg. speech) [1] to multi-modal data (eg. speech and face images) [2, 3], with the latter systems exhibiting higher performance. In current multi-modal verification systems, the separate modalities are processed by specially designed *modality experts*, where each expert gives an opinion value of the claimed identity. A high opinion indicates the person is a true claimant, while a low opinion suggests the person is an impostor. The opinions from the modality experts are used by a *decision stage* (sometimes referred to as a *fusion stage*). It considers the opinions and makes the final decision to either accept or reject the claim.

The decision stage can be a binary classifier processing n-dimensional opinion vectors. The vector is comprised of opinion values from each of the n modality experts. The classifier is trained with example opinions of known impostors and true claimants and classifies a given opinion vector as belonging to either the impostor or true claimant class.

The performance of a verification system is measured in terms of False Acceptance rate (FA%) and False Rejection rate (FR%), defined as:

$$\text{FA} = \frac{I_a}{I_t} \times 100\% \qquad \text{FR} = \frac{C_r}{C_t} \times 100\%$$

where I_a is the number of impostors classified as true claimants, I_t is the total number of impostor classification tests, C_r is the number of true claimants classified as impostors, and C_t is the total number of true claimant classification tests.

To quantify the performance into a single number, two measures are often used: Total Error, defined as TE = FA + FR, and Equal Error Rate (EER), where the system is configured to operate with FA = FR.

The performance of a multi-modal verification system can degrade rapidly when one of the experts is processing noise corrupted signals, eg. speech with background noise [4]. This occurs due to a mismatch between training and testing (verification) conditions. One way to alleviate the degradation is to adapt the experts to noisy conditions. Another way is to adapt the decision stage, which is the focus of this paper. Previous work [3] has shown that for a system using speech as one of its modalities, it is possible to alleviate performance degradation in noisy conditions by finding

optimum parameters for the binary classifier for each condition. During verification, the quality of the speech signal is measured (which can be interpreted as a measure of the mismatch between training and testing conditions) and classifier parameters are chosen that best match the given condition. While this approach is quite effective, the usefulness of the system is limited since all of the noisy conditions need to be seen a priori.

It has been shown in [2] that a Minimum Cost Bayesian Classifier (MCBC) obtained the best performance when compared to other binary classifiers. In this paper we propose to modify the MCBC to take into account the reliability of each expert, hence allowing the de-emphasis of the contribution from the expert affected by noise. We also propose a method to find the reliability of the speech expert without directly modeling the noisy speech or finding the reliability a priori for each condition.

The paper is structured as follows: The database used for experiments is described in Section 2. The speech and face experts are described in Sections 3 and 4 respectively. The modified MCBC is shown in Section 5 followed by the proposed method to find the reliability in Section 6. Experiments evaluating the performance of the adaptive system are shown in Section 7.

2. DIGIT DATABASE

The database is comprised of video and corresponding audio recordings of 37 subjects (16 female and 21 male), divided into *train*, *validation* and *test* sections. While wearing different clothes for each section, the subjects were asked to perform the following:

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

The video, recorded at 25 fps using a broadcast quality digital camera, is stored as a sequence of JPEG files with a resolution of 280 $\times$ 260. The audio data is stored in 32 kHz, 16-bit mono format. In total, the database occupies approximately 7 Gigabytes. For more information about the database please visit: *http://spl.me.gu.edu.au/digit/*

3. SPEECH MODALITY EXPERT

The speech modality expert is based on the Gaussian Mixture Model (GMM) approach [1]. The given speech signal, sampled at 16 kHz and quantized over 16 bits, is analyzed every 10 ms using a 20 ms Hamming window. For each window the energy is measured, and if it is above a set threshold (which is set to a value so as to discard silent parts), 12th order cepstral parameters are derived from Linear Prediction Coding (LPC) parameters [5]. Each set of extracted parameters can be treated as a 12-dimensional feature vector. Delta cepstral parameters are then computed using neighbouring windows [6] and appended to the feature vector, extending it to 24 dimensions.

