Analysing the importance of different visual feature coefficients

Danny Websdale and Ben Milner

University of East Anglia

{d.websdale, b.milner}@uea.ac.uk

Abstract

A study is presented to determine the relative importance of different visual features for speech recognition which includes pixel-based, model-based, contour-based and physical features. Analysis to determine the discriminability of features is performed through F-ratio and J-measures for both static and temporal derivatives, the results of which were found to correlate highly with speech recognition accuracy ($r = 0.97$). Principal component analysis is then used to combine all visual features into a single feature vector, of which further analysis is performed on the resulting basis functions. An optimal feature vector is obtained which outperforms the best individual feature (AAM) with $93.5\,\%$ word accuracy.

Index Terms: Visual features, speech recognition, F-ratio, J-measure, PCA

1. Introduction

Over the course of research into visual speech processing, many different visual feature representations have been proposed and applied to a wide range of applications. The range of visual features can be broadly grouped into four types: pixel-based, model-based, contour-based and physical [1]. Pixel-based features used pixel intensities from the speaker's mouth and have low computation complexity, such as 2D discrete cosine transform (DCT) features [2]. Model-based features create a model to extract visual information and are generally of higher complexity, and include the widely used active appearance model (AAM) features [3]. Contour-based features use a pixel boundary around the mouth to produce a shape signature which may then undergo further transformation, such as Fourier descriptors [4]. Finally, physical features measure geometric properties of the mouth such as simple measurements of height and width.

Several studies have compared visual features and reveal model and pixel based methods to provide best performance [5, 6, 7]. These tests are typically performed by building speech recognisers using different visual features and comparing performance. In this work we aim to identify which specific visual coefficients offer most discrimination in classification tasks. This is achieved by first measuring the F-ratios of individual visual coefficients taken from a large range of different visual feature types and secondly using J-measures to measure the discriminability of an entire visual feature and comparing this to the equivalent speech recognition accuracy. Following this detailed analysis, the effect of applying principal component analysis (PCA) to a range of visual features is explored and experiments used to see how best to include temporal derivatives.

The remainder of this paper is organised as follows. Section 2 provides an overview of the four visual features analysed in this work. Section 3 then uses F-ratios and J-measures to analyse the discriminability of both static and temporal features. PCA is then applied to the visual features in Section 4 and the resulting basis functions of the transforms observed.

2. Visual features

Four visual features are considered, 2D-DCT, AAM, Fourier descriptors and geometric, and are now described briefly.

2.1. Two-dimensional DCT

Two-dimensional DCT (2D-DCT) features are pixel-based and extracted from a $N \times M$ matrix of pixel intensities $\mathbf{P}$, where in this work $N = 90$ and $M = 110$ [2]. The mouth centre is generated from tracked landmarks and features extracted from $\mathbf{P}$ by applying a 2D-DCT

$$\mathbf{q}_{u,v} = W_u W_v \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} p_{i,j} \cos\left(\frac{u\pi(2i+1)}{2N}\right) \cos\left(\frac{v\pi(2j+1)}{2M}\right) \quad (1)$$

$$W_u = \begin{cases} \sqrt{1/N} & \text{if } u = 0 \\ \sqrt{2/N} & \text{otherwise} \end{cases} \quad W_v = \begin{cases} \sqrt{1/M} & \text{if } v = 0 \\ \sqrt{2/M} & \text{otherwise} \end{cases} \quad (2)$$

where $p_{i,j}$ refers to the pixel intensity in row i and column j, producing $\mathbf{q}_{u,v}$. Energy from the image is concentrated into the lower coefficients of $\mathbf{q}$, of which the first 23 are extracted in a zigzag order producing visual vector $\mathbf{c}_t$ for time t

$$\mathbf{c}_t = [\mathbf{q}_{0,0}\ \mathbf{q}_{0,1}\ \mathbf{q}_{1,0}\ \mathbf{q}_{2,0}\ \mathbf{q}_{1,1}\ \mathbf{q}_{0,2}\ \mathbf{q}_{0,3}\ \mathbf{q}_{1,2}\ ...\ \mathbf{q}_{5,1}] \quad (3)$$

2.2. Active appearance model

AAM features are model-based and a combination of shape and appearance. Although shape and appearance could be used separately, their combination using principle component analysis (PCA) has been demonstrated in [5] to produce higher performance by creating a more compact and de-correlated feature set. AMMs require labelled data with landmarks to generate features and use a model to perform this task automatically. The model requires hand labelled training images to learn the variation in mouth shapes and in this work 43 training images were used with 101 landmarks tracked. Forty-six and 20 landmarks represent the outer and inner lip respectively, with the extra landmarks for the eyes and jaw line, which assist the model in locating the face and fitting landmarks. A new model is produced by selecting only the mouth landmarks, and is used to produce AAM features, $A_t = [\mathbf{s}_t\ \mathbf{a}_t]$, that comprise shape, $\mathbf{s}_t$, and appearance, $\mathbf{a}_t$, components for time t.

