AN EXAMINATION OF AUDIO-VISUAL FUSED HMMS FOR SPEAKER RECOGNITION

David Dean[*]*, Tim Wark*[†] *and Sridha Sridharan*[*]

[*]Speech, Audio, Image and Video Research Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane 4001, Australia
ddean@ieee.org, s.sridharan@qut.edu.au

[†]Queensland University of Technology &
CSIRO ICT Centre
Brisbane 4001, Australia
tim.wark@csiro.au

ABSTRACT

Fused hidden Markov models (FHMMs) have been shown to work well for the task of audio-visual speaker recognition, but only in an output decision-fusion configuration of both the audio- and video-biased versions of the FHMM structure. This paper looks at the performance of the audio- and video-biased versions independently, and shows that the audio-biased version is considerably more capable for speaker recognition. Additionally, this paper shows that by taking advantage of the temporal relationship between the acoustic and visual data, the audio-biased FHMM provides better performance at less processing cost than best-performing output decision-fusion of regular HMMs.

1. INTRODUCTION

The aim of audio-visual speaker recognition (AVSPR) is to make use of complementary information between the acoustic and visual domains to improve the performance of traditional acoustic speaker recognition. Most current approaches to AVSPR either combine the output of individual hidden Markov models (HMMs) in each modality (late fusion), or use a single HMM to classify both modalities (early fusion). Because the decisions or scores are combined at the whole-utterance level, late fusion cannot take true advantage of the temporal dependencies between the two modalities. While early fusion has the advantage that it can take advantage of these dependencies, it often suffers from problems with noise, and has difficulties in modelling the asychronicity of audio-visual speech [1]. The problems with performing AVSPR with early or late fusion have led to the development of middle-fusion methods, or models that accept two streams of input and combine the streams *within* the model to produce a single score or decision.

Most existing approaches to middle-fusion use coupled HMMs, which combine two single-stream HMMs by linking the dependencies of their hidden states. However, due to the small number of hidden states in each modality, these dependencies are often not strong enough to capture the true dependency between the two streams [2]. Fused HMMs (FHMMs) were developed by Pan *et al* [3] by attempting to design a model that maximises the mutual information between the two modalities within a multi-stream HMM. Pan *et al* found that the optimal multi-stream HMM design would result from linking the hidden states of one HMM to the observations of the other, rather than linking the hidden states together, as in a coupled HMM.

This configuration means that FHMMs can be biased towards either modality, and the configuration chosen for AVSPR will depend upon which modality is judged to be more reliable. Additionally the two biased FHMMs can be combined together using late fusion if the comparative reliability of each modality is less clear. In the introductory paper for FHMMs [3], Pan *et al* found that a 50/50 fusion of the two biased FHMMs performed significantly better than a number of alternative AVSPR modelling techniques.

In this paper, we propose to look at the performance of the each of the biased FHMMs individually, rather than in the decision-fusion configuration used by Pan *et al*. By studying the suitability of each of the biased FHMMs to both acoustic and visual degradation, future audio-visual speech research can take advantage of the idiosyncrasies of each biased FHMM. In particular, if recognition can be performed adequately using only a single biased FHMM, the processing required is half that of the fusion of two biased FHMMs. In addition, the performance of the biased FHMMs will be compared to the decision fusion of normal single-stream HMMs.

2. AUDIO-VISUAL FUSED HMMS

2.1. Modelling

Consider two tightly coupled time series $\mathbf{O}^A = \left\{\mathbf{o}_0^A, \mathbf{o}_1^A, \ldots, \mathbf{o}_{T-1}^A\right\}$ and $\mathbf{O}^V = \left\{\mathbf{o}_0^V, \mathbf{o}_1^V, \ldots, \mathbf{o}_{T-1}^V\right\}$, corresponding to audio and video observations respectively. Assume that $\mathbf{O}^A$ and $\mathbf{O}^V$ can be modelled by two HMMs with hidden states $U^x = \left\{u_0^x, u_1^x, \ldots, u_{T-1}^x\right\}$, where x is A or V, respectively. In the FHMM framework, an optimal solution for $p\left(\mathbf{O}^A; \mathbf{O}^V\right)$ according to the maximum entropy

