



COVER SHEET

This is the author-version of article published as:

Dean, David and Sridharan, Sridha and Wark, Tim (2006) Audio-visual speaker verification using continuous fused HMMs. In *Proceedings HCSNet Workshop on the Use of Vision in HCI, Canberra, Australia.*

Copyright 2006 [Australian Computer Society](#)

Accessed from <http://eprints.qut.edu.au>

Audio-Visual Speaker Verification using Continuous Fused HMMs

David Dean¹

Sridha Sridharan¹

Tim Wark^{1,2}

¹Speech, Audio, Image and Video Research Laboratory, Queensland University of Technology

²CSIRO ICT Centre

Brisbane, Australia

ddean@ieee.org, s.sridharan@qut.edu.au, tim.wark@csiro.au

Abstract

This paper examines audio-visual speaker verification using a novel adaptation of fused hidden Markov models, in comparison to output fusion of individual classifiers in the audio and video modalities. A comparison of both hidden Markov model (HMM) and Gaussian mixture model (GMM) classifiers in both modalities under output fusion shows that the choice of audio classifier is more important than video. Although temporal information allows a HMM to outperform a GMM individually in video, this temporal information does not carry through to output fusion with an audio classifier, where the difference between the two video classifiers is minor. An adaptation of fused hidden Markov models, designed to be more robust to within-speaker variation, is used to show that the temporal relationship between video observations and audio states can be harnessed to reduce errors in audio-visual speaker verification when compared to output fusion.

Keywords: audio-visual speaker recognition (AVSPR), fused hidden Markov model (FHMM)

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 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 modeling the asynchronicity of audio-visual speech (Chibelushi, Deravi & Mason 2002). The problems with performing AVSPR with early or late fusion have led to the development of middle-fusion methods, or mod-

els that accept two streams of input and combine the streams *within* the model to produce a single score.

Most existing approaches to middle-fusion use coupled HMMs (Nefian, Liang, Fu & Liu 2003), 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 relationship between the two streams (Brand 1999). Fused HMMs (FHMMs) were developed (Pan, Levinson, Huang & Liang 2004) by attempting to design a model that maximises the mutual information between the two modalities within a multi-stream HMM. Pan et al. (2004) 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.

In this paper, we first introduce a novel adaptation of Pan et al's FHMMs, designed to be more robust to within-speaker variation. A comparison of a number of different audio-visual output-fusion configurations is performed to obtain an insight into the temporal information available in both audio and video, individually and combined for the purposes of speaker verification. Finally we examine the ability of our FHMM model to take better advantage of the temporal dependencies between the modalities than is possible with output fusion alone.

2 Fused Hidden Markov Models

2.1 Theory

Consider two tightly coupled time series $\mathbf{O}^A = \{\mathbf{o}_0^A, \mathbf{o}_1^A, \dots, \mathbf{o}_{T-1}^A\}$ and $\mathbf{O}^V = \{\mathbf{o}_0^V, \mathbf{o}_1^V, \dots, \mathbf{o}_{T-1}^V\}$, corresponding to audio and video observations respectively. Assume that \mathbf{O}^A and \mathbf{O}^V can be modeled by two HMMs with hidden states $U^x = \{u_0^x, u_1^x, \dots, u_{T-1}^x\}$, where x is A or V , respectively. In the FHMM framework, an optimal solution for $p(\mathbf{O}^A; \mathbf{O}^V)$ according to the maximum entropy principle (Pan, Liang & Huang 2001) is given by

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

where $\mathbf{w} = g_A(\mathbf{O}^A)$, and $\mathbf{v} = g_V(\mathbf{O}^V)$ are transformations designed such that $p(\mathbf{w}, \mathbf{v})$ is easier to calculate than $p(\mathbf{O}^A, \mathbf{O}^V)$, 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. Pan et al. (2001) also showed that the minimum distance between $\tilde{p}(\mathbf{O}^A; \mathbf{O}^V)$ and the ground truth $p(\mathbf{O}^A, \mathbf{O}^V)$