Client models are generated by pooling training data for a given person and constructing an 8-mixture GMM using the Expectation Maximization algorithm [7]. During verification, the expert, using the GMM of the claimed identity, calculates a score y_s on the claim s using:

$$y_s = \frac{1}{N}\sum_{i=1}^{N}\log[p(\vec{x}_i|\lambda_s)] \tag{1}$$

$$\text{where} \quad p(\vec{x}|\lambda) = \sum_{m=1}^{M} p_m \mathcal{N}(\vec{x}, \vec{\mu}_m, \Sigma_m) \tag{2}$$

$$\text{and} \quad \lambda = \{p_1, ..., p_M, \vec{\mu}_1, ..., \vec{\mu}_M, \Sigma_1, ..., \Sigma_M\} \tag{3}$$

Here N is the number of feature vectors, $\vec{x}_i$ is the i-th feature vector, λ_s is the model for person s, M is the number of mixtures, p_m is the mixture weight for mixture m, and $\mathcal{N}(\vec{x}, \vec{\mu}, \Sigma)$ is a multi-variate Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix Σ. The score y_s is transformed into an opinion z_s by cohort normalization [8] as follows:

$$z_s = y_s - \frac{1}{C-1}\sum_{i=1, i\neq s}^{C} y_i \tag{4}$$

where C is the number of client models.

4. FACE MODALITY EXPERT

The face modality expert is based on the Principal Component Analysis (PCA) approach [9] combined with the GMM approach. Given a grey-scale image of a person from the Digit Database, the location of the face is found by correlation with a template of an average face. Locations of eyes and nose are found similarly and are used by an affine transformation to normalize the distance between the eyes and the distance between the eye line and the nose. Next, a 85 $\times$ 65 pixel "face" window is extracted, containing the forehead, eyes and the nose, with the locations of the eyes and nose fixed at pre-determined locations. To normalize any lighting/brightness differences between "face" windows, an offset is added to all pixels inside the window so that their median is equal to a pre-determined value.

By concatenating the rows of the "face" window, a 5525-dimensional "face" vector is constructed. Since processing vectors with such high dimensions is computationally infeasible, PCA is used to reduce the "face" vector to a 50-dimensional feature vector.

The training and verification is similar to the speech modality expert with the following differences: the client models are single mixture GMMs and the minimum of score y_s is limited to a pre-determined value to reduce the effect of outliers.

5. MODIFIED MINIMUM COST BAYESIAN CLASSIFIER

Let us define a data set D of M n-dimensional opinion vectors belonging to two classes labelled as -1 and $+1$, indicating impostor and true claimant classes respectively:

$$D = \{(\vec{x}_k, y_k) \mid k \in \{1, ..., M\}, \vec{x}_k \in \mathbb{R}^n, y_k \in \{-1, +1\}\}$$

It has been shown in [2] that a Minimum Cost Bayesian Classifier can be used to map the vectors from their data space to their label space, ie.,

$$f(\vec{x}) = \begin{cases} +1, & \text{if } \prod_{i=1}^{n} \frac{p(x_i|\lambda_{i,+1})}{p(x_i|\lambda_{i,-1})} > 1 \\ -1, & \text{otherwise} \end{cases} \quad (5)$$

where $p(x_i|\lambda_{i,k})$ is the likelihood of opinion x_i from expert i belonging to class k. By taking the log of the above likelihood ratio test, we obtain:

$$f(\vec{x}) = \begin{cases} +1, \text{ if } \sum_{i=1}^{n} \log p(x_i|\lambda_{i,+1}) - \\ \qquad \sum_{i=1}^{n} \log p(x_i|\lambda_{i,-1}) > 0 \\ -1, \text{ otherwise} \end{cases} \quad (6)$$

Assuming a 2 modality system, (6) becomes:

$$f(\vec{x}) = \begin{cases} +1, \text{ if } \frac{\log p(x_1|\lambda_{1,+1}) + \log p(x_2|\lambda_{2,+1})}{\log p(x_1|\lambda_{1,-1}) + \log p(x_2|\lambda_{2,-1})} > 1 \\ -1, \text{ otherwise} \end{cases} \quad (7)$$

