ROBUST SPEAKER VERIFICATION VIA FUSION OF SPEECH AND LIP MODALITIES

T. Wark[†]*, S. Sridharan*[‡] *and V. Chandran*

Speech Research Laboratory
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
[†]t.wark@qut.edu.au [‡]s.sridharan@qut.edu.au

ABSTRACT

This paper investigates the use of lip information, in conjunction with speech information, for robust speaker verification in the presence of background noise. It has been previously shown in our own work, and in the work of others, that features extracted from a speaker's moving lips hold speaker dependencies which are complementary with speech features. We demonstrate that the fusion of lip and speech information allows for a highly robust speaker verification system which outperforms the performance of either sub-system. We present a new technique for determining the weighting to be applied to each modality so as to optimize the performance of the fused system. Given a correct weighting, lip information is shown to be highly effective for reducing the false acceptance and false rejection error rates in the presence of background noise.

1. INTRODUCTION

Speaker verification can be thought of as person authentication using the class of information which arises from the production of speech. Within this class, the most obvious source of features is speech information itself. In ideal or clean conditions, automatic speaker recognition (ASR) systems perform very well using speech characteristics alone. However, considerable decreases in performance are observed as a result of adverse variables such as background noise, channel distortion or reverberation [1].

A less obvious source of information related to speech production is that of visual lip information. Lip movement is a natural by-product of the various positions the oral cavity must take to produce the range of phonetic sounds we understand as speech. In noisy conditions, a listener makes considerable use of lip information to aid in the speech intelligibility process. We have shown in our previous work that speaker recognition of reasonable accuracies can be obtained by using lip information only [2].

Previous work in acoustic-labial speaker verification has been performed via the use of Hidden Markov Model (HMM) classifiers using *fixed* acoustic conditions [3]. Other recent audio-visual authentication work has considered the fusion of facial and speech information, however once again the fusion systems assume fixed acoustic and visual conditions [4][5].

The work presented in this paper considers the fusion of speech and lip information given that audio conditions can differ greatly from training to testing. We develop an algorithm for the automatic determination of weights to be applied to audio and visual classifiers, so as to maximize verification performance over a range of operating conditions.

This work was supported in part by a Dept. of Defence Science & Technology Organization (DSTO) research contract.

2. SYSTEM FEATURE EXTRACTION

2.1. Audio Sub-System

The audio sub-system feature extraction is quite standard, with mel-cepstral features [6] being extracted from the speech. Mel-cepstral features have been shown in the past to be well suited for speaker identification purposes [1], hence their use in this application.

2.2. Visual Sub-system

To extract features from moving lips, a system must be able to automatically locate and track the lip contour. This is by no means a simple task and much research has gone into the topic of lip tracking in itself.

We have presented in detail [2] a new method for lip tracking using a combined chromatic-parametric approach, where the parametric lip contour model is derived directly from chromatic information. This technique provides computational advantages as no minimization procedure is required to fit the contour model to the lips.

3. AUDIO AND VISUAL SYSTEMS

3.1. Audio and Visual Classifiers

Classification of both audio and visual data was achieved via the use of the Gaussian Mixture Model (GMM). These models have been used extensively in the past for the modelling of the output probability distribution of speech features for a particular speaker [6]. The multi-modal nature of the model allows it to cater for a wide range of voice characteristics for each speaker.

Experiments also showed that the *distribution patterns* of features from a speaker's moving lips, over a period of time, held speaker dependent qualities [2]. The Gaussian mixture density for a given model λ_i is given by:

$$p(\vec{x}|\lambda_i) = \sum_{m=1}^{M} p_{im}\Gamma(\vec{x}, \mu_{im}, \Sigma_{im}) \quad (1)$$

where $\vec{x}$ is the observation vector, p_{im} is the mixture weight for mixture m, of M mixtures, for speaker i, and $\Gamma(\vec{x}, \mu, \Sigma)$ is a multivariate Gaussian function with mean μ and covariance matrix Σ.

