INTER-FRAME CONTEXTUAL MODELLING FOR VISUAL SPEECH RECOGNITION

Adrian Pass   Ji Ming   Philip Hanna   Jianguo Zhang   Darryl Stewart

School of Electronics, Electrical Engineering and Computer Science
Queens University Belfast
Belfast BT7 1NN, UK
{apass01, j.ming, p.hanna, j.zhang, dw.stewart}@qub.ac.uk

ABSTRACT

In this paper, we present a new approach to visual speech recognition which improves contextual modelling by combining Inter-Frame Dependent and Hidden Markov Models. This approach captures contextual information in visual speech that may be lost using a Hidden Markov Model alone. We apply contextual modelling to a large speaker independent isolated digit recognition task, and compare our approach to two commonly adopted feature based techniques for incorporating speech dynamics. Results are presented from baseline feature based systems and the combined modelling technique. We illustrate that both of these techniques achieve similar levels of performance when used independently. However significant improvements in performance can be achieved through a combination of the two. In particular we report an improvement in excess of 17% relative Word Error Rate in comparison to our best baseline system.

***Index Terms*—** Contextual modelling, lipreading, speech dynamics, AVASR .

1. INTRODUCTION

Research and development of Automatic Speech Recognition (ASR) has been going on for some time, and modern audio only based speech recognition systems are sometimes capable of achieving near perfect levels of word recognition accuracy, given clean audio conditions. As a result ASR technology is now widely commercially available, with many home computers and even mobile phones being operable by voice command. It was shown by the work of Petajan [1] that by also combining visual information of the speakers lip movements with the audio, recognition rates can be improved further and ultimately the system can be made more robust to the effects of noise corruption in the audio stream. As such, visual speech recognition has become the focus of numerous research projects. Work includes investigations into modelling techniques and feature stream combination [2][3][4][5] feature extraction [6][7][8] and more recently head pose invariant lip-reading [9][10]. One of the factors limiting the recognition accuracy of visual only speech recognition is the small number of possible lip shapes/movements in relation to the range of corresponding vocal sounds. This is demonstrated through phoneme to viseme mapping; for example the phonemes /g/, /ŋ/, /k/ all appear to share the same corresponding viseme. Using a window based HCRF in a speaker dependent isolated digit recognition task in [5] we demonstrated that visual speech recognition performance can be improved by adopting a contextual approach to visual speech recognition. Due to excessive training times however, this technique was found to be impractical for a larger speaker independent task.

The standard approach to modelling speech is through the use of Hidden Markov Models (HMMs) [11] which use a hidden underlying state structure to model the temporal variation, with observations being generated at each state. Due to this generative modelling approach however, HMMs generally assume conditional independence between observations in order to make inference computationally tractable, and thus long range dependencies between observations are not directly modelled. Instead dynamic speech information is typically accounted for in the feature vectors themselves that make up the observations, through either feature derivatives or inter-frame concatenation. The former and more common approach [11] is achieved by combining the 'per frame' static feature vectors with their corresponding first and possibly second derivatives calculated over neighbouring frames, to yield a larger set of 'dynamic' features. The latter approach on the other hand involves the actual concatenation of the current frame's feature vector with neighbouring feature vectors, centred on the current frame. This in itself can result in excessively large vector sizes so the concatenation is usually followed by a dimension reduction step such as Principle Component Analysis (PCA) or Linear Discriminant Analysis (LDA) [12].

As an alternative to feature combination, a system was developed in [13] that instead combines models in an attempt to better capture dynamic information. More specifically a standard multiple-mixture HMM is combined with a segment based Inter-Frame Dependent model (IFD) resulting in an IFDHMM, which allows for effective modelling of the context of a particular point in an observation sequence based on preceding and succeeding frames. It was shown in [13] that for audio speech recognition using phonemes, this technique outperforms the standard HMM using dynamic features. We carry this and the work in [5] forward by applying the system to the task of isolated digit, *visual* speech recognition, and evaluate the performance against the standard feature based approaches.