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

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

is established when the mutual information between \mathbf{w} and \mathbf{v} is maximised:

$$(\hat{\mathbf{w}}, \hat{\mathbf{v}}) = \arg \max_{(\mathbf{w}, \mathbf{v}) \in \theta} \mathcal{I}(\mathbf{w}, \mathbf{v}) \quad (2)$$

In their audio-visual FHMM paper, Pan et al. (2004) chose \mathbf{w} and \mathbf{v} empirically from the following set (θ):

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

$$\mathbf{w} = \hat{\mathbf{U}}^A, \quad \mathbf{v} = \hat{\mathbf{U}}^V \quad (4)$$

$$\mathbf{w} = \mathbf{O}^A, \quad \mathbf{v} = \hat{\mathbf{U}}^V \quad (5)$$

where $\hat{\mathbf{U}}^x$ is an estimate of the optimal state sequence of HMM x for output \mathbf{O}^x . By invoking (2) over the set θ and invoking the following inequality in information theory

$$\mathcal{I}(x, f(y)) \leq \mathcal{I}(x, y) \quad (6)$$

And that estimated hidden state sequences can be viewed as a function of the output ($\hat{\mathbf{U}}^x = f_x(\mathbf{O}^x)$), Pan et al. (2004) concluded that

$$\mathcal{I}(\hat{\mathbf{U}}^A, \hat{\mathbf{U}}^V) = \mathcal{I}(\hat{\mathbf{U}}^A, f_V(\mathbf{O}^V)) \leq \mathcal{I}(\hat{\mathbf{U}}^A, \mathbf{O}^V) \quad (7)$$

$$\mathcal{I}(\hat{\mathbf{U}}^A, \hat{\mathbf{U}}^V) = \mathcal{I}(f_A(\mathbf{O}^A), \hat{\mathbf{U}}^V) \leq \mathcal{I}(\mathbf{O}^A, \hat{\mathbf{U}}^V) \quad (8)$$

Therefore the transforms (3) and (5) produce better estimates of $\tilde{p}(\mathbf{O}^A; \mathbf{O}^V)$ than (4). By invoking (3) in $p(\mathbf{O}^A; \mathbf{O}^V)$:

$$\begin{aligned} p_A(\mathbf{O}^A; \mathbf{O}^V) &= p(\mathbf{O}^A) p(\mathbf{O}^V) \frac{p(\hat{\mathbf{U}}^A, \mathbf{O}^V)}{p(\hat{\mathbf{U}}^A) p(\mathbf{O}^V)} \\ &= p(\mathbf{O}^A) p(\mathbf{O}^V | \hat{\mathbf{U}}^A) \end{aligned} \quad (9)$$

where $p(\mathbf{O}^A)$ can be obtained from the regular audio HMM and $p(\mathbf{O}^V | \hat{\mathbf{U}}^A)$ 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 (5) to arrive at the *video-biased* FHMM gives:

$$p_V(\mathbf{O}^A; \mathbf{O}^V) = p(\mathbf{O}^V) p(\mathbf{O}^A | \hat{\mathbf{U}}^V) \quad (10)$$

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 used concurrently and combined using output fusion, as in Pan et al. (2004).

2.2 Continuous FHMMs

In the original implementation of FHMMs (Pan et al. 2004), the subordinate modality features were treated as discrete symbols through vector-quantisation codebooks to simplify the calculation of the coupling parameters. However this simplification caused problems with within-speaker session variability, especially when the video was the subordinate modality. While audio-biased FHMMs (A-FHMMs) worked well in experiments on the CUAVE database (Dean,

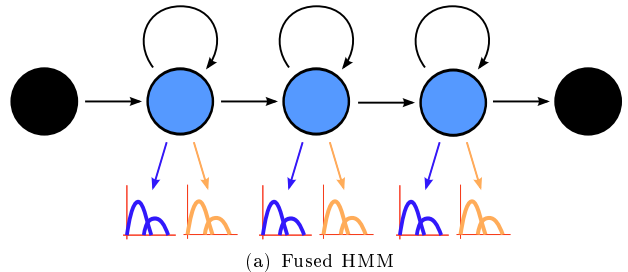


Figure 1: State diagram representation of a FHMM. (Compare to a regular HMM in figure 2.)