This research was supported by a grant from the Australian Research Council (ARC) Linkage Project LP0562101

principle [3] is given by

$$p\left(\mathbf{O}^A;\mathbf{O}^V\right) = p\left(\mathbf{O}^A\right)p\left(\mathbf{O}^V\right)\frac{p\left(\mathbf{w},\mathbf{v}\right)}{p\left(\mathbf{w}\right)p\left(\mathbf{v}\right)} \quad (1)$$

where $\mathbf{w} = g_A\left(\mathbf{O}^A\right)$, and $\mathbf{v} = g_V\left(\mathbf{O}^V\right)$ are transformations designed such that $p\left(\mathbf{w},\mathbf{v}\right)$ is easier to calculate than $p\left(\mathbf{O}^A,\mathbf{O}^V\right)$, but still reflects the statistical dependence between the two streams. The final term in (1) can therefore be viewed as a correlation weighting, which will be high if $\mathbf{w}$ and $\mathbf{v}$ are related, and low if they are mostly independent.

In [3], Pan *et al* showed that according to maximum mutual information criterion, the transformations g_A and g_V can result in either of the following:

$$\mathbf{w} = \hat{\mathbf{U}}^A, \qquad \mathbf{v} = \mathbf{O}^V \quad (2)$$

$$\mathbf{w} = \mathbf{O}^A, \qquad \mathbf{v} = \hat{\mathbf{U}}^V \quad (3)$$

where $\hat{\mathbf{U}}^x$ is an estimate of the optimal state sequence of HMM x for output $\mathbf{O}^x$.

By invoking (2) in $p\left(\mathbf{O}^A;\mathbf{O}^V\right)$:

$$\begin{aligned} p_A\left(\mathbf{O}^A;\mathbf{O}^V\right) &= p\left(\mathbf{O}^A\right)p\left(\mathbf{O}^V\right)\frac{p\left(\hat{\mathbf{U}}^A,\mathbf{O}^V\right)}{p\left(\hat{\mathbf{U}}^A\right)p\left(\mathbf{O}^V\right)} \\ &= p\left(\mathbf{O}^A\right)p\left(\mathbf{O}^V\middle|\,\hat{\mathbf{U}}^A\right) \end{aligned} \quad (4)$$

where $p\left(\mathbf{O}^A\right)$ can be obtained from the regular audio HMM and $p\left(\mathbf{O}^V\middle|\,\hat{\mathbf{U}}^A\right)$ is the likelihood of getting the video output sequence given the estimated audio HMM state sequence which produced $\mathbf{O}^A$. This equation represents the *audio-biased* FHMM as the main decoding process is the audio HMM.

Similarly, invoking (3) to arrive at the *video-biased* FHMM gives:

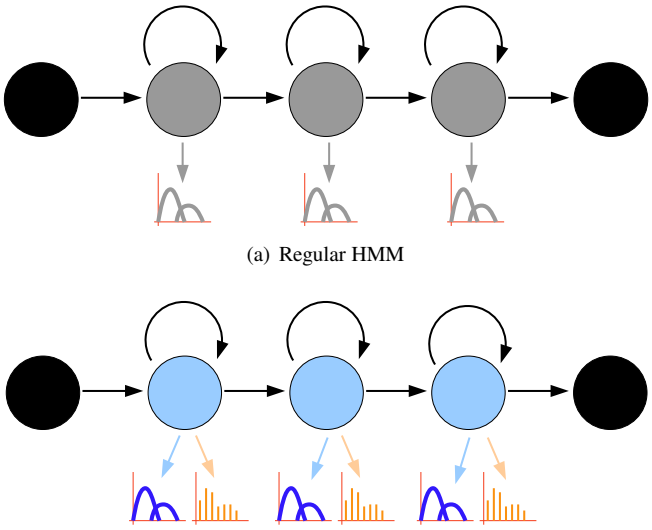
$$p_V\left(\mathbf{O}^A;\mathbf{O}^V\right) = p\left(\mathbf{O}^V\right)p\left(\mathbf{O}^A\middle|\,\hat{\mathbf{U}}^V\right) \quad (5)$$