2.2.1. Shape

The shape feature, $\mathbf{s}$, is obtained by concatenating n x and y coordinates that form a two-dimensional mesh of the mouth, $\mathbf{s} = (x_1 y_1, ..., x_n y_n)^T$. A model that allows linear variation in

shape is produced using PCA,

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} p_i \mathbf{s}_i \tag{4}$$

where $\mathbf{s}_0$ is the base shape, $\mathbf{s}_i$ are the shapes corresponding to the m largest eigenvectors and p_i are shape parameters. Coefficients comprising 90 % of the variation are selected, resulting in a vector size of 8 shape coefficients, $\mathbf{s}_t$.

2.2.2. *Appearance*

The appearance feature, $\mathbf{a}$, is obtained from the pixels that lie inside the base mesh, $\mathbf{s}_0$ [8]. As with the shape model, an appearance model, $\mathbf{a}$, can also be expressed with linear variation,

$$\mathbf{a} = \mathbf{a}_0 + \sum_{i=1}^{m} q_i \mathbf{a}_i \tag{5}$$

where $\mathbf{a}_0$ is the base appearance, $\mathbf{a}_i$ are the appearances that correspond to the m largest eigenvectors and q_i are appearance parameters. Coefficients comprising 95 % of the variation are selected, giving a vector size of 15 appearance coefficients, $\mathbf{a}_t$.

2.3. Fourier descriptor-based visual features

Fourier descriptors are contour-based features generated by applying a Fourier transform to a shape signature, obtained from the pixel boundary of a mouth region. Possible shape signatures are complex coordinates, curvature function, cumulative angular function and centroid distance, and were compared in [4]. This revealed centroid distance performs best, and as such is selected for this work.

The centroid distance is calculated for the outer and inner lip contours separately, and consists of finding the Euclidean distance between lip contour and mouth centre (x_c, y_c), (Figure 1(a)), producing shape signature $\mathbf{r}$ as shown in Figure 1(b),

$$r(i) = \sqrt{(x(i) - x_c)^2 + (y(i) - y_c)^2} \tag{6}$$

$$\text{where, } x_c = \frac{1}{N}\sum_{i=0}^{N-1} x(i) \quad , \quad y_c = \frac{1}{N}\sum_{i=0}^{N-1} y(i) \tag{7}$$

The waveform $\mathbf{r}$ is split into two halves: the first half is the upper lip showing a double peak around the philtrum, and second half is the lower lip with a smooth contour. An FFT is applied to $\mathbf{r}$ and the magnitude, $|\mathbf{r}|$, calculated as illustrated in Figure 1(c). This is truncated to 10 coefficients which is sufficient to describe the shape effectively [4], and produces the final Fourier descriptor feature vector, $F_t = [\mathbf{f}_t^{out}\ \mathbf{f}_t^{in}]$

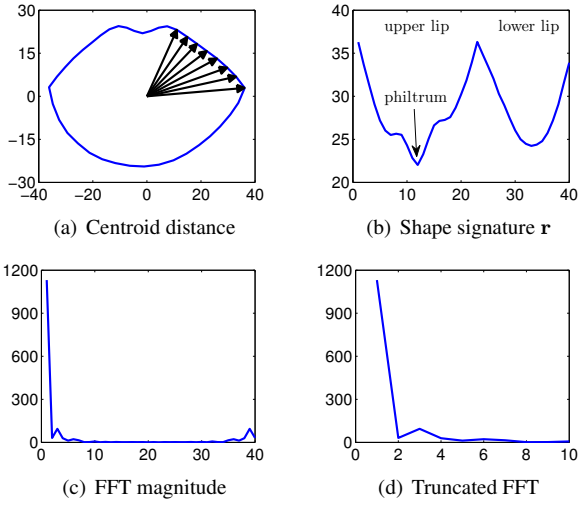
$$\mathbf{f}_t^{out} = [|r_0^{out}|\ |r_1^{out}|\ |r_2^{out}|\ ...\ |r_9^{out}|] \tag{8}$$

$$\mathbf{f}_t^{in} = [|r_0^{in}|\ |r_1^{in}|\ |r_2^{in}|\ ...\ |r_9^{in}|] \tag{9}$$

2.4. Geometric visual features

Geometric features are physical features representing properties of the mouth and comprise height, width, perimeter and area, for both the outer and inner lip. These are extracted from the tracked landmarks discussed in Section 2.2. An example of geometric features for the outer lip is shown in Figure 2.