2. IFDHMM THEORY

A general outline of the combined model system follows next;

The likelihood function of a conventional HMM can be expressed as;

$$p(o|\lambda) = \sum_{S} \pi_{S_0} \prod_{t} a_{S_{t-1}S_t} \cdot b_{S_t}(o_t) \tag{1}$$

The summation is over all state sequences S, whilst a represents the state transition matrix. The state dependent observation density b_{S_t} is a combination of all K Gaussian mixture components, summed as follows;

$$b_{S_i}^{hmm}(x) = \sum_{k=1}^{K} w_{ik} g_{ik}(x) \tag{2}$$

Fig. 1. Images from XM2VTS database showing sample variation

where $g_{ik}(x)$ is the kth Gaussian mixture component and w_{ik} the corresponding weight. The likelihood function of the combined IFDHMM then can similarly be expressed as;

$$p(o|\lambda) = \sum_S \pi_{S_0} \prod_t a_{S_{t-1}S_t} \prod_m b^m_{S_t}(o_t) \tag{3}$$

where b^m represents the observation density of the m^{th} component model, i.e. equivalent to the linear combination of the individual component model likelihoods in the logarithmic domain. The segment based model captures dynamic information by assuming a dependence between the current frame and pre-defined neighbouring frames such that the state dependent observation density can be written;

$$b^{ifd}_{S_i}(x|x_1.....x_N) = \sum_{n=1}^{N} c_{in} g_{in}(x|x_n) \tag{4}$$

It can be seen that this is treated similarly to the Gaussian mixture components in 2, except the Gaussian g_{in} now represents the conditional probability of observation x given x_n and N represents the number of neighbouring frames to be included. c_{in} is the corresponding weight where $\sum_n c_{in} = 1$. Succeeding and preceding frames are modelled by separate forward and backward IFD components, these being combined along with the HMM according to 3 to give the combined likelihood function;

$$\begin{aligned} p(o|\lambda) = \sum_S \pi_{S_0} \prod_t a_{S_{t-1}S_t} &\cdot b^{hmm}_{S_t}(o_t) \\ &\cdot b^{ifd-}_{S_t}(o_t|o_{t-\tau(1)}...o_{t-\tau(N)}) \\ &\cdot b^{ifd+}_{S_t}(o_t|o_{t+\tau(1)}...o_{t+\tau(N)}) \end{aligned} \tag{5}$$

By retaining the general HMM structure in this way, the combined IFDHMM allows for efficient training and decoding using the standard Baum Welch and Viterbi algorithms. For a more detailed explanation of the mathematical foundations the reader is directed to [13].

3. EXPERIMENTAL SETUP

Both the HMM and IFDHMM used in this work were implemented using in-house software.

3.1. Train/test data

For these experiments, video of speakers uttering the digits zero to nine from the XM2VTS [14] database was used, see Figure 1, recorded at 25 frames per second. All tests were speaker independent, making use of all available 295 speakers. This was split into 200 speakers for training and 95 speakers for testing, with 8 sessions per speaker. As each digit is uttered twice in a session this gave around 32000 isolated digits for training and 15200 for evaluation.

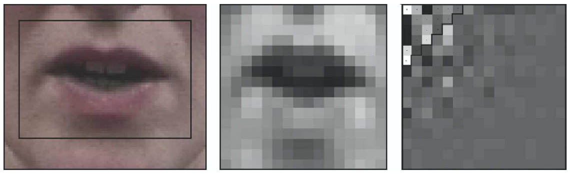

Fig. 2. Feature extraction from left to right: mouth ROI cropping, sub-sampling Y channel to 16 by 16 pixels, 2D DCT coefficients showing 5 by 5 triangular mask.