		Configuration											
		1	2	3	4	5	6	7	8	9	10	11	12
Session	1	Train	Train	Train	Train	Eval	Test	Eval	Test	Eval	Test	Eval	Test
	2	Train	Eval	Test	Eval	Test	Train	Train	Train	Train	Test	Eval	Test
	3	Eval	Test	Train	Test	Eval	Train	Test	Eval	Train	Test	Eval	Train
	4	Test	Eval	Test	Eval	Train	Test	Eval	Train	Train	Train	Train	Train

Table 1: XM2VTS dataset configurations used in these experiments

Wark & Sridharan 2006), the change in codebook values caused by a change in session outweighed that due to a change in speaker, rendering the discrete FHMM worse than the underlying HMM when used in a multi-session database like XM2VTS.

To allow the FHMM structure to more robustly model the subordinate modality, we proposed modeling the relationship between the dominant states and the subordinate observations using an extra GMM within each of the dominant states. Therefore our *continuous* FHMM (as opposed to Pan et al’s *discrete* FHMM) can be viewed as a regular HMM with two GMM-based output probability distributions instead of one in a normal HMM, as shown in Figure 1.

3 Experimental Setup

3.1 Training and Testing Datasets

For this experiment, training, testing and evaluation data were extracted from the digit-video sections of the XM2VTS database (Messer, Matas, Kittler, Luetten & Maitre 1999). The training and testing configurations used for these experiments was based on the XM2VTSDB protocol (Luetten & Maitre 1998), but adapted to allow more tests than provided by the protocol. Each of the 295 speakers in the database has four separate sessions of video where the speaker speaks two sequences of two sentences of ten digits. In each of the configurations, two sessions were used for training, one for evaluation and one for testing, allowing for 12 configurations in total, as shown in Table 1. By comparison, the XM2VTSDB protocol only allows for the first configuration.

These experiments were performed as verification experiments, where the speaker would attempt to enter the system by claiming the identity of a particular client. To perform this task, the speakers were split into two groups: clients, who claimed their own identity; and imposters, who claimed the identity of one of the clients.

As per the XM2VTSDB protocol, 200 speakers were designated clients, and 95 were used as imposters. For each client testing sequence (2 per session), 20 sequences were chosen at random from the imposter set allowing for a total of 400 (200×2) client tests and 8000 ($200 \times 2 \times 20$) imposter tests for each configuration. Over all 12 configurations, 4800 client tests and 96000 imposter tests are performed.

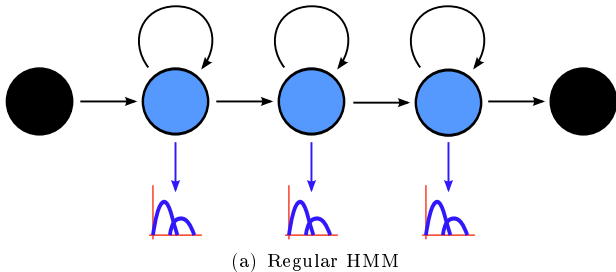


Figure 2: Regular HMM. The output probability of each state is implemented as a GMM.

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 the 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.