3.2. Verification Decisions

In any verification system the aim is to determine whether to accept or reject a speaker based on how well their data fits the model of the claimed speaker. We can categorize the verification decision as a two class problem where the classes H_0 and H_1 are the acceptance and rejection classes respectively. The simplest approach is to compare the score from the model to a threshold and make a class decision as:

$$P(X^{mode}|\lambda_{claim}^{mode}) \geq \mathcal{T}_{mode} \Rightarrow H_0 \quad (2)$$

$$P(X^{mode}|\lambda_{claim}^{mode}) < \mathcal{T}_{mode} \Rightarrow H_1 \quad (3)$$

where:

$$P(X^{mode}|\lambda_{claim}^{mode}) = \frac{1}{T}\sum_{t=1}^{T_{mode}} \log p(x_t^{mode}|\lambda_{claim}^{mode}) \quad (4)$$

where λ_{claim}^{mode} is the model for the claimed speaker, T is the number of frames for input features x_t^{mode}, $\mathcal{T}_{mode}$ is the threshold value and $mode \subset [aud, vis]$.

3.3. Background Normalization

In general, superior verification performance can be obtained via the use of background normalisation or cohort speaker models [7]. Rather than only the *claimed* speaker model score being used for thresholding purposes, we also make use of background model scores. A *normalised* score is calculated as:

$$u(X^{mode}|s_{claim}) = \log p(X^{mode}|\lambda_{claim}^{mode}) - \\ - \log \sum_{b \in \mathcal{B}(i)} p(X^{mode}|\lambda_b^{mode}) \quad (5)$$

where s_{claim} is the claimed speaker and $mode \subset [aud, vis]$ as before, and $\mathcal{B}$ is the background speaker set.

To increase the robustness of each client's model to both similar and dissimilar impostors, we incorporate both *near* and *far* speakers into our background speaker cohort selection. We follow a procedure similar to [6] where we select maximally-spaced speakers from a close set, and maximally spaced speaker's from a far set, thus decreasing redundancy in the choice of background speaker characteristics.

The final normalized score u is calculated as:

$$u(X^{mode}|s_{claim}) = \log p(X^{mode}|\lambda_{claim}^{mode}) - \\ - \log \sum_{b \in \mathcal{C}(i)} p(X^{mode}|\lambda_b^{mode}) - \\ - \log \sum_{b \in \mathcal{F}(i)} p(X^{mode}|\lambda_b^{mode}) \quad (6)$$

where $\mathcal{C}$ and $\mathcal{F}$ are the close and far cohort sets for the claimed speaker respectively.

In the case of our experiments we chose close and far cohorts sets of 5 speakers each from initial groups of 10 close and 10 far speakers. Hence our final cohort set contained 10 speakers.

4. AUDIO-VISUAL FUSION SYSTEM

4.1. System Structure

Two main approaches can be taken for fusion, being that of *direct* fusion, and *output* fusion [8]. In direct fusion features from each source are combined *prior* to classification, whereas in output fusion, features from each source are separately classified, with the classifier outputs then being combined. Past research [9] has shown that *output* fusion is in general superior for audio and visual fusion.

The basic structure of our fusion system is that of *asynchronous linear output* fusion. Here the verification decision H is based upon a linear combination of outputs from the audio and visual classifiers. This can be expressed for the general case [10] as:

$$assign \quad H \rightarrow H_j \quad for\ j = 0, 1 \quad if$$

$$\alpha P(H_j|\mathbf{X}_{aud}) + (1-\alpha)P(H_j|\mathbf{X}_{vis}) = \\ = \max_{k \subset [0,1]} \{\alpha P(H_k|\mathbf{X}_{aud}) + (1-\alpha)P(H_k|\mathbf{X}_{vis})\} \quad (7)$$

where H_0 and H_1 are the *accept* and *reject* classes respectively, $\mathbf{X}_{mode}$ are the input features, $\alpha \in [0, 1]$, and we assume the *a priori* class probabilities $P(H_0)$ and $P(H_1)$ are equal.