3.2. Static feature extraction

The XM2VTS database contains lip tracking results for each video, by means of tracking files describing a spline around the outer contour of the speaker's mouth. These coordinates can be used to extract a mouth ROI for each frame by calculating the centre point and the average widths and heights of the mouth over all frames of a given recording. We made use of video data containing pre-cropped mouth ROIs from previous work [15] which was created using the above method. The cropped video frames were firstly converted to grayscale by retaining only the Y channel (luminance) from YUV colour space, then sub-sampled to 16 by 16 pixels and standard Type-II 2D DCT coefficients for each frame were calculated. A triangular mask was used to extract the 15 highest energy / low frequency components per frame. Figure 2 illustrates the feature extraction process.

3.3. Dynamic features

In order to evaluate the IFDHMMs ability to capture dynamic visual speech information, dynamic features were also calculated from the static features above to use with a standard baseline HMM. Three groups of dynamic features were created in total. The first group was created by concatenating the static features with both their first derivatives and first and second derivatives, calculated using a cubic spline across each utterance. These shall be called *derivative* features. The second group was created by concatenating the feature vector at each time slice with neighbouring feature vectors using a sliding window, centred at the current time slice. This was performed using various window sizes. These shall be called *window* features. The final group was created identically to the window features, with the addition of a final dimension reduction PCA step, with a varying final feature vector size. These shall be called *PCA window* features.

3.4. Model topologies

Both the standard HMM and the HMM component of the IFDHMM (where used) are of standard left to right topology with one model being trained for each digit. Each model used five states with four Gaussian mixtures per state, this being found optimum through preliminary testing. The IFD components of the IFDHMM use single Gaussian mixtures to model each individual time slice.

4. RESULTS

4.1. Derivative, window and PCA window features

To assist with comparisons between the different techniques, we introduce the model complexity ***p*** in each case. This gives an indication of the number of *observation* parameters relative to a standard five state, four mixture HMM using an input feature vector length of 15 (mean and diagonal covariance, $5*4*15*2 = 600$ parameters).

Table 1. WER achieved using a standard HMM with various combinations of static, first and second derivative features

Features	Vector length	*p*	WER%
Static only	15	*1*	19.72
Static + Δ	30	*2*	11.91
Static + Δ + $\Delta\Delta$	45	*3*	**11.35**

Table 2. WER achieved using a standard HMM with windowed features of various window sizes

Window size	Vector length	*p*	WER%
1	45	*3*	18.33
2	75	*5*	17.81
3	105	*7*	16.54
4	135	*9*	15.65
5	165	*11*	15.11
6	195	*13*	14.57
7	225	*15*	14.02
8	255	*17*	13.39
9	285	*19*	13.10
10	315	*21*	12.66
11	345	*23*	**12.39**
12	375	*25*	13.39

Table 3. WER achieved using a standard HMM with PCA windowed features of various feature vector length per time slice

Vector length	*p*	WER%
15	*1*	21.58
45	*3*	13.75
75	*5*	11.47
105	*7*	11.14
135	*9*	**11.09**
165	*12*	12.00

To set a baseline for comparison, tables 1 and 2 show the word error rates (WER) achieved using a standard HMM with both derivative and window features respectively. The window size parameter in table 2 refers to the number of preceding *and* succeeding feature vectors concatenated onto the current one, i.e. a window size of 2 corresponds to the feature vectors from the range t-2 to t+2, giving 5 time slices centred at the current. As would be expected, table 1 illustrates the significance of dynamic information for visual speech recognition, even when using only static and first order derivative features. This is also true when dynamic information is derived from the concatenation of neighbouring feature vectors as evidenced by the results in table 2, albeit to a lesser extent. It should be noted at this point that the average frame length for a digit is around 11 frames. This may explain the optimum window size of 11 which ensures that the entire utterance is accounted for at any given frame, and thus increasing the window size further incorporates additional redundant information.

By examining the relative model complexities it can be seen that the window features require some 7-8 times the number of parameters to achieve optimum results than the derivative features. It is clear in this case at least that the use of first and first derivative features is a considerably more efficient method of incorporating dynamic information than the use of a window.

