AN INVESTIGATION INTO FEATURES FOR MULTI-VIEW LIPREADING

Adrian Pass   Jianguo Zhang   Darryl Stewart

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

ABSTRACT

For the first time in this paper we present results showing the effect of speaker head pose angle on automatic lip-reading performance over a wide range of closely spaced angles. We analyse the effect head pose has upon the features themselves and show that by selecting coefficients with minimum variance w.r.t. pose angle, recognition performance can be improved when train-test pose angles differ. Experiments are conducted using the initial phase of a unique multi view Audio-Visual database designed specifically for research and development of pose-invariant lip-reading systems. We firstly show that it is the higher order horizontal spatial frequency components that become most detrimental as the pose deviates. Secondly we assess the performance of different feature selection masks across a range of pose angles including a new mask based on Minimum Cross-Pose Variance coefficients. We report a relative improvement of 50% in Word Error Rate when using our selection mask over a common energy based selection during profile view lip-reading.

***Index Terms*—** AVASR, pose invariance, feature extraction, discrete cosine transform

1. INTRODUCTION

Research and development of Audio-Visual Automatic Speech Recognition (AVASR) systems has come a long way since the pioneering work of Petajan [1], however there are still a few significant challenges to overcome before such systems can be practically realisable. The addition of a visual modality can improve robustness to the effects of noise corruption in the audio channel and indeed help to visually disambiguate confusable phonemes such as /s/ and /f/ or /b/ and /d/. State of the art AVASR achieves visual only Word Error Rates (WERs) as low as 15-20% [2] in isolated unit recognition tasks and ideal conditions. However the visual modality also brings with it a much greater scope for a train-test mismatch than the audio, through factors such as poor mouth ROI localisation, local and global changes in illumination and variations in head pose which can each significantly degrade performance [3]. It is the latter of these problems that is considered in this paper.

The first known use of non-frontal video data for AVASR can be found in [4][5]. In this work the authors combine audio speech features with visual features extracted from profile view lip images, demonstrating that useful speech information may still be gained from non-frontal visual features. In separate works [6] and [7], profile based AVASR systems are compared to frontal based systems with conflicting results as to the superior. In the former it is the frontal view that yields the best performance; the authors use appearance based DCT features for both views in the visual front end, with the profile mouth ROI containing background information from the room itself. In the second paper, the authors make use of simple geometric visual features, with the profile features benefiting from an additional dimension over the frontal. Coupled with the lack of destructive background information in the profile mouth ROI, it is the profile based features that yield superior results. Interestingly both works report that the best performance is achieved when combining feature streams from both views, indicating that there may exist useful speech information unique to each view. A similar comparison is made in [8] between frontal and 45 degree video data, again showing the frontal based system to yield the best performance.

Fewer works exist on tackling the problem of pose invariant AVASR. Perhaps the most practical contribution can be found in [9][10], where the authors adopt a viewpoint transform approach in the feature domain to project features from one viewpoint into those of another. This allows for a model to be trained using a single viewpoint, such that features obtained from an alternative viewpoint may be mapped onto the ‘correct’ feature space. The results demonstrate this to be a potentially viable approach although it firstly requires the estimation of pose [11] in order to select the appropriate transform.

As there is such a vast quantity of frontal view speech data available for training a real world system, it is desirable to develop a lip-reading system that can operate across a range of pose angles despite training on a single view. The work in this paper focusses on the development of such systems with a view to omitting the pose estimation step, by using a Minimum Cross-Pose Variance (MCPV) analysis technique to highlight the DCT feature components most robust to changes in head pose. We choose DCT as it provides an efficient feature representation that has been shown to outperform other common visual speech features [12][13]. Using visual only, speaker dependent, isolated digit recognition tasks we present baseline results demonstrating the decline in performance as the pose angle deviates using a common energy based feature selection method. We then compare the performance of an alternative selection method based on MCPV coefficients to this baseline and also a third selection method from the literature [14], shown to be robust to small changes in head pose/rotation. A significant factor behind the limited research in this area lies in the lack of suitable multi-pose AVASR databases available, therefore another contribution of this work is in the creation of such a database named QuLips. In section 2 we describe MCPV feature analysis, followed by methodology and database details in section 3 and results in section 4.

