Fused HMM-Adaptation of Multi-Stream HMMs for Audio-Visual Speech Recognition

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

[*]Speech, Audio, Image and Video Research Laboratory, Queensland University of Technology
[†]CSIRO ICT Centre
Brisbane, Australia
ddean@ieee.org, {p.lucey, s.sridharan}@qut.edu.au, tim.wark@csiro.au

10.21437/Interspeech.2007-285

Abstract

A technique known as fused hidden Markov models (FHMMs) was recently proposed as an alternative multi-stream modelling technique for audio-visual *speaker* recognition. In this paper we show that for audio-visual *speech* recognition (AVSR), FHMMs can be adopted as a novel method of training synchronous MSHMMs. MSHMMs, as proposed by several authors for use in AVSR, are jointly trained on both the audio and visual modalities. In contrast our proposed FHMM adaptation method can be used to adapt the multi-stream models from single-stream audio HMMs, and in the process, better model the video speech in the final model when compared to jointly-trained MSHMMs. By experiments conducted on the XM2VTS database we show that the improved video performance of the FHMM-adapted MSHMMs results in an improvement in AVSR performance over jointly-trained MSHMMs at all levels of audio noise, and provide significant advantage in high noise environments.

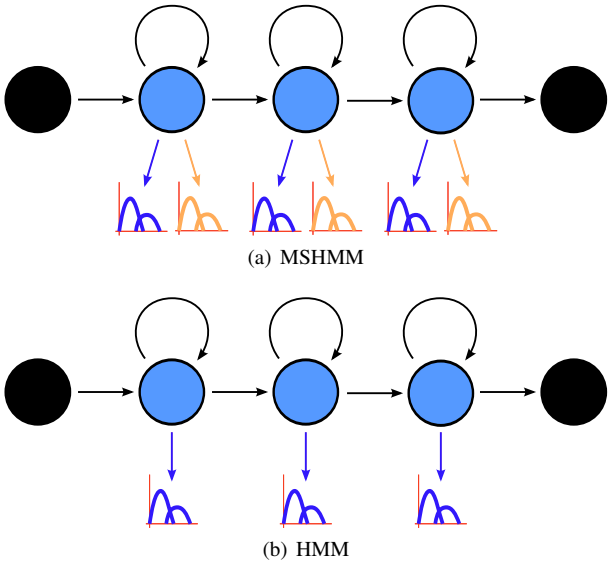
Index Terms: audio-visual speech recognition, fused hidden Markov models, multi-stream hidden Markov models

1. Introduction

Human speech is inherently bimodal in nature [1], and the aim of audio-visual speech recognition (AVSR) is to exploit this bimodality by using the complementary information between the acoustic and visual domains to increase the performance of traditional acoustic speech recognition, particularly in noisy acoustic conditions. In terms of modelling the relationship between the two modalities, MSHMMs can be seen as providing a middle ground for AVSR between feature fusion and asynchronous HMMs [2]. Unlike feature fusion, MSHMMs can model the reliability of each stream independently, but they cannot model the asynchronicity between the two streams as asynchronous HMMs can [3]. However, the small performance benefit of modelling the asynchronicity may not be worth the increase in model complexity, such as in embedded environments where processing power or memory may be limited.

In this paper we will look at an alternative technique of training MSHMMs based on a modelling technique called fused HMMs (FHMMs), originally introduced as an alternative multi-stream model design for audio-visual *speaker* recognition [4]. By using a method extended from the original FHMM model design [5], we will show that by adapting a MSHMM from an audio-only HMM, we can increase the AVSR performance, particularly in noisy acoustic environments.

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



(a) MSHMM

(b) HMM

Figure 1: *State diagram representation of a MSHMM compared to a regular HMM*

2. Multi-stream hidden Markov models

A MSHMM can be viewed as a regular single-stream HMM, but with two observation-emission Gaussian mixture models (GMMs) for each state–one for audio, and one for video–as shown in Figure 1. In the existing literature, MSHMMs have been trained in one of two manners: Two single-stream HMMs can be trained independently and combined, or the entire MSHMM can be jointly-trained using both modalities. Because the combination method makes an incorrect assumption that the two HMMs were synchronous before combination, better performance can be obtained with the joint-training method [6].