Visual features were extracted from a manually tracked lip region-of-interest (ROI) from 25 fps (40 milliseconds / frame) video data. Manual tracking of the locations of the eyes and lips were performed every 50 frames, and the remainder of the frames were interpolated from the manual tracking. The eye locations were used to normalise the rotation of the lips. A rectangular region-of-interest, 120 pixels wide and 80 pixels tall, centered around the lips was extracted from each frame in the video. Each ROI was then reduced to 20% of its original size (24×16 pixels) and converted to grayscale. Finally the ROI was reduced to 20 dimensions using discrete cosine transformation (DCT) (Heckmann, Kroschel, Savariaux & Berthommier 2002). First and second time derivatives of these features were added to form a 60 dimensional feature vector.

4 Audio-Visual Speaker Verification using Output Fusion

4.1 Training

Two classifier-types were used for each modality, for a total of four output-fusion experiments. The two classifiers used were Gaussian mixture models (GMMs), which are good at modeling static, or time-independent, variables, and HMMs, which are better at modeling temporal variables. This can be observed by examining a standard HMM design: HMMs are commonly implemented as a chain of GMMs, as shown in Figure 2, where the HMM controls the likelihood of moving between states, and the GMM-states control the likelihood of outputting certain features when in an emitting state. Conversely, a GMM can be viewed as HMM with only one emitting state.

Both HMM and GMM speaker-dependent models were generated by adapting background models to each individual speaker. The background models were generated using the training sequences for each configuration over both clients and impostors. These models were then adapted to each individual client speaker’s training sequences using maximum a posteriori (MAP) adaptation (Lee & Gauvain 1993).

GMM models were trained over all training sequences, whereas HMM models were trained for each word. Empirical experiments were performed on

Model	Mixtures	States
Audio HMM	9	7
Audio GMM	256	-
Video HMM	16	7
Video GMM	8	-

Table 2: Best performing topologies for each classifier.

a single configuration to determine the best topology, shown in Table 2. HMM training was performed using the HTK toolkit (Young, Evermann, Kershaw, Moore, Odell, Ollason, Povey, Valtchev & Woodland 2002), and GMM training with internally developed utilities.

4.2 Testing

For each of the four client models trained in the previous section, two client sequences and 40 impostor sequences were verified using that model for each configuration. Scores obtained from the client models were normalised for length and environment by subtracting the background-model score for the same sequence.

In addition to the individual models, the four possible output-fusion combinations of audio and video classifiers were also examined, as listed below:

- Audio HMM + Video HMM
- Audio HMM + Video GMM
- Audio GMM + Video HMM
- Audio GMM + Video GMM

Given that the parameters of the score-distribution vary considerably between classifiers, the evaluation session of each configuration is used to get an estimation of each classifier’s score distribution, which is used to normalise the scores.

$$Z_i(s_i) = \frac{s_i - \hat{\mu}_i}{\hat{\sigma}_i} \quad (11)$$

Where s_i is the score from classifier i and $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the estimated mean and standard deviations of classifier i ’s score distribution. Therefore the output-fusion score for each combination is calculated as

$$s_F = \frac{Z_a(s_a) + Z_v(s_v)}{2} \quad (12)$$

Where a is the audio classifier and v is the video classifier.

4.3 Results

Detection error trade-off (DET) plots showing the performance of both the individual classifiers and the four output-fusion combinations for speaker verification are shown in Figure 3.

From a comparison of the HMM and GMM performance for each modality, it can be clearly seen that there is temporal information in both the audio and video features. Whilst the audio GMM performs nearly as well as the audio HMM, it is only through using a much higher number of mixtures (256 vs 9). However, in the video we found that the GMM performance could not be made to match the HMM’s, regardless of the number of mixtures used.