Rather than attempting to compute the *a posteriori* probabilities $P(H_k|\mathbf{X}_{mode})$ the verification decision is based upon a speaker independent thresholding of cohort normalised scores u from each modality. This can be expressed mathematically as:

$$assign \quad H \rightarrow H_0 \quad if$$

$$\alpha.u(\mathbf{X}_{aud}|s_{claim}) + (1-\alpha)u(\mathbf{X}_{vis}|s_{claim}) \geq \mathcal{T} \quad (8)$$

$$assign \quad H \rightarrow H_1 \quad if$$

$$\alpha.u(\mathbf{X}_{aud}|s_{claim}) + (1-\alpha)u(\mathbf{X}_{vis}|s_{claim}) < \mathcal{T} \quad (9)$$

where $\mathcal{T}$ is the score threshold value, and $u(\mathbf{X}_{mode}|s_i)$ are defined in Equation 6.

Thus we first calculate the cohort normalised scores for each modality, and then combine these scores via a linear weighting before thresholding the final value.

4.2. Determination of Optimal Classifier Weightings

In any classification system, the output probability is really an estimate of the true *a posteriori* probability with an associated error factor. Hence we can express the output estimate $\hat{P}(H_k|\mathbf{X}_i)$ as:

$$\hat{P}(H_k|\mathbf{X}_{mode}) = P(H_k|\mathbf{X}_{mode}) + \epsilon_{mode} \quad (10)$$

where $mode \subset [aud, vis]$ and $k \subset [0, 1]$.

We seek to find a way to automatically allocate the optimum weighting $\alpha \in [0, 1]$ to classifiers so as to minimise the error contributions ϵ_{mode}, to the overall verification problem. To determine the resulting confidences for each classifier, we treat the problem as a large-sample test of the hypothesis for the difference between two sample means. In our case, the two sample means μ_0 and μ_1

represent the means of the normalised scores u given true clients and given true impostors respectively.

Hence we are testing the hypothesis:

$$H_0 : \frac{1}{m}\sum_{i=1}^{m} u(\mathbf{X}_{mode}|client_i) - \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{X}_{mode}|impos_i) \geq 0 \tag{11}$$

It can be shown statistically, that the *standard error* ξ for this estimate is:

$$\xi_{mode} = \sigma^{mode}_{\bar{X}_0 - \bar{X}_1} = \sqrt{\frac{\sigma_0^2}{m} + \frac{\sigma_1^2}{n}} \tag{12}$$

where m and n are the number of client and imposter tests respectively, and σ_0^2 and σ_1^2 are the sample class variances determined from the training set.

We assume that the standard error for a classifier gives a relative indication of the ability of the classifier to consistently separate client scores and imposter scores. The less variation there is in client and imposter scores, the lower the standard error for that classifier will be, and the better the verification performance.

Based on this we determine an "optimal" value of α as:

$$assign \quad \alpha^{optim}_{mode} \rightarrow \alpha_{mode} \quad if \quad \alpha_{mode} \propto \frac{1}{\xi_{mode}} \tag{13}$$

where $mode \subset [aud, vis]$.

Hence based on the assignment of α in Equation 7, we determine α as:

$$\alpha = \frac{\xi_{vis}}{\xi_{aud} + \xi_{vis}} \tag{14}$$

5. EXPERIMENTS

5.1. Experiment Details

We trained and tested the audio and visual verification systems using the M2VTS multi-modal database [11]. The database consists of over 27000 colour images of 37 subjects counting from *zero* to *neuf* in French over a number of different sessions, with a week between each session. We used the first three recording sessions as training data, and the fourth session as test data.

