IMPROVING SPEAKER IDENTIFICATION PERFORMANCE IN REVERBERANT CONDITIONS USING LIP INFORMATION

T. Wark[†] *and S. Sridharan*[‡]

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 considers the improvment of speaker identification performance in reverberant conditions using additional lip information. Automatic speaker identification (ASI) using speech characteristics alone can be highly successful, however problems occur with mis-matches between training and testing conditions. In particular, we find that ASI performance drops dramatically when given anechoic training but reverberant test speech. Previous work [1][2] has shown that speaker dependant information can be extracted from the static and dynamic qualities of moving lips. Given that lip information is uneffected by reverberation, we choose to fuse this additional information with speech data. We propose a new method for estimating confidence levels to allow adaptive fusion of the audio and visual data. Identification results are presented for increasing levels of artificially reverberated data, where lip information is shown to provide excellent ASI peformance improvement.

1. INTRODUCTION

Room reverberation of speech occurs to some extent in almost any enclosed area. As sound is reflected off walls back to the source, the resulting speech spectrum is smeared, reducing both speech intelligibility and speaker dependent qualities. We can mathematically express a reverberated signal $r(n)$ as the convolution of the original signal $s(n)$ with the room impulse response $h(n)$:

$$r(n) = s(n) * h(n) \tag{1}$$

The effects of speech reverberation on Automatic Speaker Recognition (ASR) has not been studied extensively in the past, however work which has been done demonstrate a considerable drop in recognition performance. The case of Automatic Speaker Verification (ASV) under varying reverberant conditions has been considered [3], and it is shown that ASV peformance degrades sharply as reverberation time is increased and/or the enclosure size is decreased.

Other researchers [4] have considered the use of acoustic array processing and spectral normalisation to develop a more robust ASR system in reverberant conditions. Some performance improvement can come about as a result of these steps.

In this paper we propose the use of lip information as an additional source of information to fuse with reverberated speech features for robust ASI. As visual lip information is uneffected by reverberant conditons, we are interested in particular to evaluate the extent to which lip information can improve identification performance as reverberation of speech increases.

For experiments we use the M2VTS multi-modal database [5] consisting of 37 subjects counting from *zero* to *neuf* in French over five different recording sessions.

2. 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, hence their use in this application.

2.2. Visual Sub-system

We have presented in detail [1] a new method for lip tracking using a combined chromatic-parametric approach, however, unlike past approaches [7][8], the parametric lip contour model is derived *directly* from the chromatic information, with no minimization procedure required to fit the model to the lips. Features are extracted via colour profiles taken around the lip contour. As the contour model follows the moving lips, the chromatic features are consistent with respect to the lip position. This is illustrated in Figure 1.

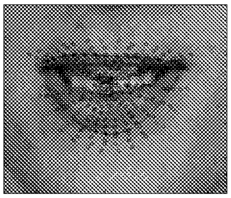

Figure 1: Colour Profile Vectors

Features are reduced via the use of Principal Component Analysis (PCA), followed by Linear Discriminant Analysis (LDA). In this way, lip features are chosen which provide the greatest discrimination between speakers. We describe these feature reduction steps in more detail elsewhere [1].

This work was supported by a CSIRO research contract.

3. PRIMARY 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 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, as well as the actual static features themselves. Based on this, we chose also to use the multi-modal nature of the GMM to allow it to model the the wide variation in features from a speakers moving mouth.

The decision rule for identifying a speaker, based on Bayes' rule [6], is defined as:

$$\hat{s} = \arg \max_{1<s<S} \sum_{t=1}^{T} log\ p(x_t|\lambda_s) \tag{2}$$

where x_t is an input feature vector at time t, and λ_s is the audio or visual model for speaker s of S total speakers.

4. AUDIO-VISUAL FUSION SYSTEM

4.1. System Structure

The aim of any fusion system is to combine information from various sources so that, in the case of identification, the resulting performance is greater than or equal to the performance of the best individual source. Anything less than this is termed as *catastrophic fusion* [9], and is of course undesirable for a speaker identification problem.

