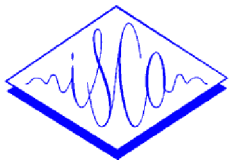
ISCA Archive
http://www.isca-speech.org/archive



AVSP 2010 - International Conference
on Audio-Visual Speech Processing
Hakone, Kanagawa, Japan
September 30-October 3, 2010

Improving Visual Features for Lip-reading

Yuxuan Lan[1], *Barry-John Theobald*[1], *Richard Harvey*[1], *Eng-Jon Ong*[2] *and Richard Bowden*[2]

[1]School of Computing Sciences, University of East Anglia, UK
[2]School of Electronics and Physical Sciences, University of Surrey, UK
{y.lan,b.theobald,r.w.harvey,}@uea.ac.uk,{e.ong,r.bowden}@surrey.ac.uk

Abstract

Automatic speech recognition systems that utilise the visual modality of speech often are investigated within a speaker-dependent or a multi-speaker paradigm. That is, during training the recogniser will have had prior exposure to example speech from each of the possible test speakers. In a previous paper we highlighted the danger of not using different speakers in the training and test sets, and demonstrated that, within a speaker-independent configuration, lip-reading performance degrades dramatically due to the speaker variability encoded in the visual features. In this paper, we examine feature improvement techniques to reduce speaker variability. We demonstrate that, by careful choice of technique, the effects of inter-speaker variability in the visual features can be reduced, which improves significantly the recognition accuracy of an automated lip-reading system. However, the performance of the lip-reading system still is significantly below that of acoustic speech recognition systems, and an analysis of the confusion matrices generated by the recogniser suggests this largely is due to the number of deletions apparent in a visual-only system.

Index Terms: lip-reading, feature extraction, feature comparison, speaker variability.

1. Introduction

Automatic lip-reading systems are trained using visual features extracted from regions of interest in images and video sequences. These visual features typically are either shape-based, texture-based, or a combination of both, and usually the region of interest from which they are extracted contains only the mouth. Unlike acoustic features used in automatic speech recognition (typically MFCCs), visual features usually are data-driven and so a problem is that they are highly speaker-dependent. To overcome this, many systems are presented either using a speaker-dependent or a multi-speaker configuration — a recogniser will have had prior exposure to all possible test speakers. In [1], we showed that both speaker-dependent and multi-speaker recognition can achieve a level of performance close to that of acoustic speech recognition on an isolated letter recognition task. However, we also showed that compared with speaker-independent acoustic speech recognition, speaker-independent lip-reading performance degrades dramatically.

In this work we attempt to reduce the effects of speaker variability in the visual features and thus improve the discriminative power of the visual features. Feature improvement techniques considered in this work include: per speaker z-score normalisation and HiLDA[2].

2. Data Capture

We use an audiovisual corpus of 12 speakers, 7 male and 5 female, each reciting 200 sentences selected from the Resource Management Corpus [3]. The database has a vocabulary size of approximately 1000 words, and was recorded in full-frontal view using using a tri-chip Thomson Viper FilmStream high-definition camera. The speakers were instructed to keep their head relatively still, and the recording of each speaker was done in a single sitting to ensure constant illumination.

3. Visual Features for Lip-reading

Two forms of visual feature for lip-reading are compared. The first features are low-level, image-based features extracted using a 2D Discrete Cosine Transform (DCT) on the HSV images and the pseudo-hue space [4]. The latter of these features were developed for lip-tracking. The second features are derived from a higher-level, model-based approach that utilises shape and appearance information. These are based on active appearance models (AAMs), and in our previous experiments [5], we have found these to significantly outperform low-level features.

3.1. Active Appearance Models

The *shape*, $\mathbf{s}$, of an AAM is formed by concatenating the x and y-coordinates of a set of n vertices that delineate the features of interest on an object: $\mathbf{s} = (x_1, y_1, \ldots, x_n, y_n)^T$. A compact model that allows a linear variation in the shape is given by,

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

where $\mathbf{s}_0$ is the mean shape and $\mathbf{s}_i$ are the eigenvectors corresponding to the m largest eigenvectors of the covariance matrix. The coefficients p_i are the shape parameters that define the contribution of each eigenvector in the representation of $\mathbf{s}$. The model usually is computed by applying Principal Component Analysis (PCA) to a set of shapes hand-labelled in a corresponding set of images.

The *appearance*, A, of an AAM is defined by the pixels that lie inside the base mesh $\mathbf{s}_0$. AAMs allow linear appearance variation, so A can be expressed as a base appearance A_0 plus a linear combination of l appearance images A_i,

$$A = A_0 + \sum_{i=1}^{l} \lambda_i A_i, \tag{2}$$

where λ_i are the appearance parameters. As with shape, the base appearance A_0 and appearance images A_i are usually computed by applying PCA to the shape normalised training images [6]. A_0 is the mean shape normalised image and the vec-

tors A_i are the (reshaped) eigenvectors corresponding to the l largest eigenvalues.