FHMMs were introduced as an alternative to other multi-stream modelling techniques for audio-visual *speaker* recognition, designed to maximise the mutual information between the two modalities. As originally implemented FHMMs consisted of a continuous HMM for the dominant modality combined with a discrete vector-quantisation classifier for the subordinate modality within each state [4]. The subordinate classifiers were trained based on the forced-alignment of the dominant HMM on the training set. This original design was extended by the present authors in [5] to improve the modelling of

the subordinate modality by using a continuous classifier. This resulted in two continuous GMMs inside each state of the original dominant HMM, which can be seen to be identical to the multi-stream model shown in Figure 1(a). Therefore it can be concluded that rather than being an alternative model type, FHMMs can be regarded as an alternative way of training a regular MSHMM by adaptation from the dominant single-stream HMM rather than jointly-training on both modalities. The choice of the dominant modality for FHMM-adaptation should be based on the more reliable modality, which for speech processing will generally be the acoustic one [5].

3. Audio-visual speech recognition

3.1. Speech recognition task

Experiments were conducted to evaluate the speech recognition performance of our new FHMM-adaptation approach to MSHMM training and compared with other standard modelling techniques commonly used for AVSR. All models were tested for the task of small-vocabulary (digits only) continuous speech recognition on the XM2VTS dataset as outlined in Section 3.2. Speech recognition was performed on a simple word-loop with word-insertion penalties calculated for each system on the evaluation session. Speech recognition results were reported as a word-error-rate (WER) calculated by

$$\left(1 - \frac{H - I}{N}\right) \times 100\%$$

Where H is the number of correctly estimated words, I is the number of incorrectly inserted words, and N is the total number of actual words.

3.2. Training and testing datasets

Training, testing and evaluation data were extracted from the digit-video sections of the XM2VTS database [7]. The training and testing configurations used for these experiments was based on the XM2VTSDB protocol [8], but adapted for the task of speaker-independent speech recognition. 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. The first two sessions were used for training, the third for tuning/evaluation, and the final for testing. As per the XM2VTSDB protocol, 200 speakers were designated 'clients', and 95 were designated 'impostors'. Training of the speaker-independent models were performed on the 'client' speakers and tested on the 'imposters' to ensure that none of the test speakers were used in training the models.

The data in the testing sessions were also artificially corrupted with speech-babble noise in the audio modality at levels of 0, 6, 12 and 18 dB signal-to-noise ratio (SNR) to examine the effect of train/test mismatch on the experiments. Video degradation through decreasing the JPEG quality factor was also considered, but was found to have little effect on the final speech recognition performance and, as such, will not be reported in this paper.

3.3. Feature extraction

Perceptual linear prediction (PLP) based cepstral features were used to represent the acoustic features in these experiments. Each feature vector consisted of the first 13 PLPs including the zeroth, and the first and second time derivatives of those 13 features resulting in a 39 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.

Following the ROI extraction, the mean ROI over the utterance is removed. Our mean normalisation is similar to that of Potamianos et al [2], where the authors have used an approach called 'feature mean normalisation' for visual feature extraction which resembles the cepstral mean subtraction (CMS) method commonly used with audio features. However in our approach we perform normalisation in the image domain instead of the feature domain. A two-dimensional, separable, discrete cosine transform (DCT) is then applied to the resulting mean-removed ROI, with the 20 top DCT coefficients according to the zig-zag pattern retained, resulting in a 'static' visual feature vector. Subsequently, to incorporate dynamic speech information, 7 neighboring such features over ± 3 adjacent frames were concatenated, and were projected via an *inter*-frame linear discriminant analysis (LDA) cascade to 20 dimensional 'dynamic' visual feature vector. The delta and acceleration coefficients of this vector were then incorporated, resulting in a 60 dimensional visual feature vector.

3.4. Baseline systems

As well as the FHMM-adapted MSHMMs, four other baseline systems were also trained and tested for comparison:

- Audio-only HMMs
- Video-only HMMs
- Feature-fusion HMMs
- Jointly-trained MSHMMs

All baseline HMMs, including the MSHMMs, were trained with the HTK Toolkit [9] over the two training sessions. To ensure a fair comparison of the baselines systems with the FHMM-adapted system, all models were trained with the same topology of 13 states and 10 mixtures (for audio and/or video). This topology was determined empirically based on the best performing audio HMMs on the evaluation sessions. By keeping the topology the same for the individual models and the corresponding fusion models, we can more easily evaluate the ability of the fusion models to model each stream.

For the training and testing of the feature-fusion and MSHMM systems, the closest video feature vector was chosen for each audio feature vector and appended to create a single 99-dimensional feature-fusion vector. No interpolated estimation of the video features between frames was performed.

For the purposes of these experiments, stream weightings for the MSHMMs were defined to be α for the audio stream and $1 - \alpha$ for the video stream. From previous experiments, we had determined that the best performance from joint training of MSHMMs was obtained when $\alpha = 0.9$.

