A COMPARISON OF STATIC AND DYNAMIC CLASSIFIER PERFORMANCE FOR MULTI-MODAL SPEAKER VERIFICATION

Timothy Wark Sridha Sridharan Vinod Chandran
Speech Research Laboratory, RCSAVT
Research Concentration in Speech, Audio and Video Technology
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
{t.wark,s.sridharan,v.chandran}@qut.edu.au

ABSTRACT: This paper compares the performance of two techniques for the fusion of speech and lip information for robust multi-modal speaker verification. The first approach uses static speech and lip information via GMM classifiers in order to make a speaker verification decision, whilst the second approach uses dynamic information via the use of HMM classifiers. Verification experiments are performed on the M2VTS database which show that the dynamic system significantly outperforms the static system over a range of operating conditions.

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 (Mammone et al., 1996).

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 (Wark et al., 1998).

Previous work in acoustic-labial speaker verification has been performed via the use of Hidden Markov Model (HMM) classifiers using *fixed* acoustic conditions (Jourlin et al., 1997). 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 (Duc, 1997; Dieckmann et al., 1997).

The work presented in this paper compares the performance of two previous multi-modal approaches we have developed. Both approaches use output fusion of speech and lip information in order to make a verification decision however one uses static information, and the other uses dynamic information. Speaker verification experiments are performed using the M2VTS multi-modal database (Pigeon, 1996) and the results are compared over a range of operating conditions.

SYSTEM FEATURE EXTRACTION

Audio Sub-System

The audio sub-system feature extraction is standard, with mel-cepstral features (Reynolds, 1995) being extracted from the speech. Silence is first removed from the speech via a low-energy thresholding. This is followed by the calculation of the magnitude spectrum of 32ms speech segments. The spectrum is then pre-emphasised and processed by a mel-scale filterbank. Finally the filterbank coefficients are cosine transformed to produce the cepstral coefficients.

Visual Sub-system

We have presented in detail (Wark et al., 1998) 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, and no prior manual labeling is required to obtain lip information.

The original algorithm has since been extended to allow the technique to perform more robustly under varying skin and lighting conditions, as well as performing inner lip tracking. A sample of tracking results over two speakers are shown in Figure 1.

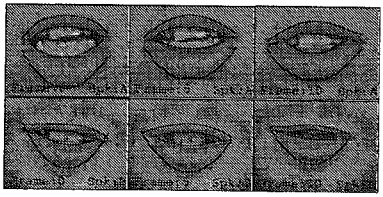

Figure 1: Examples of tracking performance over two speakers with outer/inner lip parametric model.

Raw features are extracted via colour profiles taken around the lip contour. As the contour model follows the moving lips, the chromatic features will be consistent with respect to the lip position. Features are then reduced via the use of Principal Component Analysis (PCA), followed by Linear Discriminant Analysis (LDA). A complete description of these feature reduction steps may be found in Wark et al. (1998).

STATIC MULTI-MODAL SYSTEM

Audio and Visual Classifiers

Classification of both static 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 (Reynolds, 1995). The multi-modal nature of the model allows it to cater for a wide range of voice characteristics for each speaker.

The Gaussian mixture density for a given model λ_i is given by:

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

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.

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:

$$P(X^{\xi}|\lambda^{\xi}_{claim}) \geq \mathcal{T}_{\xi} \Rightarrow H_0 \tag{2}$$

$$P(X^{\xi}|\lambda^{\xi}_{claim}) < \mathcal{T}_{\xi} \Rightarrow H_1 \tag{3}$$

where:

$$P(X^{\xi}|\lambda^{\xi}_{claim}) = \frac{1}{T}\sum_{t=1}^{T_{\xi}} \log p(x^{\xi}_t|\lambda^{\xi}_{claim}) \tag{4}$$

where λ^{ξ}_{claim} is the model for the claimed speaker, T is the number of frames for input features x^{ξ}_t, $\mathcal{T}_{\xi}$ is the threshold value and $\xi \subset [aud, vis]$.

Background Normalization

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 Reynolds (1995) 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:

$$u(X^{\xi}|s_{claim}) = \log p(X^{\xi}|\lambda^{\xi}_{claim}) - \log \sum_{b\in\mathcal{C}(i)} p(X^{\xi}|\lambda^{\xi}_{b}) - \log \sum_{b\in\mathcal{F}(i)} p(X^{\xi}|\lambda^{\xi}_{b}) \tag{5}$$

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.

DYNAMIC MULTI-MODAL SYSTEM

System structure

The aim of the dynamic approach is to develop a multi-modal system that can model the *temporal* speaker-dependencies within a speaker's lip movements and combine this with the associated temporal speech dependencies. This section presents the structure of the system which enabled speakers to be verified based upon incoming temporal audio and visual information.