The choice of the audio- or video-biased FHMM should be chosen upon which individual HMM can more reliably estimate the hidden state sequence for a particular application. Alternatively, both versions can be use concurrently and combined using decision fusion, as in Pan *et al.*

2.2. Training

The training of a biased FHMM is a three step process:

1. The dominant individual HMM is trained independently
2. The best hidden state sequence of the trained HMM is found for each training observation using the Viterbi process [4]



(a) Regular HMM

(b) Fused HMM

Fig. 1. State diagram representations of (a) a regular HMM and (b) a fused HMM

3. The coupling parameters are determined between the hidden state sequences and the subordinate observations

Step 1 establishes the model parameters of the dominant HMM, and step 2 gives the estimate state sequence, $\hat{\mathbf{U}}^d$,of the dominant HMM that produces the dominant training observations, $\mathbf{O}^d$.

The calculation of the coupling parameters are determined as follows:

$$\hat{B}^{d,s} = \arg\max_{B^{d,s}} p\left(\mathbf{O}^s\middle|\,\hat{\mathbf{U}}^d\right) \quad (6)$$

Since the subordinate observations, $\mathbf{O}^s$ and the dominant hidden state sequences, $\hat{\mathbf{U}}^d$ are known, (6) is a typical maximum-likelihood problem. If $\mathbf{O}^s$ is a sequence of discrete symbols, then the subordinate-observation-emission probability-density-function (pdf) for state i is given as:

$$b_i^{d,s}\left(k\right) = \frac{\sum_{t=0}^{T-1}\delta\left(\mathbf{o}_t^s - k\right)\delta\left(\hat{u}_t^d - i\right)}{\sum\delta\left(\hat{u}_t^d - i\right)} \quad (7)$$

where k is a particular observation from the set of possible discrete observations.

2.3. Decoding

Generalising (4) and (5) we can see that:

$$p_d\left(\mathbf{O}^d,\mathbf{O}^s\right) = p\left(\mathbf{O}^d\right)p\left(\mathbf{O}^s\middle|\,\hat{\mathbf{U}}^d\right)$$

As $p\left(\mathbf{O}^d\right) = \sum_{\mathbf{U}^d}p\left(\mathbf{O}^d,\mathbf{U}^d\right)$, and the aim of the decoding process is to find the optimal $\mathbf{U}^d$ by maximising the likeli-

hood, we find the optimal state sequence is given by:

$$\hat{\mathbf{U}}^d = \arg\max_{\mathbf{U}^d} p\left(\mathbf{O}^d, \mathbf{U}^d\right) p\left(\mathbf{O}^s | \hat{\mathbf{U}}^d\right) \quad (8)$$

This can be viewed a special type of HMM that has two observation-emission probability-density-functions for each state, one being the continuous dominant-observation-emission pdf of the regular HMM, and the second being the discrete subordinate-observation-emission pdf derived in (7). An state diagram representation of a FHMM showing both pdfs, and with comparison to a normal HMM, is shown in Figure 1. As each state still provides a single probability within the Viterbi process, the decoding process is otherwise unaffected.

3. EXPERIMENTAL SETUP

3.1. Training and Testing Datasets

For this experiment, training and evaluation data was extracted from the individual speaker section of Clemson University's CUAVE audio-visual database [5]. This database was chosen because, although relatively new, it is the only freely available audio-visual database for researchers to use. The freely available nature of this data makes it ideal for forming benchmarks and comparing research.

Each of the 36 individual speakers in the CUAVE database has a single MPEG2 file containing 16 separate digit sequences. For these experiments the files were split into the individual sequences, and only the isolated-word sequences were used. Of the 5 isolated sequences for each speaker, 4 were used for training, and 1 for testing. These sequences consisted of the speaker saying the digits '*zero*' to '*nine*'.