3.5. Fused HMM adaptation

Our FHMM method of adapting a MSHMM from an audio-only HMM is a two step process:

1. For each audio training observation, we find the best hidden-state alignment of the audio HMM by force-aligning the training transcriptions.
2. We next train additional video GMMs for each state based on the video observations lining up with the best hidden-state alignment in (1)

The FHMM-adapted MSHMM used in these experiments was based on the baseline acoustic HMM. Once the video observations that overlapped a particular state in the acoustic HMM were determined, a 10 mixture GMM was trained on those observations and the *video* GMM was added next to the state's already existing *acoustic* GMM. Once this had been performed for each state in each acoustic HMM, the result was a new set of MSHMMs of the same topology as the baseline jointly-trained MSHMMs. Because the video GMMs were trained separately to the audio models, the aligned video features used were not required to be up-sampled as they were for joint-training in a convential MSHMM.

3.6. MSHMM decoding

Decoding of the jointly-trained and FHMM-adapted MSHMM systems was performed in an identical manner. By keeping the decoding process of the two systems identical, the difference imparted by training with the FHMM-adaptation method versus joint-training can be examined.

As every audio observation needed to be matched with a corresponding video observation, the closest video feature vector was appended to each audio feature vector, in the same manner as during joint-training of the baseline MSHMMs in Section 3.4. These concatenated features were used to test the decoding of both MSHMM systems.

Because the distribution of scores varied considerably between the two stream-classifiers, zero-normalisation [10] was performed for each frame, within the MSHMM states, during decoding. The normalisation parameters were determined by performing the speech recognition task on the evaluation session with stream weight parameter, α, set such that only the modality of interest was being tested (ie. $\alpha = 0$ and $\alpha = 1$) and the scores of the best path were recorded on a frame-by-frame basis to determine the score-distribution of that modality. Once the estimated normalisation parameters (standard deviation, mean) $(\hat{\sigma}_a, \hat{\mu}_a)$ and $(\hat{\sigma}_v, \hat{\mu}_v)$ were determined for the audio and video respectively the final log-likelihood score s_f returned by a MSHMM state to the Viterbi decoding process is given by

$$s_f = \alpha Z_a(s_a) + (1-\alpha) Z_v(s_v)$$

Where s_a and s_v are the audio and video GMM log-likelihood scores respectively, and

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

is the zero-normalisation function for modality i.

This normalisation can be seen as essentially pre-weighting the two streams such that the variances and means are equal. Because speech recognition is a comparative task, a change in the mean log-likelihood scores on a stream-wide basis will have no effect on the ability to discriminate between individual word-models. Therefore, mean-normalisation is not strictly necessary but was kept for the sake of completeness. Once the stream models are normalised, the final α is a closer reflection of the true weights for each stream than an unnormalised α would be. This is because the unnormalised α would be performing two tasks: variance normalisation and stream weighting. For these experiments the post-normalisation α was chosen as 0.5 as it had the highest speech performance on the evaluation session.

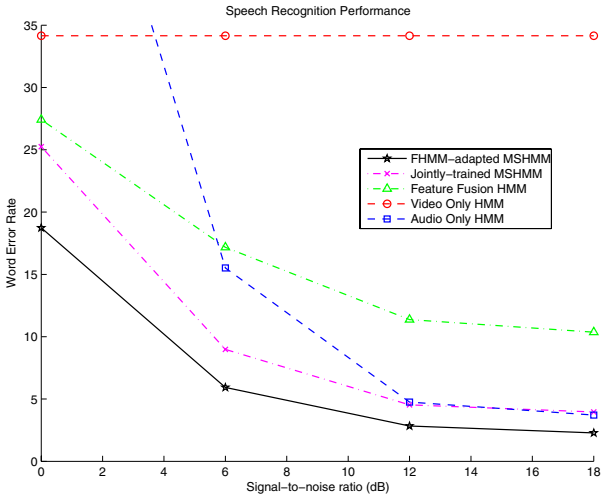


Figure 2: *Speech recognition performance for FHMM-adapted MSHMM and baseline systems over audio noise.*

4. Results and Discussion

The results of the speech recognition experiments on the testing session are shown in Figure 2. Of the three fused models tested, our FHMM-adapted MSHMMs performed best followed by the jointly-trained MSHMMs. The feature-fusion system provided the worse performance. The jointly-trained system outperforms the feature-fusion sytem for most of the chart because of its ability to consider (and normalise) each stream independently. However the FHMM-adapted system clearly provides the minimum WER at all noise levels, particular at equal levels of speech and noise (0 dB SNR), where the reduction in error is around 7% over the jointly-trained MSHMM.

To further examine the ability of the two differently-trained MSHMM systems in recognising speech over both modalities, the performance of each system in each modality was examined by adjusting the weights of the MSHMMs to the extremes of audio-only ($\alpha = 1$) and video-only recognition ($\alpha = 0$). This approach is essentially extracting the single-stream HMMs and evaluating their ability to recognise speech in the relevant modality. The results of these experiments are shown in Table 1.

