# Feature Selection for Pose Invariant Lip Biometrics

*Adrian Pass, Jianguo Zhang, Darryl Stewart*

School of Electronics, Electrical Engineering and Computer Science
Queens University Belfast
Belfast BT7 1NN, UK

{apass01, jianguo.zhang, d.w.stewart}@qub.ac.uk

## Abstract

For the first time in this paper we present results showing the effect of out of plane speaker head pose variation on a lip biometric based speaker verification system. Using appearance DCT based features, we adopt a Mutual Information analysis technique to highlight the class discriminant DCT components most robust to changes in out of plane pose. Experiments are conducted using the initial phase of a new multi view Audio-Visual database designed for research and development of pose-invariant speech and speaker recognition. We show that verification performance can be improved by substituting higher order horizontal DCT components for vertical, particularly in the case of a train/test pose angle mismatch. We show that the best performance can be achieved by combining this alternative feature selection with multi view training, reporting a relative 45% Equal Error Rate reduction over a common energy based selection.

**Index Terms**: Lip biometrics, speaker verification, pose invariance, mutual information, discrete cosine transform.

## 1. Introduction

The audio produced during speech can be considered to contain two broad and distinct categories of information. The first and most obvious is in the message itself, i.e. *what* is being said. The second is very much speaker dependent and thus may be used to help verify a speakers identity, i.e. *who* is saying it. It has been known for some time [1] that visual cues from a speakers lip movements may be combined with the audio in Audio Visual Automatic Speech Recognition (AVASR) systems to improve performance where there may be noise corruption in the audio. Work in recent years has also shown that this visual modality also contains speaker specific information, thus making it suitable as an additional modality in speaker verification [2, 3, 4]. This allows a verification system to be more robust to both audio corruption and indeed false identity claims from impostors, due to the increased difficulty of impersonating an additional dynamic biometric. The visual modality however can also be prone to corruption unique to the domain such as local or global changes in illumination, poor mouth ROI localization and variations in out of plane head pose which can each significantly degrade lip-reading performance [5]. It is the latter of these problems that is considered here.

There are a number of works available which investigate the use of non-frontal video for AVASR, namely profile view [6, 7, 8, 9] and 45 degree [10] data. While each show that useful speech information may be obtained from alternative viewpoints, the latter works also demonstrate this information to be complementary to that obtained from the frontal view. The number of works tackling the problem of pose invariant AVASR are fewer still, where perhaps the most practical contribution can be found in [11]. In this work the authors propose a viewpoint transform approach following a pose estimation step [12], allowing a single view model to operate across multiple views.

In this paper we examine the problem of out of plane variations in pose angle on a lip biometric *speaker* verification system, such as a doorway access control system. In such a scenario it is reasonable to assume the speaker to be fairly cooperative, i.e. holding a steady, approximately frontal pose towards the camera. It is therefore only necessary to ensure such a system is robust to a small range of *out* of plane pose angles, for example if the speaker isn't fully aware of the precise location of the camera. Unlike AVASR however where there may only be frontal view video data available for training, a speaker verification system requires an enrolment procedure for each subject, where non-frontal viewpoints may be incorporated into training. We use a Mutual Information (MI) analysis technique, similar to that used in [13] for frontal view visual *speech* recognition, to highlight the DCT coefficients with the highest information content w.r.t. *speaker identity* across a range of horizontal viewing angles. We use appearance based DCT features as they provide a simple and compact feature representation that have been shown to outperform other feature types when lip reading for *speech* recognition [14, 15]. In [16] the authors approximate the MI based selection of [13] for visual *speech* recognition by retaining only the even columns of an energy based selection. They show that by forcing vertical symmetry in the feature domain in this way it is possible to correct for small variations of *in* plane variations in pose. As we show further on, the properties of the DCT also lend themselves well to correction for *out* of plane pose angle. Using visual only speaker verification experiments based on isolated digits we present baseline results showing the effect of pose angle during training and testing on Equal Error Rate (EER) when using a common energy based DCT selection technique. We then compare these results to those obtained using an alternative feature selection based on the cross pose MI analysis, as well as a third selection similar to that used in [16]. As part of ongoing research we are collecting a new multi pose Audio-Visual Speech database named QuLips which has been used in this work. In section 2 we describe the cross pose MI feature analysis, followed by methodology and database details in section 3 and results in section 4.