However, the clear improvement of using a video HMM over a video GMM does not appear to translate over to output fusion. The main differences in output fusion appears to be related to the audio classifier chosen and not the video. The video HMM does appear to improve output fusion slightly in areas of

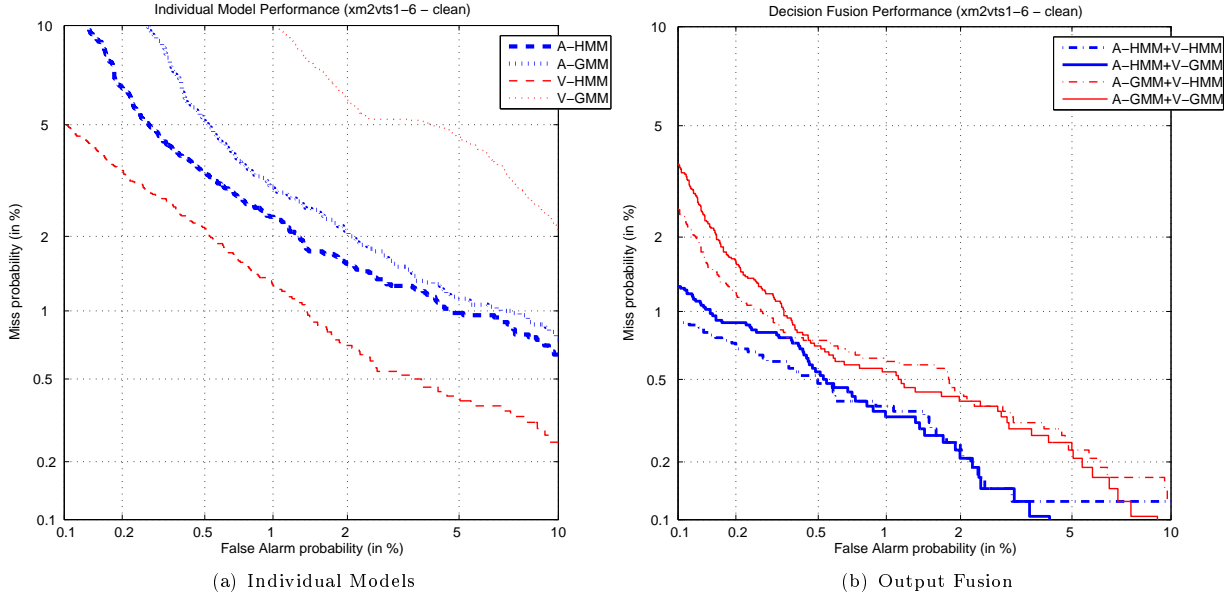


Figure 3: Detection error trade-off (DET) plots for output-fusion speaker verification.

low false alarm, but it does not provide a major improvement that the difference of the two classifiers in video alone might indicate. So, while the video HMM clearly takes advantage of temporal video information when compared to the video GMM, this temporal information provides little benefit in output fusion where a static GMM would work almost as well. It is also clear that output fusion cannot take advantage of temporal dependencies between the two modalities, as the only information fused together is the classifier’s scores over an entire utterance.

5 Audio-Visual Speaker Verification using FHMMs

5.1 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 (Young et al. 2002)
3. The relationship between the hidden state sequences and the subordinate observations are modeled

For these experiments, both audio- and video-biased FHMMs were examined, so the underlying HMMs trained in Step 1 were the audio HMM and the video HMM as trained in Section 4.1, respectively.

The relationship between the hidden state sequences and the subordinate observations is contained in $p(\mathbf{O}^s | \hat{\mathbf{U}}^d)$ where d represents the dominant modality, and s the subordinate. This is basically defined as the likelihood of getting a subordinate observation when in a particular dominant state. Once the estimated hidden state sequence, $\hat{\mathbf{U}}^d$, for the training data was determined in Step 2, the subordinate training observations were segmented based on the word and state boundaries. Each speaker’s GMM (trained in Section 4.1) was then adapted for each word and state within their training sequences to form

the FHMM’s subordinate GMMs. The background GMM was also adapted to each word and state and added to the background HMM to form the background FHMM. The optimal number of mixtures for the subordinate GMMs was found empirically to be the same as that for individual GMM classifiers, that being 256 for the audio and 8 for the video.

The training sequence was performed twice, once with audio as the dominant modality, and once with video dominant to form the audio- and video-biased FHMMs respectively.