2. MINIMUM CROSS-POSE VARIANCE ANALYSIS

Consider two simultaneously recorded video streams of a speaker V_1 and V_2, captured from two different horizontal viewing angles. Each

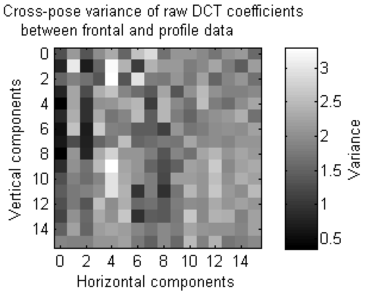

Fig. 1. Cross pose variance of raw 16x16 DCT coefficients using simultaneously recorded frontal and profile visual speech data.

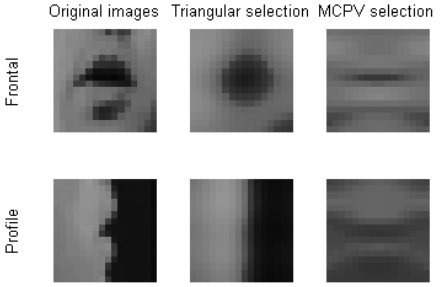

Fig. 2. Left; Simultaneous frontal and profile sub-sampled mouth ROI images. Middle; Reconstructed from DCT using 5*5 high energy triangular selection (15 coeffs). Right; Reconstructed using 15 best MCPV DCT coefficients.

of these video streams has identical resolution $M * N$ pixels and T frames. Raw 2D DCT features are then calculated for each frame and for each video stream providing two raw feature streams F_1 and F_2, each with $M * N$ components per frame and T frames. Assuming that individual DCT components have been mean and variance normalised as per section 3.3, a 'difference' feature stream F_{1-2} is then calculated by subtracting corresponding DCT components from one another across F_1 and F_2 for each time frame t;

$$F_{1-2}(m,n,t) = F_1(m,n,t) - F_2(m,n,t) \quad \forall m,n,t \qquad (1)$$

The cross-pose variance C_{1-2} of each DCT coefficient is then the variance of each corresponding element of F_{1-2} w.r.t. time t.

$$C_{1-2}(m,n) = \sum_{t=1}^{T} \frac{1}{T}(F_{1-2}(m,n,t) - \mu_{F_{1-2}}(m,n))^2 \quad \forall m,n \qquad (2)$$

It is then assumed that MCPV feature components are those components of C_{1-2} with the lowest values.

Figure 1 shows the result of the MCPV feature analysis on our database (detailed in section 3.1) between frontal and profile pose angles (0 and 90 degrees) after performing feature extraction as per section 3.3. It is the darkest areas that represent MCPV features and it can be seen that *it is predominantly columns zero and two of the raw DCT that are most robust to changes in pose, corresponding to high levels of detail in the vertical direction and very little detail in the horizontal.* The emphasis on even columns also indicates the importance of symmetry across the vertical plane [14]. This is intuitive in that it is the appearance of the mouth in the horizontal direction that distorts the most between profile and frontal views (see figure 4), and that whilst a frontal view mouth could be considered approximately symmetrical, the profile view is very much non-symmetrical. The implication of this analysis is illustrated in figure 2 which shows both frontal and profile sub-sampled mouth ROIs along with the same images reconstructed from 15 raw DCT coefficients using a common energy based triangular selection, and the 15 best MPVC coefficients (14 + DC to allow image reconstruction) from figure 1 respectively. While the images constructed using the triangular selection still clearly show the speakers pose, the images using the low 'cross pose' variance coefficients appear to have been normalised w.r.t. pose, resulting from a combination of the forced vertical symmetry and the smearing of horizontal detail. This result could be considered a generalisation of the findings in [14] which ultimately show that by retaining only the even columns from an energy based selection, vertical symmetry is forced in the spatial frequency domain and thus the mouth normalised to small changes in rotation or pose.