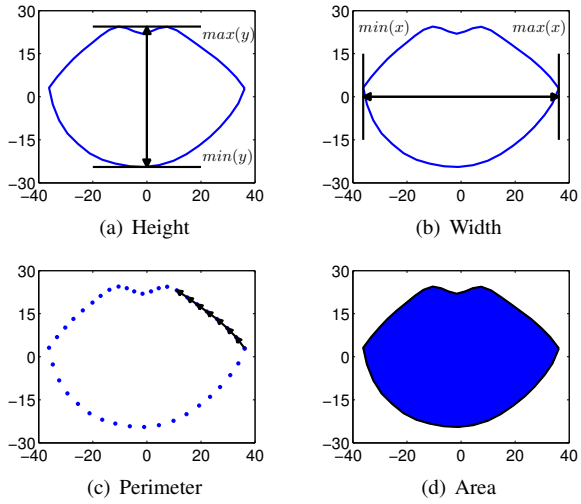
Combining all the features gives the final geometric feature vector, $G_t = [\mathbf{g}_t^{out}\ \mathbf{g}_t^{in}]$, where $\mathbf{g}_t^{out}$ and $\mathbf{g}_t^{in}$ are geometric features for outer and inner lip contours respectively,



(a) Centroid distance (b) Shape signature $\mathbf{r}$

(c) FFT magnitude (d) Truncated FFT

Figure 1: *Fourier descriptor centroid distance method for outer lip contour*



(a) Height (b) Width

(c) Perimeter (d) Area

Figure 2: *Geometric features of height, width, perimeter and area for the outer lip.*

$$\mathbf{g}_t^{out} = [\text{height}_t^{out}\ \text{width}_t^{out}\ \text{perimeter}_t^{out}\ \text{area}_t^{out}] \tag{10}$$

$$\mathbf{g}_t^{in} = [\text{height}_t^{in}\ \text{width}_t^{in}\ \text{perimeter}_t^{in}\ \text{area}_t^{in}] \tag{11}$$

2.5. Combining visual features

For analysis purposes a single 74 dimensional vector, $\mathbf{z}_t$, comprising all coefficients is produced by concatenating the previous visual features

$$\mathbf{z}_t = [\mathbf{c}_t\ \mathbf{s}_t\ \mathbf{a}_t\ \mathbf{f}_t^{out}\ \mathbf{f}_t^{in}\ \mathbf{g}_t^{out}\ \mathbf{g}_t^{in}] \tag{12}$$

3. Analysis of visual features

This section analyses the discriminability of the visual coefficients using F-ratios and J-measures. Both static and temporal derivatives are considered and comparisons are made with speech recognition accuracy.

3.1. Speech database

The GRID audio-visual speech database was used for the visual feature tests and contains recordings from 34 speakers who each produced 1000 sentences [9]. Each sentence comprises six words and follows the grammar shown in Table 1. Speaker 12 was selected for the analysis, with 800 sentences selected for the training set, and 200 for the test set.

Table 1: GRID sentence grammar.

command	colour	preposition	letter	digit	adverb
bin	blue	at	A-Z	1-9	again
lay	green	by	minus W	zero	now
place	red	in			please
set	white	with			soon

3.2. Baseline speech recognition results

To examine the effectiveness of the different visual features, baseline speech recognition tests are first performed. Fifty-one hidden Markov models (HMMs) are trained to model the words in the database and an additional HMM is trained for non-speech movement. The HMMs have a left-right topology with diagonal covariance matrices. An exploratory search found the best speech recognition configuration for each visual feature type by varying the number of states (1-25) and modes (1-4). Velocity (Δ) and acceleration ($\Delta\Delta$) temporal derivatives were augmented to the static features and z-score normalisation applied [10]. The highest recognition found for each visual feature is shown in Figure 3. AAM features attain best results with 92.33 %, followed by 2D-DCT with 91.17 %. Geometric features have lowest performance of 63.58 %, attributed in part to containing fewer coefficients compared to AAM and 2D-DCT.