When the first expert is processing noise corrupted signals, the reliability of its opinion has decreased. To de-emphasise its contribution we introduce a weighting factor, α, as follows:

$$f(\vec{x}) = \begin{cases} +1, \text{ if } \frac{\alpha \log p(x_1|\lambda_{1,+1}) + (1-\alpha) \log p(x_2|\lambda_{2,+1})}{\alpha \log p(x_1|\lambda_{1,-1}) + (1-\alpha) \log p(x_2|\lambda_{2,-1})} > 1 \\ -1, \text{ otherwise} \end{cases} \quad (8)$$

We define α as:

$$\alpha = \frac{\rho_1}{\rho_1 + \rho_2} \quad (9)$$

where $\rho_k \in [0, 1]$ is the reliability measure of expert k. In our experiments, we use the Gaussian density to model the distribution of opinions, ie.,

$$p(x_i|\lambda_{i,k}) = \mathcal{N}(x_i, \mu_{i,k}, \sigma^2_{i,k}) \quad (10)$$

where $\mu_{i,k}$ and $\sigma^2_{i,k}$ are the mean and variance of opinions from expert i for class k.

6. SPEECH EXPERT RELIABILITY

It has been observed [10] that when white noise is added to the speech signal, the magnitude of LPCC feature vectors is decreased. This causes a decrease in the variance of the feature vectors and can be exploited to detect the amount of noise present (ie., quality). This in turn can be used to determine the reliability of the speech expert. Let us model clean speech from all speakers enrolled in the verification system by a single mixture GMM, where the model parameters are described by λ_{clean} and the model is referred to as *global speech model*. Speech parameterization and model training is done in similar fashion as in Section 3. Given a speech utterance, we define its earmark, e, as:

$$e_x = \frac{1}{N} \sum_{i=1}^{N} \log[p(\vec{x}_i|\lambda_{clean})] \quad (11)$$

where x_i is the i-th feature vector and $p(\vec{x}|\lambda)$ is described in Equation (2).

When white noise is added, the decrease of variance of feature vectors causes the corresponding earmark to be higher than for clean speech. Let us model the distribution of earmarks of all clean speech utterances with a Gaussian distribution, where the model parameters are described by $\lambda_e = \{\mu_e, \sigma^2_e\}$. We define the quality, q, of a given speech utterance as:

$$q(e|\lambda_e) = \log \mathcal{N}(e, \mu_e, \sigma^2_e) \quad (12)$$

It has been shown in [4] that the reliability of the speech expert is proportional to the quality of the speech signal. To convert the quality into reliability, which is in the [0,1] interval, we define the reliability of the speech expert for a given speech utterance as:

$$\rho_1(q) = \rho_{min} + \frac{\rho_{max} - \rho_{min}}{1 + exp\{-a(q-b)\}} \quad (13)$$

where ρ_{min} and ρ_{max} are the minimum and maximum allowable reliability values respectively, while a and b determine the rate of decrease of reliability according to the rate of decrease of quality.

7. EXPERIMENTS

7.1. Speech Signal Preparation

Let us define the Noise Floor Power (NFP) of a speech signal as the average power of non-speech segments. By dividing the signal into 20 ms windows with an overlap of 10 ms, an approximation of the NFP can be found by the mean power of 25 windows with the lowest power.

A given speech signal can be modified to have a required NFP by the following means: adjust the amplitude so that the maximum amplitude is equal to a pre-determined constant, then add a sufficient amount of white Gaussian noise.

All speech signals from the Digit Database where normalized to have an NFP of 55 dB and shall be referred to as *clean*. Versions with a NFP of 56 to 70 dB were also generated, and shall be referred to as *noisy*.

7.2. Training

The speech expert was trained on clean digit sequences from the training section. The face expert was trained on all available images from the digit sequences in the training section.

Opinion distributions for impostors and true claimants were found by testing the experts on the validation section. 20 utterances from each speaker were tested against the speaker's own model, resulting in 740 true claimant tests. Each utterance from each speaker was tested against every other speaker's model, resulting in 26640 impostor tests.

The global speech model was trained using all clean digit sequences from the training section. Distribution of earmarks was found by testing all clean digit sequences from the validation section against the global speech model. ρ_{min} and ρ_{max} were emprically set to 0.1 and 0.9 respectively. The transformation parameters a and b were empirically selected so the verification system's TE, when tested on noisy data with a NFP of 70 dB from the validation section, was below the TE of the face expert. Individual performance of the face and speech experts can be found by forcing $\alpha = 0$ and $\alpha = 1$, respectively.