The data in the testing sequences were also artificially corrupted in both modalities to examine the effect of train/test mismatch on these experiments. The acoustic data was corrupted with additive speech babble noise at a range of signal-to-noise ratios (SNR), and the visual data was corrupted by simulating poor tracking of the lip region-of-interest.

3.2. Feature Extraction

Mel frequency cepstral coefficients (MFCCs) were used to represent the acoustic features in these experiments because of their general application to both speech and speaker recognition. Each feature vector consisted of first 12 MFCCs, normalised energy coefficient, and the first and second time derivatives of those 13 features to result in a 43 dimensional feature vector. These features were calculated every 10 milliseconds using 25 millisecond Hamming-windowed speech signals.

To extract visual features, lip tracking was first performed as in our earlier paper [6] and the resulting lip regions-of-interest (ROIs) were converted to grayscale and reduced to 20 dimensions using discrete cosine transformation (DCT).

Fig. 2. A well-tracked and poorly-tracked lip ROI.

First and second time derivatives of these features were added to form a 60 dimensional feature vector. Corrupted visual data was produced by simulating poorly tracked lip ROIs, being a common problem in a real-life applications. This was achieved by randomly offsetting the accurately tracked ROIs in each frame before extracting the DCT features. An example of both a well- and poorly-tracked ROI is shown in Figure 2.

3.3. Speaker Dependent Word Modelling

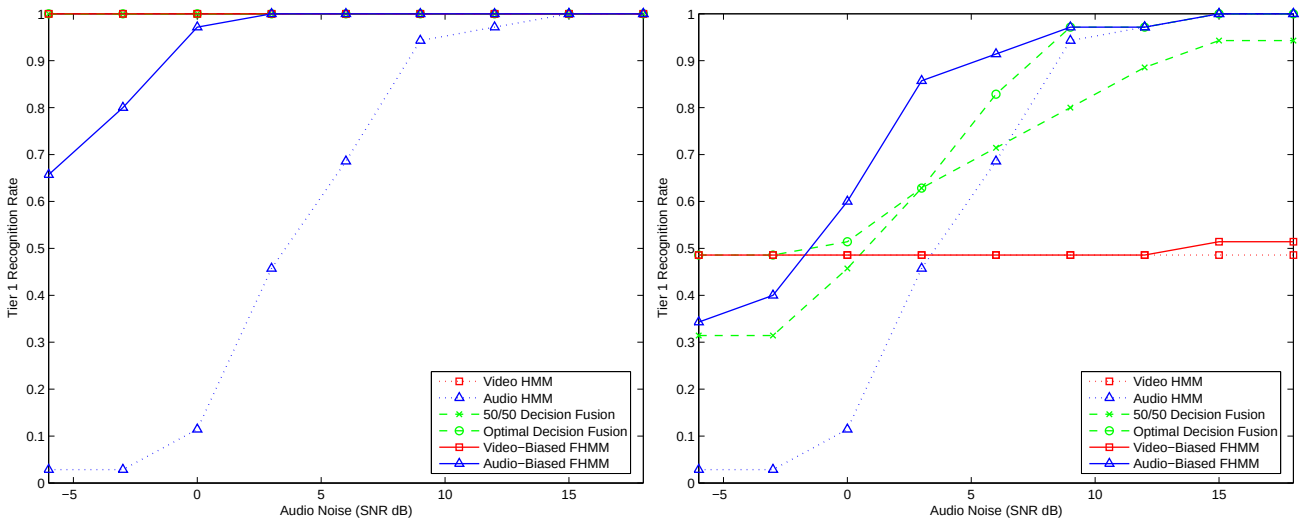
In these experiments both acoustic- and visual-biased FHMMs (A-FHMMs and V-FHMMs) are examined, with the underlying independent HMMs being speaker dependent models for each word in the training and testing sequences.

These word models HMMs were first trained in a speaker-independent manner, then adapted to each speaker using maximum likelihood linear regression (MLLR) adaption. Training was performed using the HMM Toolkit [4].