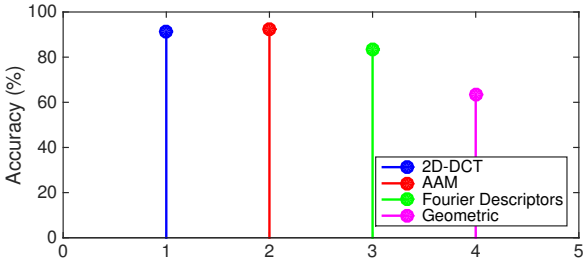


Figure 3: *Optimal baseline ASR results for visual feature types.*

3.3. Relative importance of static visual coefficients

Analysing the discriminability of each coefficient of the static visual vector $\mathbf{z}$, should reveal which are more important with regards to recognition. The discriminative ability of each coefficient is measured using the F-ratio [11]

$$F-ratio = \frac{\text{Variance of means (between} - \text{class)}}{\text{Mean of variances (within} - \text{class)}} \quad (13)$$

This is computed from 51 single mode 19 state HMMs. Each state of each HMM is considered a class from which between-class and within-class covariances are computed. Larger F-ratios suggest a more discriminant coefficient and these are plotted in Figure 4 for the individual coefficients in $\mathbf{z}$.

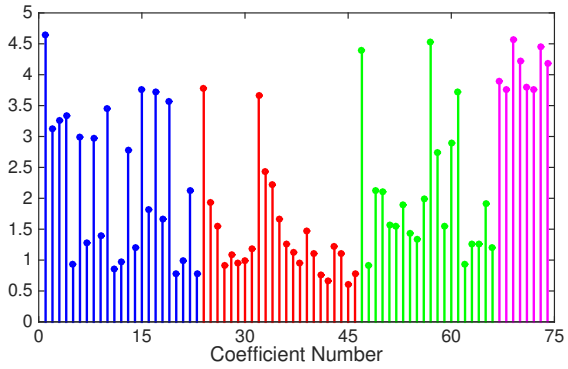


Figure 4: *F-ratios for the coefficients of visual vector* $\mathbf{z}$*. Blue = 2D-DCT, red = AAM, green = Fourier, pink = geometric.*

3.3.1. Rank order using F-ratio

Taking the F-ratios of $\mathbf{z}$ and sorting into descending order produces a re-ordered visual vector, $\tilde{\mathbf{z}}$. For illustration, the 30 most discriminative coefficients are shown in Table 2. Analysing this table reveals the importance of low order coefficients from all visual feature types, where the energy components lie. Also, all 8 geometric features are in the first 12 ranks which suggests they posses discriminative information.

Table 2: Ordered $\tilde{\mathbf{z}}$ coefficients by F-ratio rank on $\mathbf{z}$.

Rank	Feature	Rank	Feature	Rank	Feature
1	$\mathbf{c}_1$	11	$\mathbf{g}_2^{out}$	21	$\mathbf{c}_2$
2	$\mathbf{g}_3^{out}$	12	$\mathbf{g}_2^{in}$	22	$\mathbf{c}_6$
3	$\mathbf{f}_1^{in}$	13	$\mathbf{c}_{15}$	23	$\mathbf{c}_8$
4	$\mathbf{g}_3^{in}$	14	$\mathbf{c}_{17}$	24	$\mathbf{f}_4^{in}$
5	$\mathbf{f}_1^{out}$	15	$\mathbf{f}_5^{in}$	25	$\mathbf{c}_{13}$
6	$\mathbf{g}_4^{out}$	16	$\mathbf{a}_1$	26	$\mathbf{f}_2^{in}$
7	$\mathbf{g}_4^{in}$	17	$\mathbf{c}_{19}$	27	$\mathbf{a}_3$
8	$\mathbf{g}_1^{out}$	18	$\mathbf{c}_{10}$	28	$\mathbf{a}_4$
9	$\mathbf{g}_1^{in}$	19	$\mathbf{c}_4$	29	$\mathbf{c}_{22}$
10	$\mathbf{s}_1$	20	$\mathbf{c}_3$	30	$\mathbf{f}_3^{out}$

3.3.2. Ranked speech recognition results

Speech recognition is performed using the re-ordered visual vector, $\tilde{\mathbf{z}}$, by successively truncating from 74 to 1 coefficients. The HMMs have 19 states and 1 mode, and Figure 5 shows recognition accuracy using from 74 to 1 coefficients. Maximum performance was found using all 74 coefficients, achieving 77.42 %. Recognition is above 70 % until $n < 38$, drastically reducing when $n < 20$.