Table 3 shows the WERs achieved using window features that have undergone a PCA dimension reduction step. Only the window size of 11 was considered here as this yielded the best results of the window feature results in table 2. PCA was performed several times to yield a range of feature vector lengths (per time slice) and it can be seen from table 3 that a vector length of 45 or less yields a WER greater than that of the original features. The optimum vector length of 135 yielded a WER that was 1.3% lower than the original features, highlighting the detrimental impact of the redundant information removed by the PCA step. Despite the improvements reaped by removing this redundancy, it would still appear that the derivative features provide a more efficient means of incorporating speech dynamics. A like for like comparison bewteen ***p* = *1*** and ***p* = *3*** shows the derivative features to yield a 1.8% improvement in WER for each case. Nonetheless the best baseline result was achieved using PCA window features (table 3), giving a WER of 11.09%.

4.2. IFDHMM results

Although it is possible to implement the IFDHMM using any combination of preceding and/or succeeding conditional frames, it was generally found that the greatest reductions in WER were achieved using neighbouring consecutive conditional frames. As such, the parameter *N* used in the results here represents the number of neighbouring consecutive frames to the current frame being modelled. For example, *N=3* equates to the first 3 frames in a given direction from the current frame, where IFD$^-$ denotes preceding frames and IFD$^+$ denotes successive frames.

Table 4 gives the WERs for the IFDHMM using static only features. As with the window based features there is an obvious trend toward a lower WER as the length of the neighbouring segment(s) included is increased, highlighting the benefit to be gained from contextual modelling of visual speech. As before the average frame length of 11 provides an upper limit to the useful length of neighbouring segments, which may be responsible for the fluctuations in WER for larger values of *N*. The results also indicate that both preceding and succeeding frames are of approximately equal importance in providing contextual visual speech information, whilst the best results are achieved by combining both the forwards and backwards IFD and HMM models. It is to be noted however that the lowest WER of 11.68% achieved using the static only features is slightly higher than the benchmark WER using PCA window features. For a final investigation the IFDHMM tests were repeated, this time using the derivative features as input. These were chosen in favour of the PCA window features to keep model complexity to a minimum, as it was shown in the previous section that derivative features were the most efficient for a given feature vector length. Again the combination of both the forwards and backwards IFD with HMM models was found to yield the best results, so only these are shown in table 5. By combining derivative based features with contextual modelling, it is possible to reduce the WER by almost a further 2% over the benchmark result given by the PCA window features (table 3), a 17.04% reduction *relative* to the benchmark. It can be seen however that even using a value of *N=2* performance is improved over the other techniques alone whilst retaining a relatively low model complexity (***p* = *6***). These results confirm that the contextual information about a particular observation gained from the segmental conditional probabilities, does indeed go some way toward accounting for the temporal qualities of visual speech missed by standard HMM based

Table 4. WER achieved using various configurations of IFDHMM for a range of consecutive frame lengths with static features

N	HMM IFD$^+$ IFD$^-$		HMM IFD$^+$		HMM IFD$^-$	
	p	WER%	***p***	WER%	***p***	WER%
1	*1.5*	14.34	*1.25*	14.80	*1.25*	14.41
2	*2*	12.90	*1.5*	13.90	*1.5*	13.94
3	*2.5*	12.62	*1.75*	13.61	*1.75*	13.64
4	*3*	12.27	*2*	13.49	*2*	13.49
5	*3.5*	12.02	*2.25*	13.26	*2.25*	13.24
6	*4*	11.95	*2.5*	13.38	*2.5*	13.29
7	*4.5*	11.92	*2.75*	13.32	*2.75*	13.17
8	*5*	11.87	*3*	13.25	*3*	13.08
9	*5.5*	11.76	*3.25*	13.10	*3.25*	**12.93**
10	*6*	11.78	*3.5*	12.96	*3.5*	13.12
11	*6.5*	**11.68**	*3.75*	**12.95**	*3.75*	13.06