To calculate the cross-stream coupling parameters for each biased FHMM, the dominant HMM's state sequence for each of the training utterances was estimated using the Viterbi process. A vector-quantisation (VQ) codebook was then generated from the training data in the subordinate stream, and the occurrence of each discrete VQ value within each state of the dominant HMMs was recorded to arrive at an estimate of $p\left(\mathbf{O}^s | \hat{\mathbf{U}}^d\right)$. Decoding of the biased FHMM was then performed by simply multiplying the emission probability of each HMM's state by the estimated $p\left(\mathbf{O}^s | \hat{\mathbf{U}}^d\right)$ within the Viterbi process.

4. SPEAKER RECOGNITION USING FHMMS

Speaker recognition experiments were conducted using both FHMMs, regular single-stream HMMs, and output decision-fusion of regular HMMs. The speaker recognition experiments were performed in a text-dependent manner by scoring each utterance against each speaker's word models for a static word network ('*zero one ... nine*') using either the FHMM or HMM model. The tier 1 recognition performance was calculated as the number of utterances where the correct speaker-models produced the highest score (the 'tier 1' score) as a fraction of the total number of utterances.



(a) Well tracked video

(b) Poorly tracked video

Fig. 3. Comparison of FHMM performance with decision fusion. (In the well tracked results, video, video-biased and decision-fusion are all at 100%)

codebook	well-tracked video		poorly-tracked video	
size	A-FHMM	V-FHMM	A-FHMM	V-FMM
50	93.1%	100%	78.0%	49.4%
100	94.3%	100%	80.6%	49.4%
150	92.0%	100%	79.4%	49.4%
200	91.1%	100%	78.6%	49.4%

Table 1. Average A-FHMM and V-FHMM tier 1 recognition rate for VQ codebook size

Preliminary experiments were first performed at a number of different VQ codebook sizes, as shown in Table 1, and a codebook of size 100 was determined to be suitable for both audio and video.

In addition to comparing FHMMs with regular HMM recognition, output decision-fusion of the regular HMMs was implemented to serve as a baseline for comparison. Each score for a particular utterance from each modality was combined using weighted-sum decision-fusion:

$$\hat{s}_F \quad = \quad \alpha \times \hat{s}_A + (1-\alpha) \times \hat{s}_V$$

where $\hat{s}_F$ is the fusion score and $\hat{s}_A$ and $\hat{s}_V$ are the acoustic and visual HMM scores respectively. The choice of α denotes the perceived reliability of each modality, with $\alpha = 0$ being video only, and $\alpha = 1$ is audio only.

The FHMM performance was found to have no impact of the well-tracked video data, as the video performance was perfect in that case, and there was no need for the acoustic modality at all. However, as the manually-assisted tracking of the video data is not indicative of what would be experienced in a 'real world' scenario, we used the artificially mistracked video data to simulate such a scenario.

The response of each system to both types of video, and varying levels of speech-babble noise is shown in Figure 3. Two decision-fusion scenarios have been included: 50/50, or $\alpha = 0.5$, and best-performing decision-fusion. The best-performing scenario output indicates the best performance for any α value at each level of acoustic noise. This is the performance of an theoretically optimal *adaptive* decision-fusion system, one which can determine the noise level and adjust the α-value accordingly.

5. DISCUSSION

The results of these experiments have shown that there is a major improvement provided by the incorporation of the video stream into an A-FHMM over the underlying audio HMM. However, there is little improvement provided by incorporation of the acoustic stream into a V-FHMM over the video HMM alone.

The main reason for this discrepancy is the poor underlying state sequence of the video HMM. While video data is very good at recognising speakers, it is comparatively poor at recognising speech [6]. If the video HMM's hidden state sequences do not consistently line up with similar speech events, then the coupling between these states and the acoustic data will not reflect the true relationship between the two modalities.