## 2. Mutual Information

The formal definition of MI for discrete random variable $X$ and the corresponding class labels $C$ may be expressed as;

$$I(X;C) = \sum_{x \in X} \sum_{c \in C} p(x,c) log \left( \frac{p(x,c)}{p(x)p(c)} \right) \qquad (1)$$

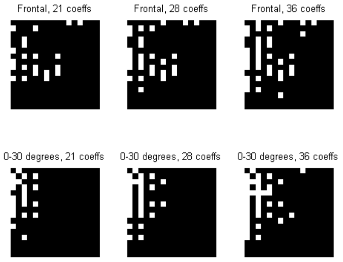

Figure 1: DCT selection masks based on top scoring coefficients from MI analysis using frontal only and 0-30 degree video data

It can be seen that if $x$ and $c$ are random w.r.t. one another then the joint probability $p(x, c)$ will be equal to the product of the individual marginal probabilities $p(x)$ and $p(c)$, such that the right hand part of equation (1) becomes zero. In order to estimate the probabilities in (1) we adopt a similar histogram based approach to [13] which requires quantisation of the discrete variable $X$ into a number of equally spaced bins. The number of bins for non-Gaussian data can be estimated using Doane's rule [17];

$$B = log_2 T + 1 + log_2(1 + \hat{k} + \sqrt{T/6}) \qquad (2)$$

where $T$ is the number of samples and $\hat{k}$ the estimated kurtosis of variable $X$.

Using equations (1) and (2) directly allows us to select DCT coefficients with the highest MI w.r.t. speaker class by selecting those that maximise $I(X; C)$. However no provision is made for redundancy between selected coefficients which may result in less than optimal selections. Therefore as an extra step we weight the MI calculation of each coefficient by the conditional MI of $X$ and $C$ given the preselected coefficients $X'$, i.e. according to the *additional* class information it contributes. As the computational cost increases exponentially with the number of preselected coefficients, we calculate the conditional MI using each of the preselected coefficients individually, in turn, and take the mean [13]. The current selected coefficient is therefore the one that maximises;

$$I(X; C) + \frac{1}{N} \sum_{X' \in X_1..X_N} I(X; C|X') \qquad (3)$$

where $X_1..X_N$ represent the $N$ previously selected coefficients. In order to select those DCT coefficients most robust to changes in head pose whilst retaining the highest discriminative ability for speaker verification, we apply the feature selection method outlined in the previous section to our multi-pose dataset, detailed in section 3.1. We use video data of 3 speakers captured from horizontal viewing angles of 0, 10, 20 and 30 degrees to simulate the range of angles that might be encountered in an access control system or similar.

Figure 1 shows DCT feature selection masks ranging from 21 to 36 coefficients generated using the MI analysis, for comparison this analysis was performed on both frontal only and multi angle data. Despite some scattering of the coefficients it is clear in both the frontal and multi angle cases that the even columns carry the highest information content, with a lack of DC component in each case. This is in line with [13] and [16] in that the mouth can be considered approximately horizontally symmetrical, and that by using only even columns horizontal symmetry may be forced in the feature domain thus normalising for small changes of *in* plane rotation. Also in line with the same works, the MI based selections using frontal data contain coefficients spread fairly evenly around the top left corner where one would expect to find the highest energy. Interestingly however the selections based on multi angle data show a preference toward higher frequency vertical components as opposed to horizontal. This result is consistent with the fact that a mouth image suffers the greatest distortion in the horizontal plane due to changes in horizontal viewing angle. The implication of this result can be seen in figure 2 which shows sub-sampled mouth ROIs captured simultaneously from frontal and 30 degree angles and reconstructed using the 21 coefficient frontal and multi angle MI masks. Note the DC component has been included for visual clarity. In the original images the mouth appears compressed in the horizontal plane at the 30 degree view, and likewise the images constructed using the frontal based MI selection show a corresponding reduction in width. In the images reconstructed using the multi angle selection however the mouth appears to take up the full width of the bounding box for either view. This is in fact a smearing effect in the horizontal plane due to the lack of higher order horizontal frequency components, effectively normalising for the change in horizontal viewing angle.

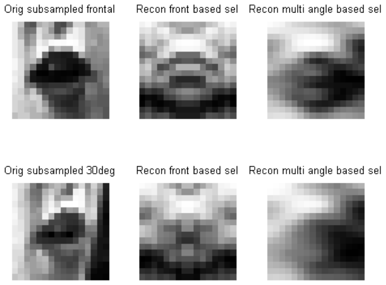

Figure 2: Left: Sub sampled mouth ROI from frontal and 30 degree viewing angles. Middle: Reconstructed using frontal derived 21 coefficient MI mask. Right: Reconstructed using multi angle derived 21 coefficient MI mask

# 3. Methodology

We now present experiments conducted to test the validity of the multi angle MI analysis performed in the previous section, followed by results and discussions in section 4. We begin this section by detailing the collection and preparation of our multi-view AV speech/speaker dataset, followed by feature extraction and details of the experimental setup.