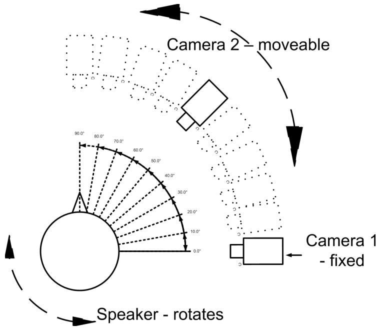

Fig. 3. Plan view of recording setup showing fixed position of camera 1, movement of camera 2 and rotation of subject. Example shows simultaneous angles of 0 and 50 degrees

3. METHODOLOGY

We now present experiments conducted to test the validity of the MPVC analysis performed in the previous section, followed by results and discussions in section 4. Pose invariant and indeed multi-pose AVASR are still relatively novel areas of research and as such there are only a handful of databases relevant to the problem, namely; HIT-AVDB-II [15], AVICAR [16], CMU AVPFV [7], data from the IBM smart rooms in [6][9] and CUAVE [17]. However none of these datasets contain a sufficient number of simultaneously captured pose angles as required for our work. As such this work was built from the ground up through the ongoing collection of a new multi-view AVASR database named QuLips. By capturing video speech data from a wide range of discrete, closely spaced and measured angles about the speakers head the resulting dataset allows for a more controlled simulation of continuous head pose. Hence we begin this section with details of data collection and preparation, followed by feature extraction and details of the experimental setup. The reader is encouraged to contact the authors regarding distribution of this dataset.

3.1. Multi-View AVASR database collection

For the initial phase of data acquisition two cameras were used and two subjects recorded. Video was captured at a rate of 25fps and a resolution of 720x576px. Audio was also captured using the internal microphones of each camera. Figure 3 shows a plan view of the setup. The floor area out from the speaker to the cameras was visually divided up into ten degree increments between zero and ninety degrees inclusive. Between recordings camera 1 was fixed at zero degrees whereas camera 2 was allowed to move around to the differ-

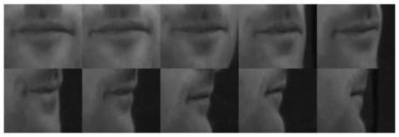

Fig. 4. Sample from QuLips database showing pose angles

ent angles. The subject was also rotated to each angle, thus allowing any pair of angles to be simultaneously recorded. The room itself was chosen as it contains no windows and consistent illumination. A blue background was used behind the speaker.

As per the XM2VTS database [18] utterances are made up of the pair of digit strings '0123456789' and '5069281374'. Considerable effort was put into recording organisation, such that the pair of digit strings is recorded from every angle and that every angle shares a simultaneous recording with every other angle. The resulting dataset allows for controlled comparisons between angles despite using only two cameras. 180 digits are available for each of the 10 angles and each speaker, giving a total of 3600 digits.

3.2. Data preparation

After data collection, mouth ROI cropping was performed via a semi-automated process. Facial feature tracking points and a mouth bounding box were manually defined in the initial frame of each video, followed by feature tracking using image correlation. The mouth ROI position was tracked based on the movement of the other features. Figure 4 shows a sample of cropped data for all angles. As per previous work [12], audio Hidden Markov Models (HMMs) were trained for each individual digit using TIDGITS [19] audio data and the Hidden Markov Toolkit (HTK) [20], enabling forced alignment to be performed on the audio from our data to obtain audio frame boundaries for each digit. For simplicity these boundaries are assumed to be common to both audio and video.

3.3. Feature extraction

Visual feature extraction follows a standard approach which has been shown to be state of the art [12]. Firstly each video frame was subsampled to 16x16 pixels and converted to gray-scale, followed by mean subtraction of the ROI in the image domain as per [10]. This was found to improve recognition accuracies over mean subtraction in the feature domain. A 2D DCT was then applied to each frame and an appropriate coefficient selection mask applied to obtain the per-frame static feature vector. First order derivative features were then calculated and concatenated onto the static vectors followed by variance normalisation across each utterance.

Three selection masks are considered in this paper (see figure5). The first of these is a standard energy based mask [12][13] intended as the baseline, obtained using a triangular zig-zag pattern from the top left corner. The second mask is made up of only even columns from the energy based triangular selection similar to [14], which was shown to force vertical symmetry and thus normalise for small changes in pose/rotation. The final mask is based on the MCPV coefficients of section 2, for simplicity it is approximated using the first two even columns of a triangular selection. These masks are respectively denoted 'Tri', 'Even' and 'MCPVa' throughout, all selection mask sizes are quoted in terms of static features.

3.4. Experiments

The HMMs used throughout are of a similar setup to [12]. All are of standard left-right topology with one model being trained for each digit using only the frontal data, 10 classes in total. Each model used 4 states with 2 Gaussian mixtures per state, this being found optimal during preliminary testing.