These audio-only and video-only performance results suggest that the main improvement of the FHMM-adaptation is in the training of the video models. Because the audio models are used to directly determine the model boundaries of the video GMMs in the final FHMM-adapted MSHMMs, the video models are trained directly. This is in contrast to the conventional jointly-trained MSHMMs where the model boundaries and the models themselves are estimated together in a re-estimation process. By removing the uncertainty about the model boundaries, an improvement in performance is gained in the FHMM-adapted models.

In fact, by looking at the video-only performance of the

	Audio-only WER over noise			
Audio Model	0dB SNR	6dB	12dB	18dB
FHMM-adapted	63.98	15.41	5.16	3.82
Audio-only HMM	64.12	15.52	4.75	3.71
Jointly-trained	68.41	18.82	4.75	3.79

Video Model	Video-only WER
FHMM-adapted	30.80
Video-only HMM	34.15
Jointly-trained	40.63

Table 1: *Performance of audio and video HMMs extracted from MSHMMs.*

FHMM-adapted system, we can see that the FHMM-adaptation process can be used to generate a video-only HMM that can outperform the baseline video system. This shows that by using the audio models to form the state boundaries, we can create a video HMM that outperforms one trained on video only, even though no audio information is used in the decoding of the FHMM-adapted video HMM.

Because the FHMM-adaptation method consists of training the video GMMs and adding them directly on top of each audio HMM, the audio-only performance of a FHMM-adapted MSHMMs is, by definition, the same as that of the audio-only HMMs. The minor differences in Table 1 are due to small differences in the implementation of the two models.

The audio-only performance of the jointly-trained system does appear to be a little degraded compared to the baseline HMM in noisier conditions. We have not investigated why this is happening for this paper, but it may be that the joint-training method introduces some degradation into the audio models. As training was performed on clean data and with a known transcription, the ability of the audio models to determine the word boundaries correctly should be very good. However, under joint-training the expectation-maximization (EM) algorithm must operate on both modalities [9] and as video is not as good at reliably estimating state alignments it is unlikely to help with the model-boundary estimation portion of the re-estimation process, and may even have a negative effect.

5. Conclusion and future research

In this paper we have shown that the FHMM-adaptation method can improve the video speech modelling ability of MSHMMs over the joint-training method, resulting in MSHMMs that are better at speech recognition and more robust to acoustic noise. Using the audio feature's superior ability to align the model boundaries during training, a more accurate alignment of the boundaries of the video models is achieved in the FHMM-adaptation process. This results in MSHMMs that are much better at modelling the audio and, in particular, the video modalities than jointly trained MSHMMs. The FHMM-adapted models improve the WER over joint-training at all noise levels, but because the importance of the video modelling increases if the audio noise increases, the greatest reduction in WER is achieved in noisy conditions.

Our proposed method of using the audio-alignment to train video models could also be extended to more complex models such as product or asynchronous HMMs [3]. For example, the tie-points of a product or asynchronous HMM could be determined solely by the audio models during training. Our future work will investigate the performance gains that could be achieved by using such approaches.

Finally, although the FHMM-adaptation method allows one to use an audio HMM and video data to create a MSHMMs, there is no reason that the audio HMM need be trained on the same sequences as the video data. The FHMM-adaptation method would allow an audio-visual MSHMM to be 'kick-started' with a previously well-trained audio HMM, allowing a MSHMM to be trained on smaller datasets than would otherwise be required. This cross-dataset adaptation can take advantage of the ready availability of large audio datasets to generate a better MSHMM model than could be jointly-trained on the limited audio-visual data available.

6. 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 [7] or at `http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/`.

7. References

[1] S. M. Thomas and T. R. Jordan, "Contributions of oral and extraoral facial movement to visual and audiovisual speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 5, pp. 873–888, 2004.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[3] S. Bengio, "Multimodal speech processing using asynchronous hidden markov models," *Information Fusion*, vol. 5, no. 2, pp. 81–9, June 2004.

[4] 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.

[5] D. Dean, S. Sridharan, and T. Wark, "Audio-visual speaker verification using continuous fused HMMs," in *VisHCI 2006*, 2006.

[6] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Johns Hopkins University, CLSP, Tech. Rep. WS00AVSR, 2000. [Online]. Available: citeseer.ist.psu.edu/neti00audiovisual.html

[7] 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.

[8] J. Luettin and G. Maitre, "Evaluation protocol for the extended M2VTS database (XM2VTSDB)," IDIAP, Tech. Rep., 1998.

[9] 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.

[10] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.