## 3.1. Multi-View AVASR database collection

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

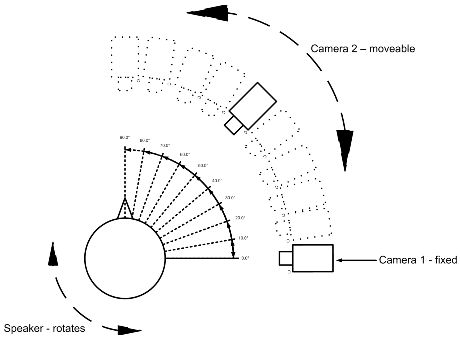

Figure 3: Plan view of recording setup showing fixed position of camera 1, movement of camera 2 and rotation of subject. Example shows simultaneous angles of 0 and 50 degrees

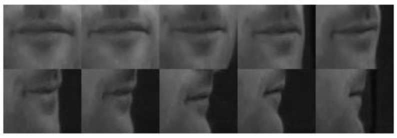

Figure 4: Sample from QuLips database showing pose angles

Recording was organised 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 per speaker for each of the 10 angles (5400 digits total). Only 4 angles (0 to 30 degrees) are considered in this work.

### 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 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 [14], 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 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 [14]. Firstly each video frame was sub-sampled to 16x16 pixels, histogram equalised then converted to grey-scale. 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 mean and variance normalisation across each utterance.

Three shapes of coefficient selection mask are considered in this work (see figure 5). The first is a baseline energy based triangular selection [2], obtained by taking coefficients in a zigzag fashion from the top left corner. The second selection, similar to that used in [16], is based on the frontal only MI analysis of section 2 using only the even columns of a triangular selection minus the DC. The third is based on the multi angle MI analysis and uses only the first 3 even columns of a triangular selection, minus DC. The masks are denoted 'tri', 'even' and '3col', with vector sizes quoted in terms of static features.

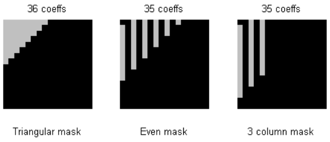

Figure 5: Left to right; 36 coefficient 'tri' mask, 35 coefficient 'even' mask and 35 coefficient '3col' mask

### 3.4. Experiments

Individual speaker modelling is performed using a 32 mixture Gaussian Mixture Model (GMM) per subject, 3 classes in total. Speaker verification likelihoods are normalised using a Universal Background Model (UBM) to obtain the likelihood ratio of the claimed identity to that of the rest of the population. The UBM itself is a 256 mixture GMM trained using the entire XM2VTS dataset.

Two sets of visual only verification experiments were conducted using isolated digits for train and test utterances. True speaker and impostor likelihood ratios were used to generate DET curves (false miss vs false alarm), from which the Equal Error Rates (EER) are obtained. The first set of experiments sets the baseline using the 'tri' selection, showing the effect of pose angle on EER. The second set compares the performance of all 3 selection masks detailed in section 3.3 w.r.t pose angle. Two scenarios are considered; training on a frontal pose only and training on combined frontal, 10, 20 and 30 degree pose angles, with each scenario tested on all four pose angles. Testing is performed using a 20 fold cross validated paradigm on the digits themselves, with 19 digits per recording used for training and 1 for testing. This provides three times as many testing samples for the frontal angle in the first scenario, so frontal EERs in this case are averaged.

## 4. Experimental results

### 4.1. Baseline results

Figure 6 shows how the EER varies using a 'tri' mask over the four pose angles for frontal only and multi-pose trained models. The frontal trained case shows a steady increase in EER as the verification pose angle deviates from the training angle, with the best 36 coefficient mask showing a significant 105% relative EER increase (from 12.2% to 25%) from 0 to just 10 degrees. An important point to note is that a smaller mask appears preferable as the angle increases, i.e. higher frequency components become increasingly detrimental. The lower plot clearly demonstrates the efficacy of incorporating multiple pose angles into the enrolment stage, the verification performance remains approximately constant across all testing angles. It is also interesting to note that although the additional training data is from non-frontal angles, the frontal view verification EERs show a significant improvement from those of the frontal view trained model, with the 36 coefficient mask showing a relative 73% EER reduction (12.2% to 3.3%).