Table 5. WER achieved using a forwards-backwards IFDHMM for a range of consecutive frame lengths with static + 1st + 2nd order derivative features

N	***p***	WER%
2	*6*	10.90
3	*7.5*	10.39
4	*9*	9.95
5	*10.5*	9.74
6	*12*	9.47
7	*13.5*	9.45
8	*15*	9.33
9	*16.5*	9.34
10	*18*	**9.20**
11	*19.5*	9.33

techniques. They also highlight however that contextual modelling alone doesn't capture all the available temporal information. The greatest reduction in WER can be gained by combining contextual information with dynamic information such as velocity and acceleration as provided by the derivative based features used here. It is to be noted at this point that as these experiments are based around digit recognition, each speech unit spans several video frames, justifying the use of such long dependencies. For smaller units such as visemes it may be more appropriate to use smaller values of *N*.

5. CONCLUSIONS

In this paper, we have presented an IFDHMM modelling approach to visual speech recognition. This approach has been shown to improve upon the performance of a standard HMM alone by incorporating contextual speech information from a range neighbouring frames, thus overcoming the assumption of conditional independence. By comparing this technique to standard dynamic feature based approaches we have demonstrated that either method in itself captures unique additional contextual information within visual speech. More importantly we have shown that by combining this additional information from both approaches, a 17.04% reduction in WER relative to our best baseline system can be achieved.

6. ACKNOWLEDGEMENTS

This work was supported by DEL and the UK EPSRC under Grant EP/E028640/1 ISIS.

7. REFERENCES

[1] E. D. Petajan, *Automatic Lip-reading to Enhance Speech Recognition*, Ph.D. thesis, University of Illinois, 1984.

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

[3] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP J.Appl.Signal Process.*, , no. 1, pp. 1274–1288, 2002.

[4] R. Seymore, D. Stewart, and J. Ming, "Audio-visual integration for robust speech recognition using maximum weighted stream posteriors," in *Interspeech*, 2007, pp. 654–657.

[5] A. Pass, J. Zhang, and D. Stewart, "Hidden conditional random fields for visual speech recognition," in *International Machine Vision and Image Processing Conference*, 2009, pp. 117–122.

[6] Alin G. Chictu, Leon J. M. Rothkrantz, Pascal Wiggers, and Jacek C. Wojdel, "Comparison between different feature extraction techniques for audio-visual speech recognition," *Journal on Multimodal User Interfaces*, vol. 1, no. 1, pp. 7–20, March 2007.

[7] Gerasimos Potamianos, Juergen Luettin, and Chalapathy Neti, "Hierarchical discriminant features for audio-visual lvcsr," in *Proc. ICASSP*, 2001, pp. 165–168.

[8] Rowan Seymour, Darryl Stewart, and Ji Ming, "Comparison of image transform-based features for visual speech recognition in clean and corrupted videos," *EURASIP J.Image Video Process.*, pp. 1–9, 2008.

[9] P. Lucey, G. Potamianos, and S. Sridharan, "A unified approach to multi-pose audio-visual asr," in *Interspeech*, 2007, pp. 650–653.

[10] P. Lucey, S. Sridharan, and D. Dean, "Continuous pose invariant lipreading," in *Conference of the International Speech Communication Association*, 2008, pp. 2679–2682.

[11] Lawrence R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc IEEE*, vol. 77, pp. 267–296, 1990.

[12] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-visual Speech Processing*. 2004, MIT Press.

[13] J. Ming, P. Hanna, D. Stewart, S. Vaseghi, and F. J. Smith, "Capturing discriminative information using multiple modeling techniques," in *International Conference on Spoken Language Processing*, Dec, 1998, pp. 2791–2794.

[14] K. Messer, J. Matas, J. Kittler, J. Lttin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *In Second International Conference on Audio and Video-based Biometric Person Authentication*, 1999, pp. 72–77.

[15] R. Seymour, *Audio-Visual Speech and Speaker Recognition*, Ph.D. thesis, Queens University Belfast, 2008.