3.3.3. J-measures

The F-ratio measures how distinct an individual coefficient is, however, to evaluate the discrimination of an entire feature vector a multi-variate extension is required which is known as J-measures [12]. As shown in [13] J_1 and J_S provide best performance for evaluation, and exhibit strong correlation with recog-

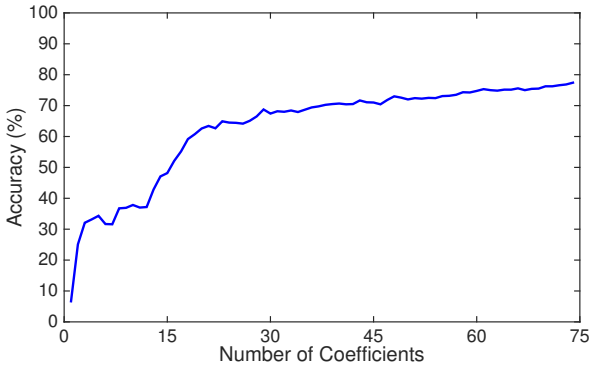


Figure 5: *Recognition accuracy while truncating* $\mathbf{z}$ *by F-ratio rank order.*

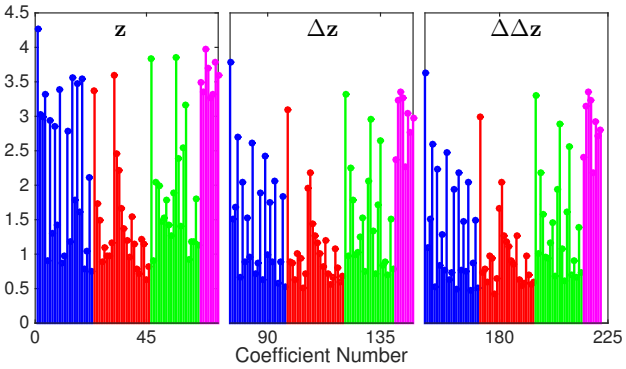


Figure 7: *F-ratio on* $\mathbf{z}'$ *coefficients showing static, velocity and acceleration components (Refer to Figure 3 for colour legend).*

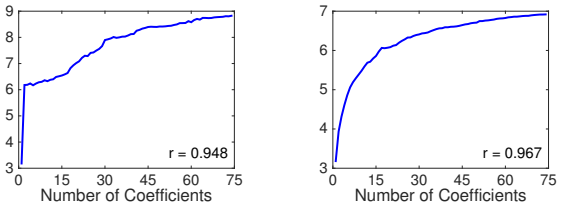
nition accuracy when applied to audio (MFCCs),

$$J_1 = tr(\mathbf{W}^{-1}\mathbf{B}) \tag{14}$$

$$J_S = \sum_{i=1}^{n} \frac{\mathbf{B}_{i,i}}{\mathbf{W}_{i,i}} \tag{15}$$

where $tr()$ indicates the trace of a matrix, $\mathbf{B}$ represents the between-class covariance matrix, and $\mathbf{W}$ is the within-class covariance matrix, for feature vector size n.

The J-measures are applied to $\tilde{\mathbf{z}}$ to investigate whether correlation with speech recognition performance can be observed regarding the reordered visual features. Figure 6 shows the *log* of the J-measures of $\tilde{\mathbf{z}}$, with correlation between J-measures and recognition accuracy visible. High correlation is observed for both measures, with J_S outperforming J_1, which agrees with the observation made in [13], suggesting J-measures can predict visual speech recognition accuracy.



(a) $log(J_1)$ (b) $log(J_S)$

Figure 6: J_1 *and* J_S *when truncating* $\tilde{\mathbf{z}}$*, with correlation coefficient to ASR shown.*

3.4. Relative importance of temporal visual coefficients

The analysis made on static visual features is now extended to consider the effect of temporal derivatives. Taking the static coefficients of vector $\mathbf{z}$ and augmenting temporal information to produce $\mathbf{z}' = [\mathbf{z}\ \Delta\mathbf{z}\ \Delta\Delta\mathbf{z}]$, and computing F-ratios reveals the discriminability of temporal information. This is shown in Figure 7 where a repeat in the pattern of static feature coefficient discrimination translates to velocity and acceleration with a slight reduction for higher order derivatives.

3.4.1. Rank order using F-ratio

Taking the result of applying the F-ratio to $\mathbf{z}'$ (Figure 7) and sorting into descending order produces feature vector $\tilde{\mathbf{z}}'$. Table 3 shows the top 30 coefficients which reveals similar coefficients picked out as in Section 3.3.1, although in a slightly different order. The temporal derivatives of the top performing static coefficients outperform previously high ranked static features. Again, geometric features perform well with 14 of the top 30 places.

Table 3: Ordered $\tilde{\mathbf{z}}'$ coefficients by F-ratio rank.