Two main approaches can be taken for fusion, being that of *direct* fusion, and *output* fusion [10]. In direct fusion features from each source are combined *prior* to classification, where-as in output fusion, features from each source are separately classified, with the classifier outputs then being combined. Past research [11] 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 identification decision is based upon a linear combination of outputs from the audio and visual classifiers after the speaker has spoken for a short period of time. This can be expressed mathematically as:

$$P(s|\vec{x}^a, \vec{x}^v) = \alpha P(\lambda_s^a|\vec{x}^a) + (1-\alpha) P(\lambda_s^v|\vec{x}^v) \tag{3}$$

where $P(s|\vec{x}^a, \vec{x}^v)$ is the probability of speaker s having generated the audio and visual feature vectors $\vec{x}^a$ and $\vec{x}^v$, given the audio and visual primary classifiers λ_s^a and λ_s^v and a weighting factor of $\alpha \in [0, 1]$.

Given this fusion structure, the main challenge is to determine an appropriate weighting to assign to the audio and visual classifier outputs. As the level of speech degradation increases due to increasing noise, we would wish to place more and more emphasis on visual information. Hence we need some way to assign a measure of confidence to the incoming audio and visual data, drawn from the data itself. The following sections present three methods for allocating weights to audio and visual data.

4.2. Equal Prior Weights

For an artificial test set, where the identity of the target speaker is known, the optimum value of α can be empirically determined by varying the level of α, according to Equation 3, between 0 and 1. This however is inconsequential for a real life speaker identification application where the identity of the speaker is of course not known.

Without making any prior assumptions about the quality of each data source, a reasonable compromise is to set $\alpha = 0.5$. In other words we weight the contribution of both audio and visual data equally for the identification problem. The results for using this technique are presented in Section 5.

4.3. Dispersion Confidence Measure

The technique presented in Section 4.2 is not capable of adapting to the surrounding environment in that the system weighting is fixed regardless of the quality of either data source. The technique presented in this section is capable of adapting the weighting factor based upon the quality of data at the time of testing.

The system achieves this by considering the dispersion of scores, or average output log-likelihoods, from the audio and visual primary classifiers. In general, we would expect that for the case of high-quality information, the GMM score assigned to the correct speaker model would be significantly higher than the scores assigned to the other speaker models.

Based on this, the confidence measure we used was taken as the *difference* in the top two speaker models scores, normalised by the *mean* of all speaker model scores u, for the audio and visual data respectively. This can be expressed for S speakers as:

$$u_{best} = \arg \max_{1 \le i \le S} \log P(\lambda_s|\vec{x}) \tag{4}$$

$$u_{nextbest} = \arg \max_{1 \le i \le S} \log P(\lambda_s|\vec{x}),\ i \neq i \mapsto u_{best} \tag{5}$$

$$u_{mean} = \frac{1}{S} \sum_{i=1}^{S} \log P(\lambda_i|\vec{x}) \tag{6}$$

$$\kappa = \frac{|u_{best} - u_{nextbest}|}{|u_{mean}|} \tag{7}$$

where κ is the confidence measure which is evaluated over both audio and visual classifier outputs, notated as κ_{aud} and κ_{vis}.

We then evaluate the weighting factor $\alpha \in [0, 1]$ as:

$$\alpha = \frac{\kappa_{aud}}{\kappa_{aud} + \kappa_{vis}} \tag{8}$$

The results from this technique are also presented in Section 5.

4.4. Secondary Classifiers

The dispersion measure is in itself a reasonable confidence measure for audio-visual fusion, however the technique breaks down when an incorrectly classified score stands out well above the other classifier scores. In this situation the confidence measure will assign a high level of confidence to the particular data while the classifier outputs are quite incorrect. We can improve on this by designing a system which can provide an indication of the *global* accuracy of the best classifier scores when comparing between the audio and visual sub-systems.

4.4.1. Mathematical Description

To achieve this we propose a *secondary* classifier stage, where the *probability distributions* of the scores of the primary classifiers are modelled themselves by uni-variate Gaussian models. In other words, we postulate that the distribution of scores u from a speaker's Gaussian Mixture Model can be adequately modelled as:

$$\varphi_i(u) = \frac{1}{(2\pi)^{1/2}|\Sigma_i|^{1/2}} exp\left[-\frac{1}{2}(u-\mu_i)'\Sigma_i^{-1}(u-\mu_i)\right] \tag{9}$$

where:

- $i \in [1, S]$ for S speakers
- u is the output log-likelihoods from audio or visual primary GMM λ_i^{aud} and λ_i^{vis}, where $u = u_t$ for $t \in [1, 2, \ldots, N_{frames}]$
- μ and Σ are the mean vector and covariance matrix respectively for the output log-likelihoods from the audio and visual primary models λ_i^{aud} and λ_i^{vis}.