The verification tests consisted of a series of both *false rejection* (FR) tests and *false acceptance* (FA) tests. The first 30 speakers were chosen to be clients, whilst the remaining 7 speakers were used as impostors only. Cohort speakers for each of the client speakers were obtained from the other remaining client speakers. For FR tests, all 30 speakers were used as clients to their own models resulting in 30 tests. For FA tests each of the 7 impostors were used against all 30 client models resulting in 210 tests.

One of the key aims of the experiments was to evaluate the effectiveness of the choice of α as speech data quality was degraded with noise. Given that the technique for choosing α_{opt}, described in Section 4.2 is optimised for clean audio and visual data, we deliberately change conditions to extreme levels to evaluate system robustness.

5.2. Results

5.2.1. Audio Tests

In Figure 1 the *receiver operating characteristic* (ROC) curves are presented for the verification system using speech data alone. Each ROC curve represents verification performance under a particular level of audio degradation.

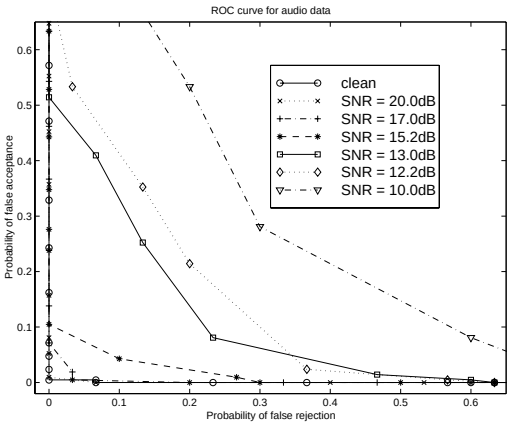

Figure 1: Audio ROC curves

5.2.2. Lip Tests

Figure 2 shows the ROC curve for verification using lip information only. For the purposes of these tests, the quality of visual information has been held constant and not degraded in any way.

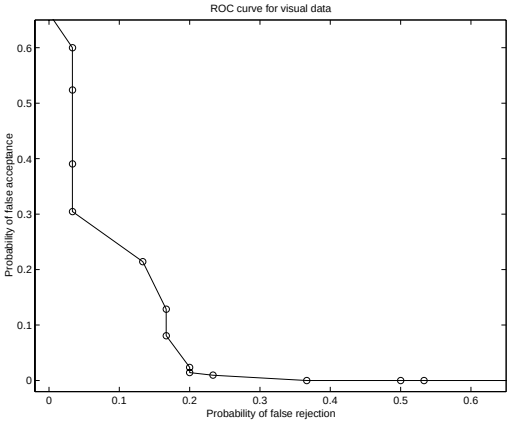

Figure 2: Visual ROC curves

5.2.3. Audio-Visual Fusion Tests

The verification results after fusion of speech and lip information are presented in Figure 3. The value of α used to form the results is determined as per Section 4.2. Given clean audio and visual training data, α_{opt} was calculated to be 0.901.

To evaluate how good the choice of α_{opt} is, Figure 4 gives a comparison of the EER's for the "optimal" system with a range of other values of $\alpha \in [0, 1]$.

For clean data, the optimal fused system can be seen to maintain the speech only EER rate of 0.47%. At very high noise levels,