While the V-FHMM does not seem to be any better than the underlying video HMM, in the poorly tracked data the A-FHMM is performing equal to, or in some cases much better than the optimal decision-fusion case in all but very noisy au-

dio. This performance increase is also given at around half the decoding cost, as only one decoding process is required for FHMMs, as compared to two (audio and video) for output decision-fusion of regular HMMs.

The primary reason for the A-FHMM performing so well in comparison to the optimal decision-fusion case is that the A-FHMM can take advantage of the relationship between the actual features in each modality on a frame-by-frame basis, while decision-fusion can only look at the relationship between scores for an entire utterance. Even when the underlying acoustic HMM would perform poorly, the audio state sequences which corresponded best with the video data were weighted up, allowing the HMM to score the utterance reliably.

Another major advantage of the A-FHMM, over the improved performance, is that it can be run 'blind', or with little knowledge of the environmental conditions. Designing a decision-fusion speaker recognition system requires that the best fusion parameter (α) must be estimated either for each noise level, or for the entire operating noise-range of the recognition environment. In addition, if the fusion parameter is estimated for each noise level, the noise level in a particular sample must then be estimated itself in some manner, which is still an ongoing area of research [7]. In comparison, the A-FHMM is running with the same parameters for all noise levels, and is providing equal or better performance for all but the noisiest acoustic levels. Even in the well tracked data, the A-FHMM only performs worse than the video HMM for acoustic signal-to-noise ratios of 0 dB or below (i.e. noise $\geq$ signal), and could be a good choice given that the quality of the video wouldn't always be known.

6. CONCLUSION AND FUTURE RESEARCH

In this paper, we have shown a clear difference between the performance of the two biased versions of the fused HMM design. While the coupling between the acoustic states of an A-FHMM and the video data can provide a drastic improvement to a normal acoustic HMM, the poor reliability of video HMM state sequence estimate leaves V-FHMMs performing no better than regular video HMMs.

In both well and poor tracked video, and under all but the noisiest acoustic conditions, the acoustic-biased FHMM has shown that it can achieve results equal to or better than the optimal decision-fusion of both modalities HMMs at less processing cost as only one decoder is required instead of two. In addition, the acoustic-biased FHMM can largely run 'blind', with no fine-tuning required to perform well under a wide range of acoustic noise levels.

As the CUAVE database is quite small for speaker recognition experiments, future research will focus on extending these experiments onto the XM2VTS [8] database. The much larger size of the XM2VTS corpus, in both number of speakers and speech length, should allow for a more thorough study of the mechanics of FHMMs, and a better idea of their applicability in real-world scenarios.

Additionally, FHMMs should prove quite suitable in other areas relating to audio-visual speech, such as speech recognition or speaker detection.

7. ACKNOWLEDGEMENTS

The authors wish to thank Clemson University for freely supplying their CUAVE audio-visual database [5] for our research.

8. REFERENCES

[1] C. Chibelushi, F. Deravi, and J. Mason, "A review of speech-based bimodal recognition," *Multimedia, IEEE Transactions on*, vol. 4, no. 1, pp. 23–37, 2002.

[2] M. Brand, "A bayesian computer vision system for modeling human interactions," in *ICVS'99*, Gran Canaria, Spain, 1999.

[3] H. Pan, S. Levinson, T. Huang, and Z.-P. Liang, "A fused hidden markov model with application to bimodal speech processing," *IEEE Transactions on Signal Processing*, vol. 52, no. 3, pp. 573–581, 2004.

[4] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, 3rd ed. Cambridge, UK: Cambridge University Engineering Department., 2002.

[5] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Cuave: a new audio-visual database for multimodal human-computer interface research," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, vol. 2, 2002, pp. 2017–2020.

[6] D. Dean, P. Lucey, S. Sridharan, and T. Wark, "Comparing audio and visual information for speech processing," in *ISSPA 2005*, Sydney, Australia, 2005, pp. 58–61.

[7] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognition*, vol. 36, no. 2, pp. 293–302, 2003.

[8] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *Audio and Video-based Biometric Person Authentication (AVBPA '99), Second International Conference on*, Washington D.C., 1999, pp. 72–77.