Rank	Feature	Rank	Feature	Rank	Feature
1	$\mathbf{c}_1$	11	$\mathbf{c}_{15}$	21	$\mathbf{c}_4$
2	$\mathbf{g}_3^{out}$	12	$\mathbf{c}_{19}$	22	$\Delta\mathbf{f}_1^{out}$
3	$\mathbf{f}_1^{in}$	13	$\mathbf{g}_1^{out}$	23	$\Delta\Delta\mathbf{f}_1^{out}$
4	$\mathbf{f}_1^{out}$	14	$\mathbf{c}_{17}$	24	$\mathbf{g}_1^{in}$
5	$\Delta\mathbf{c}_1$	15	$\mathbf{c}_{10}$	25	$\Delta\mathbf{g}_4^{out}$
6	$\mathbf{g}_3^{in}$	16	$\mathbf{s}_1$	26	$\Delta\Delta\mathbf{g}_4^{out}$
7	$\mathbf{g}_4^{out}$	17	$\mathbf{g}_2^{out}$	27	$\Delta\mathbf{g}_2^{out}$
8	$\Delta\Delta\mathbf{c}_1$	18	$\Delta\Delta\mathbf{g}_3^{out}$	28	$\mathbf{f}_5^{in}$
9	$\mathbf{a}_1$	19	$\Delta\mathbf{g}_3^{out}$	29	$\Delta\Delta\mathbf{g}_2^{out}$
10	$\mathbf{g}_4^{in}$	20	$\mathbf{g}_2^{in}$	30	$\Delta\mathbf{s}_1$

3.4.2. Re-ordered speech recognition results

Speech recognition is now applied to the re-ordered temporal visual feature, $\tilde{\mathbf{z}}'$, truncated from 222 to 1 coefficients, and the accuracy shown in Figure 8 using 19 state 1 mode HMMs. Maximum performance was found using 219 coefficients, achieving 85.25 %, which is a 7.83 % increase over that found with $\mathbf{z}$ only. Recognition is held above 70 % until $n < 46$, drastically reducing when $n < 15$. Again, a dip in performance is observed where static and temporal information for geometric area features lie.

4. Using PCA to combine visual features

The previous re-ordering and truncation of the visual feature provided useful information with respect to the discrimination of individual coefficients and effectively either retained or removed each coefficient. Applying PCA now allows a new feature vector to be created that is formed from a weighted combination of the original visual feature vector. This section investigates the effect of applying PCA to the visual features.

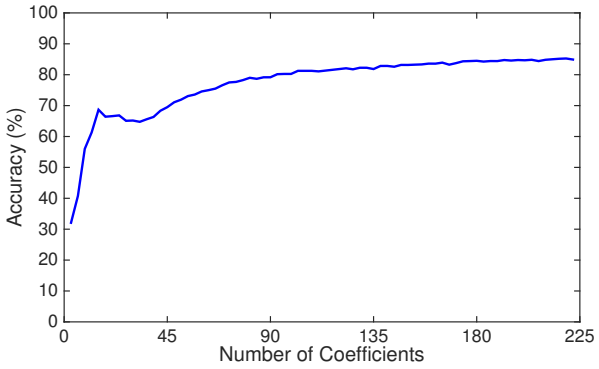


Figure 8: *Recognition accuracy when truncating* $\tilde{\mathbf{z}}'$ *by F-ratio rank order.*

4.1. PCA on static visual features z

PCA is applied to $\mathbf{z}$ to find a compact and de-correlated static feature, $\mathbf{v}$, by multiplying by transform matrix, $\mathbf{H}$,

$$\mathbf{v}_t = \mathbf{H} \times \mathbf{z}_t \tag{16}$$

where the rows, $\mathbf{h}_i$, of $\mathbf{H}$ are the PCA-derived basis functions found from the eigenvectors of the within-class covariance matrix, ranked by their eigenvalues.

4.1.1. Analysis of basis functions

Figure 9 shows the first four basis functions, $\mathbf{h}_i$, of $\mathbf{H}$. Figure 9(a) shows the first basis function, $\mathbf{h}_1$, which should produce the most discriminative coefficient of the new feature vector, $\mathbf{v}$. This provides relatively equal weighting for most coefficients, except geometric which are all given a high weight. The second basis function, $\mathbf{h}_2$, shown in Figure 9(b), notably removes weighting for geometric and low order 2D-DCT coefficients. Figure 9(c) shows the third basis function, $\mathbf{h}_3$, which provides more weighting for all AAM features and less for geometric features of the inner lip. The fourth basic function, $\mathbf{h}_4$, shown in Figure 9(d), again provides more weight for AAM, focusing on shape, and suppresses Fourier descriptors for the inner lip and all geometric features.