The distribution of the scores from a speaker's model is called the output probability distribution (OPD). The motivation for modelling OPD's with univariate Gaussian models is evident from Figure 2. The diagrams show the distribution of output probabilities for an audio primary classifier, tested on four different occasions. For the first three occasions, the classifier is tested with its own training data from three different sessions, whilst on the fourth session, the same classifier is tested with noisy test data.

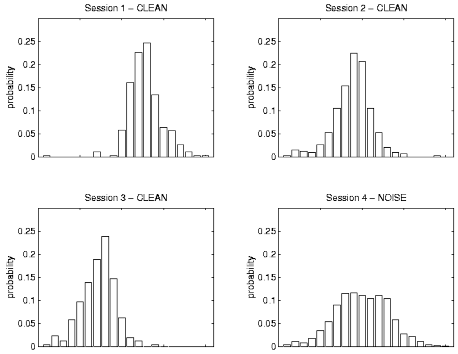

Figure 2: Output Probability Distributions (OPD)

By modelling the OPD's for high quality data for each speaker model, we have a basis to indicate the quality of incoming data to each speaker model.

The absolute values of the log-likelihoods from the primary classifiers are independent of the mean of the log-likelihoods of all classifiers. Hence before training the secondary classifiers on the primary OPD's, we *normalise* these values by dividing by the global mean of the output log-likelihoods from all speaker classifiers. We define the global mean μ_{global} for audio and visual data as:

$$\mu_{global}^{mode} = \frac{\sum_{i=1}^{S}\sum_{j=1}^{N_{frame}^{mode}} u_{ij}^{mode}}{N_{frame}^{mode} S} \tag{10}$$

where:

- $mode \subset [aud, vis]$
- S is the number of primary speaker models
- N_{frame}^{mode} is the number of frames for each speaker for either audio or video data, used to evaluate the OPD
- u_i^{aud} and u_i^{vis} are the frame output log-likelihoods from audio and visual primary classifiers for the i^{th} speaker respectively, where $u_i^{mode} \equiv u_{ij}^{mode}$ for $j \in [1, 2, \ldots, N_{frame}^{mode}]$

We then normalise the primary output log-likelihoods as:

$$y_i^{aud} = u_i^{aud} - \mu_{global}^{aud} \tag{11}$$

$$y_i^{vis} = u_i^{vis} - \mu_{global}^{vis} \tag{12}$$

for $i \in [1, 2, \ldots, S]$, where S is the number of speakers.

It is important to note that the secondary models are trained on the output log-likelihoods from the *correct* speaker's primary model only. In other words, the secondary model φ_i is trained with the log-likelihoods resulting from the primary audio and visual GMM's λ_i^{aud} and λ_i^{vis} when they are self tested with training data from speaker i. Thus at test time, we are *only* interested in evaluating the output log-likelihoods from the primary audio and visual maximally likely GMM's $\lambda_{\hat{s}}^{aud}$ and $\lambda_{\hat{s}}^{vis}$, where $\hat{s}$ is calculated as:

$$\hat{s}_{mode} = arg \max_{1<i<S} \sum_{t=1}^{T_{mode}} log\, p(x_t^{mode}|\lambda_i^{mode}) \tag{13}$$

where S is the number of speakers and T_{mode} is the number of frames of data available for testing for either the audio or visual frames of information.

In other words, we only evaluate the audio and visual secondary models φ_i^{aud} and φ_i^{vis} corresponding to the *best* audio and visual scores from the primary classifiers. Thus we determine the final audio and visual confidence measures as:

$$\nu_{aud} = P(\varphi_{s_{best}}^{aud}|y_{s_{best}}^{aud}) \tag{14}$$

$$\nu_{vis} = P(\varphi_{s_{best}}^{vis}|y_{s_{best}}^{vis}) \tag{15}$$

where $\varphi_{s_{best}}^{aud}$ and $\varphi_{s_{best}}^{vis}$ are the secondary Gaussian models corresponding to the best scores from the primary audio and visual GMM's, and $y_{s_{best}}^{aud}$ and $y_{s_{best}}^{vis}$ are the respective *normalised* primary output log-likelihoods.

The audio and visual confidence scores are finally normalised to add to one. Hence for the definition of α in Equation 3 we calculate $\alpha \in [0, 1]$ as:

$$\alpha = \frac{\nu_{aud}}{\nu_{aud} + \nu_{vis}} \tag{16}$$

5. EXPERIMENTS

We trained and tested the audio and visual identification systems using the M2VTS multi-modal database [5]. The database consists of over 27000 colour images of 37 subjects counting from *zero* to *neuf*, in French, on five different occasions. We used the first three recording sessions as training data, and the fourth session as test data.