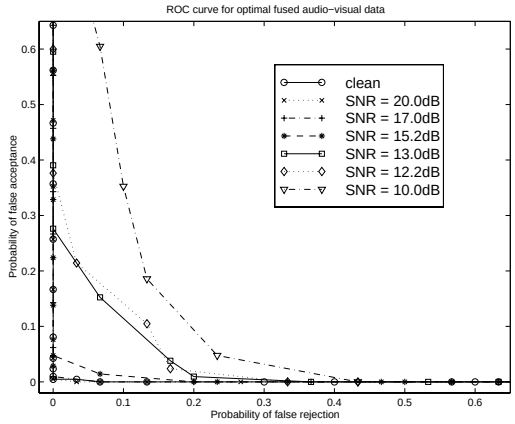

Figure 3: Audio-visual ROC curves

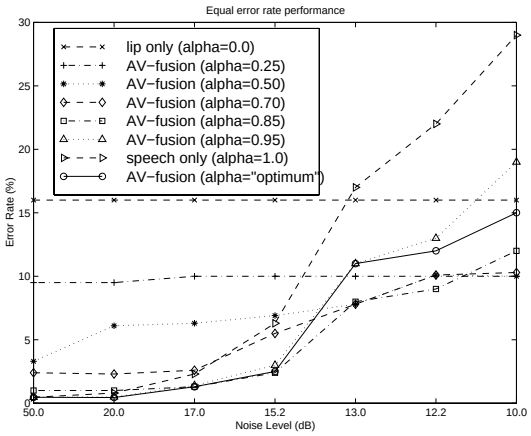

Figure 4: Comparison of EER's for varying weights

the optimal fused system reduces the EER from 29.0%, for speech only, to 15.0%.

It can be observed that values of α such as 0.70 and 0.85 outperform the optimal system at high noise levels, however for clean data the corresponding EER's for these values of α are 2.4% and 1.0%, which is a step backwards from the excellent performance using speech only.

Given that a verification system would be ideally operating in clean or low noise conditions, the choice of $\alpha_{opt} = 0.901$ made by the system does indeed appear to be almost optimal. If a system were to be continually operating in high noise conditions, we would need to determine the standard error for audio data ξ_{aud} based on highly noisy training data and find the new α_{opt} accordingly.

6. CONCLUSIONS

This paper has presented the use of lip information as a secondary source of information for robust speaker verification under varying noise conditions. We have previously shown that speaker dependent lip information can be obtained by classifying the distribution pattern of features from a speaker's moving lips over time.

Results show that speaker verification performance using speech information only, decreases considerably as background noise increases. The fusion of lip and speech information allows the system performance to remain relatively high even when speech information is highly degraded.

We present a technique for automatically determining the weighting of audio and visual classifiers to maximise overall verification performance over a range of operating characteristics. Results from experiments are encouraging and show that the technique is able to select a value of α to match the excellent performance, in clean conditions, of speech-only verification, whilst greatly improving results over speech-only in high noise.

7. ACKNOWLEDGMENT

This work was carried out in support of the European Commission ACTS Project M2VTS.

8. REFERENCES

[1] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition - a feature based approach," *IEEE Signal Processing Magazine*, pp. 58–71, Sept. 1996.

[2] T. J. Wark and S. Sridharan, "A syntactic approach to automatic lip feature extraction for speaker identification," in *Int. Conf on Acoustics Speech and Signal Processing*, vol. 6, pp. 3693–3696, May 1998.

[3] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, vol. 18, pp. 853–858, 1997.

[4] B. Duc, "Fusion of audio and video information for multi modal person authentication," *Pattern Recognition Letters*, vol. 18, pp. 835–843, 1997.

[5] U. Dieckmann, P. Plankensteiner, and T. Wagner, "Sesam: A biometric person identification system using sensor fusion," *Pattern Recognition Letters*, vol. 18, pp. 827–833, 1997.

[6] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, pp. 91–108, 1995.

[7] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proceedings of Eurospeech97*, pp. 963–966, 1997.

[8] P. Varshney, "Multisensor data fusion," *Electronics and Communication Engineering Journal*, pp. 245–253, Dec. 1997.

[9] M. Alissali, P. Deleglise, and A. Rogozan, "Asynchronous integration of visual information in an automatic speech recognition system," in *Int. Conf. on Spoken Language Processing*, 1996.

[10] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.

[11] S. Pigeon, "The m2vts database," technical report, Laboratoire de Telecommunications et Teledetection, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium, (http://www.tele.ucl.ac.be/M2VTS), 1996.