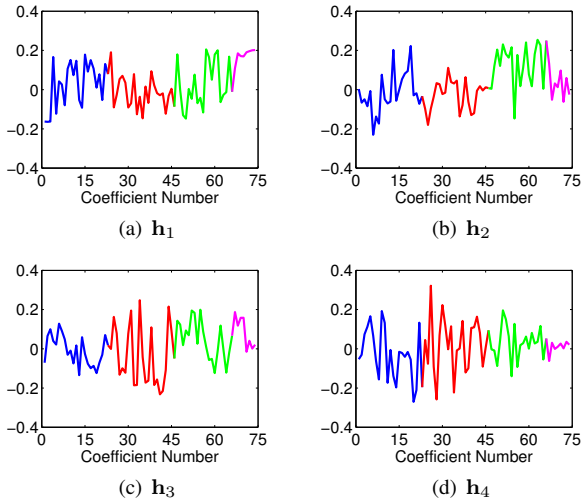
4.1.2. Speech recognition results for PCA-derived features

Speech recognition uses PCA-derived visual features, $\mathbf{v}$, which are truncated from 74 to 1 coefficients. Accuracy is shown in Figure 10, with maximum performance of 86.75 % using 40 coefficients and a sharp drop when using only 8 coefficients. Using PCA on just static features outperforms F-ratio-derived features using both static and temporal information. This is not surprising due to the flexibility of weighting individual coefficients available with PCA but not possible with the F-ratio which is limited to retaining or removing a coefficient.

4.2. PCA with temporal derivatives

Adding temporal information is known to improve speech recognition accuracy, as was shown in Section 3.4.2 when combined with F-ratios. There are two approaches to adding temporal information within the framework of PCA. The first is to apply temporal information prior to PCA, i.e. applying PCA to $\mathbf{z}'$ to give

$$\mathbf{o_t} = \mathbf{D} \times \mathbf{z'_t} \tag{17}$$



(a) $\mathbf{h}_1$ (b) $\mathbf{h}_2$ (c) $\mathbf{h}_3$ (d) $\mathbf{h}_4$

Figure 9: *Analysis of first four PCA basis functions for* $\mathbf{v}$ *from transform matrix* $\mathbf{H}$.

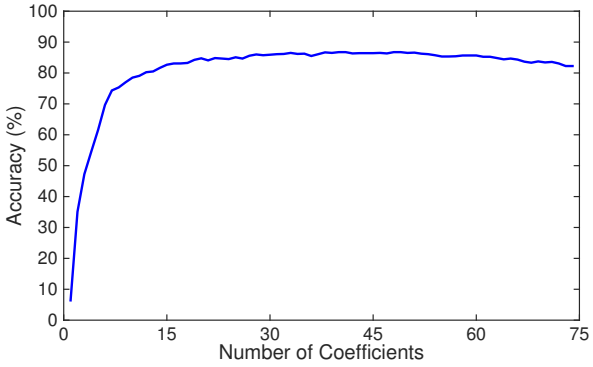


Figure 10: *Recognition accuracy when truncating PCA-derived feature,* $\mathbf{v}$ *from* 74 *to* 1 *coefficients.*

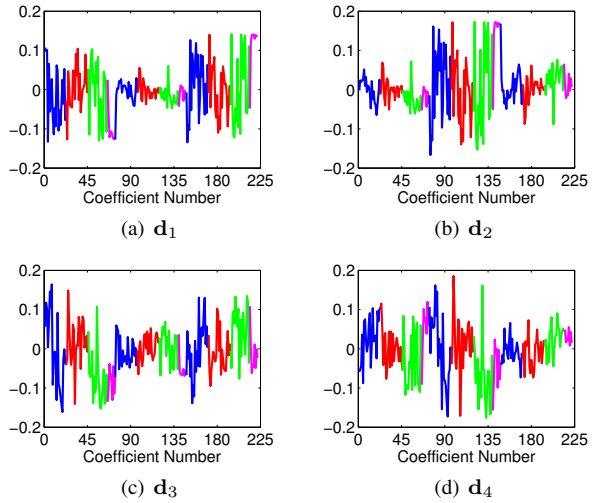
where $\mathbf{D}$ is the PCA transform used to generate feature vectors $\mathbf{o}_t$. The second method is to apply PCA to the static features (c.f. Equation 16) and then compute temporal derivatives, i.e. $\mathbf{v}' = [\mathbf{v}\ \Delta\mathbf{v}\ \Delta\Delta\mathbf{v}]$.

4.2.1. Analysis of basis function for $\mathbf{o}$

The PCA-derived basis functions of transform matrix, $\mathbf{D}$, now include both static and temporal components and Figure 11 shows the first four basis functions of $\mathbf{D}$. These appear to be split into two separate configurations, with the first and third basis functions, d_1 (Figure 11(a)) and d_3 (Figure 11(c)), providing more weight for static and acceleration features. Conversely, the second and fourth basis functions, d_2 (Figure 11(b)) and d_4 (Figure 11(d), provide more weight for the velocity features. Within these, the first and second basis function have more clear boundaries between the static, velocity and acceleration features, compared to the third and fourth basis functions.