5.2 Testing

Generalising (9) and (10) we can see that:

$$p_d(\mathbf{O}^d, \mathbf{O}^s) = p(\mathbf{O}^d) p(\mathbf{O}^s | \hat{\mathbf{U}}^d) \quad (13)$$

Where d represents the dominant modality, and s the subordinate. As $p(\mathbf{O}^d) = \sum_{\mathbf{U}^d} p(\mathbf{O}^d, \mathbf{U}^d)$, and the aim of the decoding process is to find the optimal \mathbf{U}^d by maximising the likelihood, we find the optimal state sequence is given by:

$$\hat{\mathbf{U}}^d = \arg \max_{\mathbf{U}^d} p(\mathbf{O}^d, \mathbf{U}^d) p(\mathbf{O}^s | \mathbf{U}^d) \quad (14)$$

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 GMM of the regular HMM, and the second being the continuous subordinate-observation-emission GMM trained in Section 5.1. As these scores are combined within each state, and each state still provides a single probability within the Viterbi process, the decoding process is otherwise unaffected.

Before the scores for each modality are combined within the state, they are normalised for each modality based on the evaluation data set, similar to the normalisation performed for output fusion in Section 4.2, but on a frame-by-frame basis rather than over an entire sequence. Because we found the difference in frame-scores between modalities is more significant that the difference in scores between speakers, the background dominant HMM and subordinate GMM individual models were evaluated for each

frame over the evaluation sequences for each configuration to come up with an estimate of each classifier’s score distribution which was then used to normalise the GMM scores within each FHMM state using (12). The features evaluated for each modality’s score is determined by the frame-rate of the dominant HMM, with the subordinate features chosen being the closest in time to the dominant features.

In addition to using models adapted to a specific word-state for the subordinate modality, models adapted to all states of a particular word, and just using the global speaker GMM in this role was considered. These three choices will be referred to as word-state GMMs, word GMMs and global GMMs for the remainder of this paper. By examining the difference in performance between these subordinate models in the FHMM structure, we can make some conclusions about the temporal dependencies captured by the FHMMs.

Finally, scores obtained from the client FHMM models were normalised for length and environment by subtracting the background-model FHMM score for the same sequence.

5.3 Comparison with Output Fusion

It can be seen that using the global speaker GMM should be functionally equivalent to a output fusion of the GMM and the underlying HMM. This is because at a base level the output HMM likelihood can be mathematically defined as:

$$p(\mathbf{O}) = \prod_t p_h(o_t|u_t) \quad (15)$$

Where $p_h(o_t|u_t)$ is the likelihood of the HMM outputting observation o_t whilst in state u_t at time t . Fusing the output of this HMM with a single GMM’s output ($p_g(o_t)$) results in:

$$p(\mathbf{O}^d, \mathbf{O}^s) = \prod_t p_h(o_t^d|u_t^d) \times \prod_t p_g(o_t^s) \quad (16)$$

$$= \prod_t [p_h(o_t^d|u_t^d) p_g(o_t^s)] \quad (17)$$

This is equivalent to multiplying the regular HMM and global subordinate GMM within the Viterbi process of the FHMM, assuming that the addition of the p_g term does not affect the best path chosen through the lattice, and therefore the value of u_t above. But, as the p_g term does not depend upon the value of u_t , every path in the lattice should be affected equally, and therefore the best path should remain the same.

However, there are other differences of implementation between the global subordinate-GMM FHMM and the output fusion presented above that make them slightly different for the purposes of these experiments. For the two products in (16) above to be combined to form (17), they must be multiplying over the same range of t -values, which is not the case here due to the different frame rates of each modality. Additionally, the normalisation performed in the FHMM nodes and also in the output fusion occur at different levels, introducing differences. Nevertheless, these factors could be easily controlled for, allowing output fusion to work as well as the global-subordinate-GMM-based FHMM model.