5.1. Audio Sub-system

The speech data was artificially reverberated using an image method [3], where the level of reverberation was increased by increasing the reflection coefficients of the simulated room. One of the aims of the testing process was to evaluate the improvement from training with clean and testing with reverberated speech, to training with clean and reverberated speech. To evaluate the latter case, we reverberated the training speech in a room which was different to any of the conditions the test speech was subjected to, this being a realisable step for a real-life problem.

We found that the addition of reverberation to the training data, dramatically improved results for testing with reverberated speech data, however the identification accuracy was still low for high reverberation time. The results of this are shown in Figure 3, which confirm results presented in [3].

5.2. Fused System

The speaker identification results for increasing speech reverberation are shown in Figure 3.

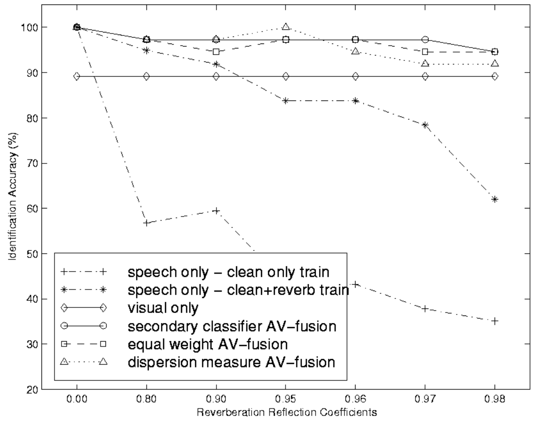

Figure 3: Speaker Identification Results

It can be seen that the fusion system is able to select linear weighting values α, for each speaker's test data so as to keep the resulting fused performance well above the best performance of either the audio or visual information. Hence we are avoiding catastrophic fusion and providing a satisfactory fusion system for speaker identification.

We observe that, with the exception of one point, the secondary classification fusion system either equals or outperforms the fusion performance of both the fixed weight and dispersion measure systems.

6. CONCLUSIONS

We have considered the performance improvement of speaker identification in reverberent conditions using lip information. Techniques for estimating confidence measures for incoming audio and visual information are proposed to allow adaptive fusion of audio and visual primary classifer outputs.

Methods proposed include confidence measures based upon a measure of score separation within primary classifiers, as well as secondary GMM classifiers trained upon output score distributions for high quality data. The use of secondary classifiers enables a "knowledge" to be built into the system as to the quality of incoming audio or visual data.

The results of experiments are encouraging and show the importance of lip information for speaker identification when speech is highly degraded due to reverberation. Whilst the amount of data per speaker is limited, results are promising for future larger-scale work in the multi-modal area.

7. ACKNOWLEDGEMENT

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

8. REFERENCES

[1] 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 Proccessing*, vol. 6, pp. 3693–3696, 1998.

[2] J. Luettin, N. Thacker, and S. Beet, "Locating and tracking facial speech features," in *Proc. Int. Conf on Pattern Recognition*, vol. I, pp. 652–656, 1996.

[3] P. J. Castellano, S. Sridharan, and D. Cole, "Speaker recognition in reverberent enclosures," in *Int. Conf. on Acoustics Speech and Signal Processing*, 1996.

[4] J. Gonzalez-Rodriguez and J. Ortega-Garcia, "Robust speaker recognition through acoustic array processing and spectral normalization," in *Int. Conf. on Acoustics Speech and Signal processing*, 1997.

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

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

[7] T. Coianiz, L. Torresani, and B. Caprile, "2d deformable models for visual speech analysis," in *Speechreading by Humans and Machines*, Springer-Verlag, 1995.

[8] M. U. R. Sanchez, J. Matas, and J. Kittler, "Statistical chromacity models for lip-tracking with b-splines," in *Int. Conf. on Audio and Video-based Biometric Person Authtication*, 1997.

[9] J. R. Movellan and P. Mineiro, "Modularity and catastrophic fusion: A bayesian approach with applications to audiovisual speech recognition," tech. rep., University of California, Jan. 1996.

[10] K. R. Farrell and R. J. Mammone, *Modern Methods of Speech Processing*, ch. 5, pp. 279–297. Kluwer Academic Publishers, 1995.

[11] 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.