For this system design, we assume a *text-dependent* enrollment scenario in which the sequence of words spoken during testing is known and fixed for each client. The modeling of speaker-dependent temporal information was obtained through the building of speaker-dependent word models, in the form of either mutli-stream HMM's or single-stream HMM's. As the testing scheme was text-dependent, each speaker Λ_i could be represented by a set of word models:

$$\Lambda_i = \{w_{i1}, w_{i2}, \ldots, w_{iN_i}\} \tag{6}$$

where w_{ij} is the j^{th} word model for speaker i, and N_i is the number of word models in the test-set for speaker i.

To illustrate more clearly the discriminative speaker information contained within each speaker's word models, a contour-plot representation of the state-likelihood distributions $b_{js}(t)$ is shown in Figure 2. The plot shows audio and video state-likelihood distributions for speaker-dependent two-state HMM's for the word "zero", over two speakers.

Verification Decisions

The verification decision is based on an output fusion of background-normalised *a posterior* speaker likelihoods, as described in Equation 5 for the static system.

The total output speaker log-likelihood $\log P(o|\Lambda_i)$ is obtained by evaluating the optimal state-sequence moving through all speaker's word models using the standard Viterbi algorithm. Given the speaker model

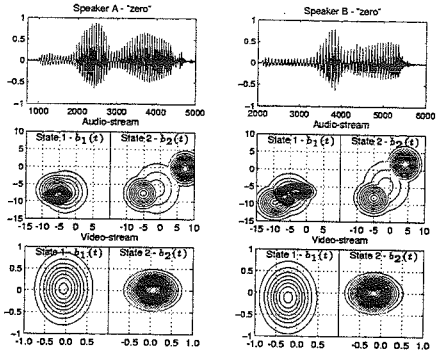

Figure 2: State distributions across an example word "zero".

$\hat{P}_i = \{w_{i1}, w_{i2}, \ldots, w_{iN_i}\}$, the total speaker output log-likelihood is calculated as:

$$\log P(\boldsymbol{o}|\Lambda_i) = \log P(\Lambda_i|\boldsymbol{o}) = \log \sum_{j=1}^{N_i} P(w_{ij}|\boldsymbol{o}) \tag{7}$$

where w_{ij} is the j^{th} word model for speaker i's test utterance, and $P(w|\boldsymbol{o})$ are the word likelihoods returned by the Viterbi algorithm.

OUTPUT FUSION OF SPEECH AND LIP INFORMATION

A linear output combination approach has been taken for the fusion of audio and visual *a posterior* likelihoods. In order to assign an *a prior* confidence level to each modality, we treat the problem as a large-sample test of the hypothesis for the difference between two sample means, where the sample means are allocated as the means of prior known client and impostor output scores, μ_c and μ_i respectively. Thus are testing the hypothesis:

$$H_0 : \frac{1}{N_c}\sum_{i=1}^{N_c} u(\mathbf{x}_\xi|client_i) - \frac{1}{N_i}\sum_{i=1}^{N_i} u(\mathbf{x}_\xi|impos_i) \geq 0 \tag{8}$$

where it can be shown statistically [ref], that the *standard error* ζ for this estimate is:

$$\zeta_\xi = \sigma^\xi_{\bar{X}_c - \bar{X}_i} = \sqrt{\frac{\sigma_c^2}{N_c} + \frac{\sigma_i^2}{N_i}} \tag{9}$$

where N_c and N_i are the number of known client and impostor sample tests available, σ_c^2 and σ_i^2 are the sample class variances for clients and impostors determined from a labeled evaluation set, and $\xi \subset \{aud, vis\}$.

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 impostor scores. The less variation there is in client and impostor scores, the lower the standard error for that classifier will be, and the better the verification performance. The distributions of client and impostor scores for an evaluation set is shown in Figure 3. The scores are extracted from clean audio data where it can be seen that there is good separation between impostor and client scores. The high peak around the impostor scores is due to the greater number of false acceptance tests performed.

Thus the final output scores $\hat{u}(\boldsymbol{x}|s_{claim})$ are determined as:

$$\hat{u}(\boldsymbol{x}|s_{claim}) = \left(\frac{\zeta_{vis}}{\zeta_{aud}+\zeta_{vis}}\right) u(\boldsymbol{x}_{aud}|s_{claim}) + \left(\frac{\zeta_{aud}}{\zeta_{aud}+\zeta_{vis}}\right) u(\boldsymbol{x}_{vis}|s_{claim}) \tag{10}$$

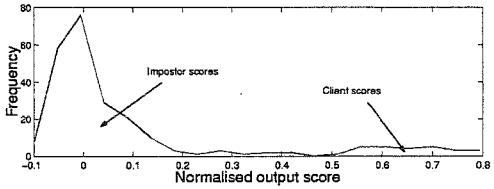

Figure 3: Distribution of known impostor and client scores from evaluation set.