In a similar manner to this, the word and word-state subordinate-GMM-based FHMM models could be viewed as almost equivalent to HMM-GMM output fusion, provided that the sequence is first segmented into words or word-states, respectively, using the underlying HMM, and the correct subordinate GMM is

chosen for each segment. This is effectively what the FHMM model is doing with the significant difference being that the score-fusion occurs within the Viterbi process, so that the boundaries of the words or word-states have the possibility of moving based upon the subordinate observations. It is not clear at this stage how much this is in effect, and this will be covered in a future paper in more detail.

5.4 Results

DET plots showing the performance of our audio- and video-biased FHMM structures are shown in Figure 4. By comparing to the output fusion of the audio and video HMM, shown in both plots, it can be seen that the audio-biased structure is clearly more powerful than the video-biased version.

For the video-biased FHMMs, the word and word-state subordinate models fare considerably worse than the global subordinate model. As the global-subordinate-model can be replicated with output fusion, as discussed in the previous section, there is therefore little need of video-biased FHMMs in this situation. However, for audio-biased FHMMs there does appear to be a small benefit in using the word-state, or word FHMM over the global FHMM, particularly around the equal-error-rate region.

The main reason for the difference in performance between the two FHMM configurations is the ability of the dominant HMM to reliably estimate its underlying state sequence. The performance of the audio-biased FHMM shows that the audio HMM can reliably segment the sequences into sections of similar video appearance, but the video HMM does not appear able to locate segments of similar audio activity. Although the performance increase in this case is not large, the improved performance of the word-state FHMM over the global FHMM does appear to show that it is taking advantage of a temporal relationship between the audio states and video features.

6 Conclusion and Future Research

In this paper we have examined output fusion using both HMM and GMM classifiers in both the audio and video modalities and found that although temporal video information is clearly useful for lip-based speaker recognition using video HMMs, under output fusion most of this information appears to be lost. The performance of output fusion appears to be based mostly on the audio-classifier chosen, with the HMM performing better, and the choice of video classifier appears to only have a minor effect.

In an attempt to take greater advantage of the temporal video information in fusion with the audio, we adapted Pan et al.’s (2004) FHMMs to improve the robustness of the subordinate models to within-speaker variations, particularly on data recorded over multiple sessions. We found that our continuous FHMM model took advantage of the temporal relationship between the video observations and audio states to improve performance over the best performing output fusion in an audio-biased configuration. However, we found that the video-biased configuration showed no useful relationship between audio observations and video states.

In the audio-biased FHMM structure, a large portion of the video subordinate-GMMs are used to recognise primarily static features, such as lip or skin colour, which do not change throughout the sequence. As this type of information cannot form a temporal relationship with audio states, its effect on the subordinate-GMMs may be swamping the more dynamic information available in the movement of the