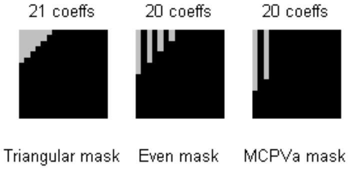

Fig. 5. Left to right; 21 coefficient 'Tri' mask, 20 coefficient 'Even' mask and 20 coefficient 'MCPVa' mask

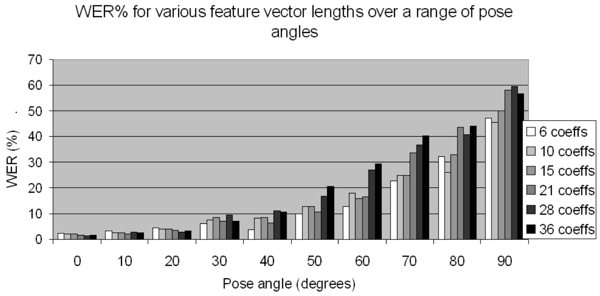

Fig. 6. Effect of head pose and feature vector length on WER% using a frontal view trained model. Triangular selection.

Two sets of visual only, speaker dependent isolated digit experiments were conducted. Isolated digits were chosen as they are relatively easy to discriminate between; this work represents an initial investigation into the effects of head pose on visual features. The first set provides a baseline result that shows how recognition performance varies across the full range of pose angles using a 'Tri' mask of varying sizes. The second compares the performance of all three feature selection masks given in section 3.3. In keeping with speaker dependent testing all results are obtained for each speaker and then averaged. For frontal only results testing was performed using the frontal data in a 9-fold cross validated fashion for each speaker. For testing on remaining angles *all* the frontal data was used for training, with testing performed using all the data for each remaining angle individually.

4. EXPERIMENTAL RESULTS

4.1. Baseline results

Figure 6 shows how the WER varies for the 'Tri' mask over a range of feature vector lengths as the pose used during testing deviates from the frontal view. The results firstly show an expected drop in performance as the train-test pose mismatch is increased. Secondly they show that a smaller feature vector length becomes preferable as the pose angle deviates further from that encountered during training, i.e. the higher frequency components appear to become detrimental. This is particularly noticeable in the 30-70 degree range where it is the smallest feature vector that yields the best performance, in contrast to the 0-20 degree range where it is the larger feature vectors.

4.2. Comparison of feature selection masks

To test the validity of our MCPV analysis in section 2, figure 7 shows the performance of all three feature selection masks given in section 3.3 over varying feature vector lengths for frontal, 30, 60 and 90 degree pose angles. Note that no account is made here for the redundancy or indeed information content of coefficients.

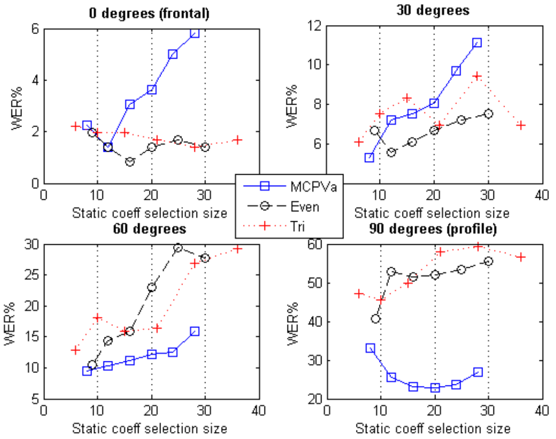

Fig. 7. Comparison of 'Tri', 'Even' and 'MCPVa' feature selection masks for varying feature vector lengths. WER% results for frontal, 30 and 60 degrees and profile view recognition, frontal trained.

In line with [14], figure 7 shows the 'Even' mask to yield the best performance for the frontal pose. For larger pose angles however forced symmetry alone is no longer sufficient to correct for pose due to the increased distortion of the mouth ROI in the horizontal direction. As such it is the 'MCPVa' mask that consistently yields the lowest WERs for pose angles of 60 and 90 degrees, i.e. *the higher order horizontal components from the 'Even' mask become detrimental to recognition when the pose angle deviates far enough from the trained frontal view*. This is evidenced by directly comparing performances of the 'Even' and 'MCPVa' masks. The smallest 'MCPVa' and 'Even' masks of 8 and 9 coefficients respectively are in fact identical selections, with the exception of one additional higher order *horizontal* component in the latter. However in the 90 degree (profile) plot this additional higher order horizontal component actually increases the WER from 33.06% to 40.56%, a relative increase of 22.69%. It is also interesting to note from figure 7 that the lowest profile view WER is achieved using a 20 coefficient 'MCPVa' mask containing the first 11 vertical spatial frequency components. This indicates that *additional higher order vertical frequency components may become preferable to horizontal components as the pose angle tends towards profile*. The lowest profile WER of 22.78% achieved using the 20 coefficient 'MCPVa' mask shows a 50% relative improvement to the lowest profile WER of 45.56% achieved using the energy based 'Tri' mask of 10 coefficients.