EXPERIMENTS

The verification experiments consisted of a series of both *false rejection* (FR) tests and *false acceptance* (FA) tests on the M2VTS multi-modal database (Pigeon, 1996). 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.

In general, verification performance can be shown graphically by the use of a DET curve. The DET curves present plots of the miss probability against the false alarm probabilities of the verification system under varying operating or threshold points. A "miss" occurs when an impostor is accepted as a client and a "false alarm" occurs where a true client is rejected as an impostor. An example DET is shown for the dynamic "speech-only" performance in Figure 4.

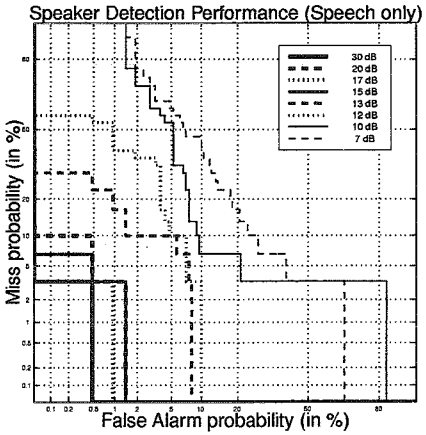

Figure 4: DET curves for speaker verification performance under varying audio noise, with speech-only information.

A more useful comparison for verification systems can be obtained by obtaining the *equal error rates* (EER) from the DET curves for each system. The EER occurs at the point on each DET curve where the false alarm and miss probabilities are equal.

A comparison of verification performances for the static and dynamic systems is given in Figure 5 for the auditory, visual and fusion static and dynamic systems. Results are plotted in terms of the verification EER corresponding to each value of the auditory noise level. It can be seen that the dynamic system produces a reduced EER over all modalities, over all values of noise. This clearly shows that the use of

temporal audio and video information results in a superior verification system, which is robust over a a range of noise levels.

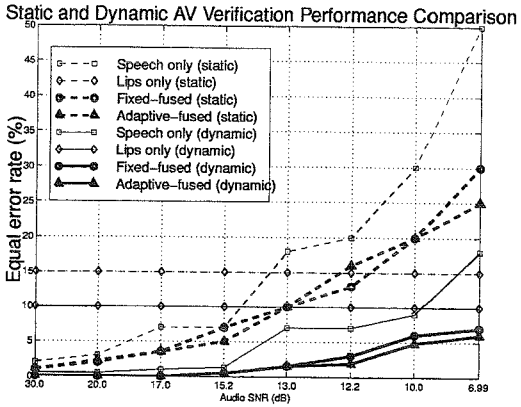

Figure 5: Comparison of verification performance with via static and dynamic audio and visual information.

CONCLUSIONS

This paper has presented a comparison of two multi-modal speaker verification systems, one using static speech and lip information and the other using dynamic information. Verification experiments are performed over the M2VTS database under varying audio conditions where it is shown that the equal error rate (EER) of the dynamic system is significantly less than the EER for the static system under the corresponding conditions.

These results demonstrate that there is important *temporal* information contained within both the visual lip information and auditory speech information that allows speakers to be discriminated in a far more effective manner than when just the static qualities of each modality are extracted. Future work would seek to extend the system to a text-independent scenario, where no fixed vocabulary is required for each client.

REFERENCES

Dieckmann, U., Plankensteiner, P. and Wagner, T. (1997), Sesam: A biometric person identification system using sensor fusion, *Pattern Recognition Letters* **18**, 827–833.

Duc, B. (1997), Fusion of audio and video information for multi modal person authentication, *Pattern Recognition Letters* **18**, 835–843.

Jourlin, P., Luettin, J., Genoud, D. and Wassner, H. (1997), Acoustic-labial speaker verification, *Audio and Video-Based Biometric Person Authentication*, number 1206 in *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, pp. 319–326. ISBN 3-540-62660-3.

Mammone, R., Zhang, X. and Ramachandran, R. (1996), Robust speaker recognition - a feature based approach, *IEEE Signal Processing Magazine* pp. 58–71.

Pigeon, S. (1996), The M2VTS database, *Proc., IEEE Aerospace Conf.*, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium. (http://www.tele.ucl.ac.be/M2VTS).

Reynolds, D. (1995), Speaker identification and verification using Gaussian mixture speaker models, *Speech Communication* pp. 91–108.

Wark, T. J., Sridharan, S. and Chandran, V. (1998), An approach to statistical lip modelling for speaker identification via chromatic feature extraction, *Int. Conf. on Pattern Recognition*, Vol. 1, pp. 123–125.