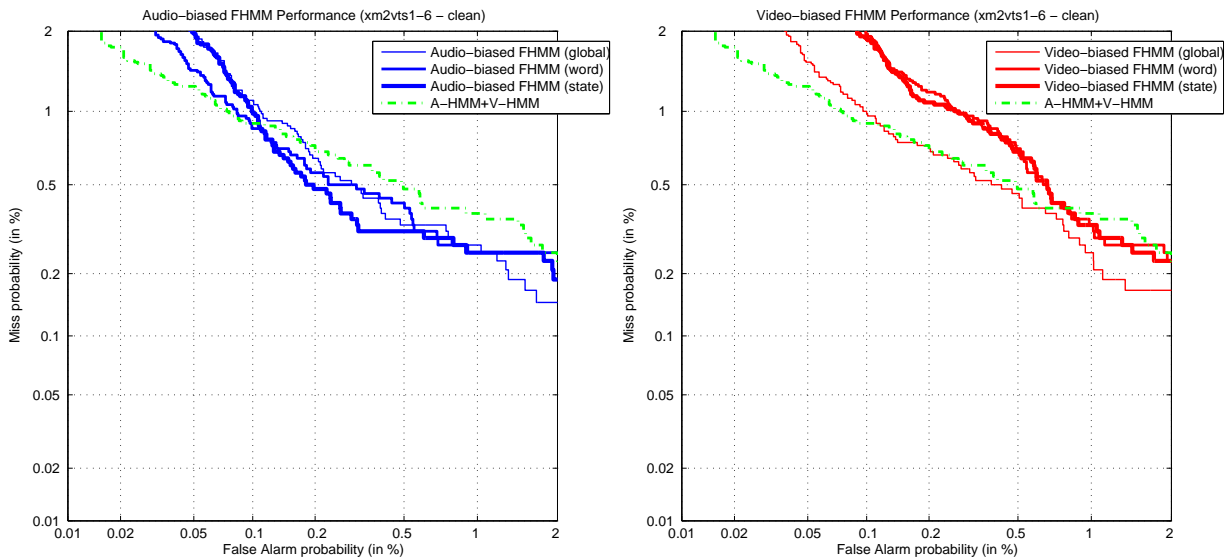


Figure 4: Detection error trade-off (DET) plots for FHMM speaker verification. (Note that the scale has changed from Figure 3.)

lips that could provide an improvement in the FHMM structure. A more efficient FHMM structure may be able to be realised by using more dynamic video features, and then performing output-fusion with a simple classifier using the static features so that the static information is not lost completely. Methods such as mean-image removal, optical flow or contour-based lip representations should provide better features to model the dynamic nature of visual speech.

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

7 Acknowledgments

The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in (Messer et al. 1999) or at <http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/>.

References

- Brand, M. (1999), A bayesian computer vision system for modeling human interactions, in 'ICVS'99', Gran Canaria, Spain.
- Chibelushi, C., Deravi, F. & Mason, J. (2002), 'A review of speech-based bimodal recognition', *Multimedia, IEEE Transactions on* **4**(1), 23–37.
- Dean, D., Wark, T. & Sridharan, S. (2006), An examination of audio-visual fused HMMs for speaker recognition, in 'MMUA 2006', Toulouse, France.
- Heckmann, M., Kroschel, K., Savariaux, C. & Berthommier, F. (2002), DCT-based video features for audio-visual speech recognition, in 'International Conf. on Spoken Language Processing', Denver, Colorado, pp. 92093–0961.
- Lee, C.-H. & Gauvain, J.-L. (1993), Speaker adaptation based on MAP estimation of HMM parameters, in 'Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on', Vol. 2, pp. 558–561 vol.2.
- Luetttin, J. & Maitre, G. (1998), Evaluation protocol for the extended M2VTS database (XM2VTSDB), Technical report, IDIAP.
- Messer, K., Matas, J., Kittler, J., Luetttin, J. & Maitre, G. (1999), XM2VTSDB: The extended M2VTS database, in 'Audio and Video-based Biometric Person Authentication (AVBPA '99), Second International Conference on', Washington D.C., pp. 72–77.
- Nefian, A. V., Liang, L. H., Fu, T. & Liu, X. X. (2003), A bayesian approach to audio-visual speaker identification, in 'Audio-and Video-Based Biometric Person Authentication (AVBPA 2003), 4th International Conference on', Vol. 2688 of *Lecture Notes in Computer Science*, Springer-Verlag Heidelberg, Guildford, UK, pp. 761–769.
- Pan, H., Levinson, S., Huang, T. & Liang, Z.-P. (2004), 'A fused hidden markov model with application to bimodal speech processing', *IEEE Transactions on Signal Processing* **52**(3), 573–581.
- Pan, H., Liang, Z.-P. & Huang, T. S. (2001), 'Estimation of the joint probability of multisensory signals', *Pattern Recognition Letters* **22**(13), 1431–1437.
- Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2002), *The HTK Book*, 3.2 edn, Cambridge University Engineering Department., Cambridge, UK.