### 4.2. Comparison of feature selection masks

Figure 7 shows how the performance of each of the 'tri', 'even' and '3col' selection masks compares w.r.t. pose angle for varying selection sizes. The plots are organised such that the top two represent frontal only training and the bottom two multi

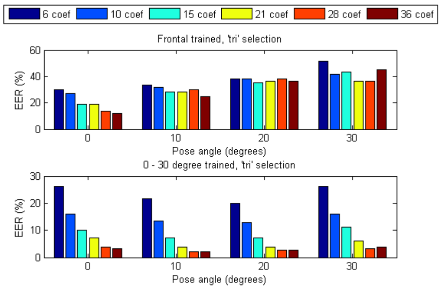

Figure 6: EERs using 'tri' selection of varying sizes across various pose angles. Top: Frontal trained models. Bottom: Multi view trained

view training. The left hand plots show frontal only verification whereas the right show verification from all four views (averaged) to represent the scenario where the pose angle may deviate slightly from one use to the next. The first point to note is that the 'even' and '3col' masks consistently outperform the 'tri' mask in each case, confirming the work of [16] in that performance can be improved by forcing lateral symmetry (removing odd columns from DCT). It is the 32 or 35 coefficient '3col' mask in each case that proves optimal, with 0% EER achieved for frontal verification using multi view trained models. The pose invariance of the '3col' mask is particularly noticeable in the case of verification from multiple angles using a frontal trained model. As the selection size increases, the higher order horizontal components of the 'even' mask suffer increased distortion. Thus by removing these components and adding higher order vertical components as per the '3 col' mask, the train test mismatch is reduced and verification performance improved. In the case of the models trained using all angles, the EER reductions from the '3col' mask and to a slightly lesser extent the 'even' mask may also be a result of reduced variance due to pose angle seen during training, thus reducing the tendency of the model to over generalise. The lowest multi view verification EER is 1.67% achieved using multi view training and a 35 coefficient '3col' mask, a 45% relative improvement on the equivalent for a 36 coefficient 'tri' mask (3.06% EER).

## 5. Conclusions

We have presented results showing the effect of horizontal viewing angle on a lip biometric speaker verification system. Using a multi view MI analysis technique we have highlighted the spatial frequency DCT components most robust to changes in pose angle and with highest class discriminability. Experiments using selection masks approximated from this analysis have shown that higher order vertical components are preferable to higher order horizontal components during verification, particularly in the case of a pose angle train/test mismatch. We have also shown that the system can be made more robust to pose angle through multi view training, and that the alternative selection masks improve performance further in this scenario. We report a 45% relative EER reduction in multi view verification when using our selection mask over a common energy based selection for multi view trained models.

## 6. Acknowledgements

This work was supported by DEL and the UK EPSRC under Grant EP/E02-8640/1 ISIS.

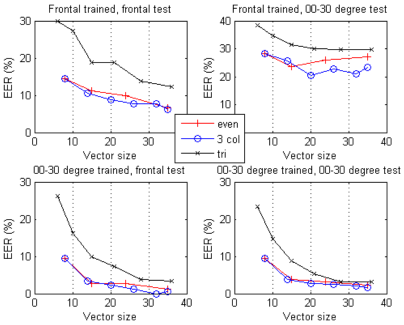

Figure 7: Comparison of selection masks. Top left: Frontal trained, frontal test. Top right: Frontal trained, 0-30 degree test. Bottom left: 0-30 degree trained, frontal test. Bottom right: 0-30 degree trained, 0-30 degree test

## 7. References

[1] E. D. Petajan, "Automatic lip-reading to enhance speech recognition," Ph.D. dissertation, University of Illinois, 1984.

[2] N. A. Fox, R. Gross, J. F. Cohn, and R. B. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *MultMed*, vol. 9, no. 4, pp. 701–714, 2007.

[3] S. L. Wang and A. W. C. Liew, "Ica-based lip feature representation for speaker authentication," in *Proc. SITIS '07*, 2007, pp. 763–767.

[4] A. G. de la Cuesta, J. Zhang, and P. Miller, "Biometric identification using motion history images of a speaker's lip movements," in *Proc. IMVIP '08*, 2008, pp. 83–88.

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

[6] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, "Audio-visual speech recognition using lip movement extracted from side-face images," in *Proc. AVSP*, 2003, pp. 117–120.

[7] ——, "Audio-visual speech recognition using new lip features extracted from side-face images," in *Proc. ROBUST*, 2004.

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

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

[10] 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*, vol. 4, 2007, pp. IV–417–IV–420.

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

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

[13] P. Scanlon, G. Potamianos, V. Libal, and S. M. Chu, "Mutual information based visual feature selection for lipreading," in *Proc. ICSLP*, 2004, pp. 4–8.

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

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

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

[17] H. H. Yang, S. V. Vuuren, and S. Sharma, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Commun.*, vol. 31, no. 1, pp. 35–50, 2000.

[18] 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.

[19] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, vol. 9, 1984, 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, Cambridge, UK, 1995.