Given these initial results it would seem that there is no optimum selection for all pose angles. Thus in order to implement a continuous pose invariant lip-reading system trained on a single pose, it may be appropriate to adopt some form of dynamic coefficient selection during recognition itself.

5. CONCLUSIONS AND FURTHER WORK

In this paper we have presented results showing the effect of head pose on automatic lip-reading performance across a wide range of angles. We have adopted a Minimum Cross-Pose Variance analysis technique for feature component selection and shown that as the speakers head pose deviates from the frontal pose used during training, the higher order horizontal spatial frequency components become increasingly detrimental. We have also shown that higher order vertical components become preferential to horizontal components as the pose tends towards profile. Using a coefficient selection approximated from this technique we report a 50% relative WER reduction over a common energy based selection method when using a frontal trained model to read profile view lips.

Using the second phase of data collection we plan to extend this work by removing redundancy within the Minimum Cross-Pose Variance feature components, evaluating over a wider range of pose angles and investigating dynamic feature component selection.

6. REFERENCES

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

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

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-visual Speech Processing*. 2004, MIT Press.

[3] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," in *Proc. EUROSPEECH*, 2003, pp. 1293–1296.

[4] Tomoaki Yoshinaga, Satoshi Tamura, Koji Iwano, and Sadaoki Furui, "Audio-visual speech recognition using lip movement extracted from side-face images," in *PROC. AVSP*, 2003, pp. 117–120.

[5] Tomoaki Yoshinaga, Satoshi Tamura, Koji Iwano, and Sadaoki Furui, "Audio-visual speech recognition using new lip features extracted from side-face images," in *Proc. ROBUST*, 2004.

[6] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," in *Multimedia Signal Processing, IEEE 8th Workshop on*, 2006, pp. 24–28.

[7] K. Kumar, T. Chen, and R. M. Stern, "Profile view lip reading," in *Proc. ICASSP 2007*, 2007, vol. 4, pp. IV–429–IV–432.

[8] K. Kumatani and R. Stiefelhagen, "State synchronous modeling on phone boundary for audio visual speech recognition and application to muti-view face images," in *Proc. ICASSP*, 2007, vol. 4, p. IV417:IV420.

[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, G. Potamianos, and S. Sridharan, "Visual speech recognition across multiple views," *Visual Speech Recognition; Lip Segmentation and Mapping*, 2009.

[11] P. Lucey and S. Sridharan, "A visual front-end for a continuous pose-invariant lipreading system," in *Proc. ICSPCS*, 2008, pp. 1–6.

[12] R. Seymour, D. Stewart, and J. Ming, "Comparison of image transform based features for visual speech recognition in clean and corrupted videos," *EURASIP - Image and Video Processing*, vol. 2008, pp. 1–9.

[13] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual lvcsr," in *ICME 2001*, pp. 825–828.

[14] G. Potamianos P. Scanlon, "Exploiting lower face symmetry in appearance based automatic speechreading," in *AVSP*, 2005, pp. 79–84.

[15] X. Hong X. Lin, H. Yao and Q. Wang, "HIT-AVDB-II: A New Multiview and Extreme Feature Cases Contained Audio-Visual Database for Biometrics," in *Proc. CVPRIP*, Dec 15-20 2008.

[16] B. Lee, M. Hasegawa-johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. Conf. Spoken Language*, 2004, pp. 2489–2492.

[17] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *In Proc. ICASSP*, 2002, pp. 2017–2020.

[18] K. Messer, J. Matas, J. Kittler, J. Luttin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," in *2nd Int. Conf. Audio and Video-based Biometric Person Authentication*, 1999, pp. 72–77.

[19] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP.*, 1984, vol. 9, pp. 328–331.

[20] J. Odell, D. Ollason, P. Woodland, S. Young, and J. Jansen, *The HTK Book for HTK V2.0*, Cambridge University Press, UK, 1995.