7.3. Performance Evaluation

The system was tested on data from the test section, with the NFP ranging from 55 to 70 dB, in 4 configurations:

1. Speech expert alone ($\alpha = 1$)
2. Face expert alone ($\alpha = 0$)
3. Non-adaptive, where reliability settings for the speech and face experts are equal, ie., $\rho_1 = \rho_2 = 0.9 \quad (\therefore \alpha = 0.5)$
4. Noise adaptive, where ρ_1 is found by Eqn. (13) and $\rho_2 = 0.9$

As it can be seen in Figure 1, the speech expert's performance rapidly degrades for noisy data. In the non-adaptive configuration, for moderate amount of noise, the TE is lower than for both the speech and face experts. Unfortunately as noise increases the TE still degrades rapidly. In the adaptive configuration, for moderate amount of noise, the performance is slightly worse than the non-adaptive counterpart. However for high amount of noise it is significantly better, with the TE being 78% lower at NFP of 70 dB.

8. CONCLUSION

We have proposed an adaptive multi-modal verification system comprised of a modified Minimum Cost Bayesian Classifier (MCBC) and a method to find the reliability of the speech expert for various noisy conditions. The modified MCBC takes into account the reliability of each modality expert, allowing the de-emphasis of the contribution of opinions from the expert affected by noise. Reliability of the speech expert is found without directly modeling the noisy speech or finding the reliability a priori for various conditions of the speech signal.

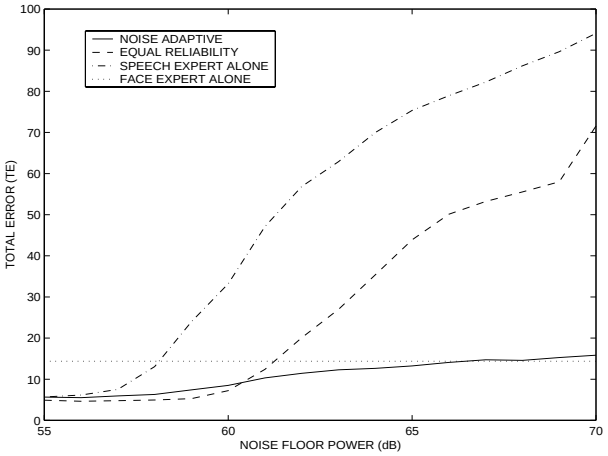

Fig. 1. Performance of the Verification System

9. REFERENCES

[1] Douglas A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", *Speech Communication 17*, 1995, pp. 91-108.

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

[3] C. Sanderson, K.K. Paliwal, "Adaptive Multi-Modal Person Verification System", *Proc. First IEEE Pacific-Rim Conference on Multimedia*, Sydney 2000, Australia.

[4] C. Sanderson, K. K. Paliwal, "Multi-Modal Person Verification System Based on Face Profiles and Speech", *Proc. Fifth Intern. Symposium on Signal Proc. Applications*, Brisbane, Australia, Aug. 1999, pp. 947-950.

[5] K. K. Paliwal, "Speech Processing Techniques", *Advances in Speech, Hearing and Language Proc.*, Vol. 1, 1990.

[6] T. Applebaum, B. Hanson, "Regression Features for Recognition of Speech in Noise", *Proc. Int. Conf. Acoustics Speech and Signal Proc.*, Toronto 1991.

[7] Todd K. Moon, "Expectation-maximization Algorithm", *IEEE Signal Processing Magazine* Vol. 13, Iss. 6, 1996.

[8] C.-S. Liu, H.-C. Wang, C.-H. Lee, "Speaker Verification Using Normalized Log-Likelihood Score", *IEEE Trans. Speech and Audio Processing*, Vol. 4, No. 1, 1996, pp. 56-60.

[9] M. Turk, A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, 1991.

[10] J.-T. Chien, H.-C. Wang, L.-M. Lee, "A Novel Projection-based Likelihood Measure for Noisy Speech Recognition", *Speech Communication 24*, 1998, pp. 298-297.