4.2.2. Speech recognition results for PCA-derived features with temporal derivatives

To compare the augmentation of temporal derivatives before and after the application of PCA, speech recognition tests are



(a) $\mathbf{d}_1$ (b) $\mathbf{d}_2$ (c) $\mathbf{d}_3$ (d) $\mathbf{d}_4$

Figure 11: *Analysis of first four PCA basis functions for* $\mathbf{o}$ *from transform matrix* $\mathbf{D}$.

performed on features $\mathbf{o}$ and $\mathbf{v}'$, respectively. As previously, the number of coefficients is reduced from 222 to 1 coefficient. Table 4 shows the highest performance achieved with each configuration, along with the result when retaining 95 % of the variation. Performance is similar between both configurations, with $\mathbf{o}$ slightly outperforming $\mathbf{v}'$. Using 95 % of the variation produces comparable results with less than half the coefficients.

Table 4: ASR results for PCA with temporal features.

Configuration	Number of Coefficients	Accuracy
$\mathbf{o}$	148	89.42 %
$\mathbf{o}_{95}$	69	88.67 %
$\mathbf{v}'$	153	88.92 %
$\mathbf{v}'_{95}$	69	87.66 %

4.3. Optimising PCA visual feature

The previous speech recognition tests all used 19 states and a single mode within each state. Using the four configurations outlined in Section 4.2.2, an exploratory search is now made to determine the optimal number of states and modes with the aim of finding best recognition performance. Table 5 shows the HMM configurations that give best performance attained for each feature. All configurations outperform the best individual visual feature, which was AAM as shown in Figure 3 with best configuration, $\mathbf{v}'$ using 153 coefficients with 12 states and 2 modes, providing 1.17 % increase. Again, retaining 95 % of the variance provides comparable results.

Table 5: ASR Results for optimal PCA with temporal features.

Configuration	States	Modes	Accuracy
$\mathbf{o}$	10	2	92.75 %
$\mathbf{o}_{95}$	11	2	92.50 %
$\mathbf{v}$	12	2	93.50 %
$\mathbf{v}'_{95}$	14	3	92.92 %

5. Conclusion

This study has shown the importance of different static and temporal visual features via F-ratio and J-measures. No single feature is more discriminative than others, suggesting a combination of the most discriminative coefficients from each feature type would provide a feature vector that could outperform standard visual features. Interestingly the analysis found that simple geometric features provided high levels of discriminability. PCA was then selected to combine the visual features into a compact, de-correlated feature, of which two approaches for augmenting temporal derivatives were compared. Computing temporal derivatives after PCA gave slightly higher accuracy, attaining 93.5 % word accuracy, an increase of 1.17 % over AAM features.

6. Acknowledgements

We wish to thank the UK Home Office for supporting this work.

7. References

[1] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," in *IEEE Trans. Multimedia*, 2000, pp. 141–151.

[2] G. Meyer, J. Mulligan, and S. Wuerger, "Continuous audio-visual digit recognition using N-best decision fusion," *Information Fusion*, vol. 5, no. 2, pp. 91–101, Jun. 2004.

[3] T. F. Cootes, C. J. Taylor *et al.*, "Statistical models of appearance for computer vision," 2004.

[4] D. Zhang and G. Lu, "A comparative study of curvature scale space and fourier descriptors for shape-based image retrieval," *Journal of Visual Communication and Image Representation*, vol. 14, no. 1, pp. 39–57, 2003.

[5] Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, and R. Bowden, "Comparing visual features for lipreading," in *International Conference on Auditory-Visual Speech Processing 2009*, 2009, pp. 102–106.

[6] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 198–213, 2002.

[7] Z. Zhou, G. Zhao, X. Hong, and M. Pietikainen, "A review of recent advances in visual speech decoding," *Image and Vision Computing 32 (2014) 590605*, vol. 32, pp. 590–605, 2014.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[9] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 150, no. 5, pp. 2421–2424, Nov. 2006.

[10] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden, "Improving visual features for lip-reading." in *AVSP*, 2010, pp. 7–3.

[11] T. Parsons, *Voice and Speech Processing*. McGraw-Hill, 1993, ISBN: 0-07-048541-0.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990, ISBN: 0-12-269851-7.

[13] S. Nicholson, B. Milner, and S. Cox, "Evaluating feature set performance using the F-ratio and J-measures," in *Eurospeech*, Rhodes, Greece, Sep. 1997, pp. 413–416.