Although the shape and the appearance components of an AAM can be used as features for lipreading separately, a combination of the two has been shown to be a more discriminative feature [5]. There usually is significant correlation in the change in shape and the change in appearance of the mouth region. For example, as the mouth opens we see the teeth and the tongue, and as the aperture of the mouth decreases the inside of the mouth darkens. To exploit this correlation, a combined model of shape and appearance is constructed. Given the labelled images used to compute the shape and the appearance components of the AAM, the shape and appearance parameters for those images are computed using:

$$\mathbf{p} = \mathbf{s}^T(\mathbf{s} - \mathbf{s}_0) \tag{3}$$

and

$$\lambda = A^T(A - A_0) \tag{4}$$

respectively. These parameters are then concatenated and reweighted to account for the difference in units that the parameters measure (the shape unit is x and y-coordinates and the appearance unit is pixel intensity):

$$\mathbf{b} = \begin{bmatrix} \mathbf{W}\mathbf{p} \\ \lambda \end{bmatrix}, \tag{5}$$

where the weights, $\mathbf{W}$, are used to adjust the overall energy in the shape parameters such that it is the same as the energy in the appearance. These concatenated vectors are subject to a third PCA to give a compact model of the form:

$$\mathbf{b} = \mathbf{Q}\mathbf{c} \tag{6}$$

where the columns of $\mathbf{Q}$ are the q leading eigenvectors of the covariance matrix defining the modes of variation of the combined shape and appearance model, and $\mathbf{c}$ is a vector of combined shape and appearance parameters. This creates a more compact model and, more importantly, de-correlates the features. An example model is shown in Figure 1.

Figure 1: The first three modes of variation of a combined shape and appearance model. The modes are shown at +3 standard deviations (top row) and -3 standard deviations (bottom row).

To obtain the shape vertices, $\mathbf{s}$, required in Equation 3 to compute the shape parameters, we use a linear predictor-based tracker [7] trained for each individual speaker.

4. Feature Improvement Methods

Conventional visual features are highly sensitive to the identity of a speaker as they are derived directly from a speaker's data [1]. Thus the visual features better encode the identity of speaker than the units of speech (e.g. phoneme or viseme) that they are supposed to represent. This characteristic is demonstrated in Figure 2 in which multi-dimensional AAM features, described in Section 3.1, are normalised using a global z-zscore normalisation and are visualised in a 2D space using a Sammon projection [8]. These features represent the same visemic class, /f/, yet there is clear distinction between the features for different speakers (encoded as different colours/symbols), which implies a large within-class variability. Likewise, in Figure 3, the (global) normalised AAM features of two visemic classes, /f/ and /u/, are shown. Rather than occupying distinct regions of the 2D space (good separability between the classes), the features form a single cloud, showing poor between-class separability.

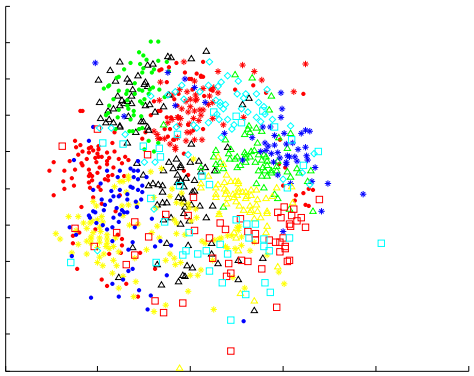


Figure 2: Sammon projection of the AAM features for 12 speakers representing the visemic class /f/ after applying a global z-score normalisation. Different symbols and colour combinations represent different speakers. Speaker-independent features would be expected to form a single cloud in this 2D space, rather than individual speakers occupying a distinct region of the feature space as is observed here.

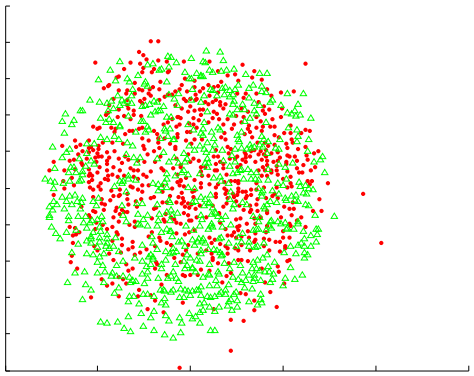


Figure 3: Sammon projection of the AAM features for 12 speakers representing the visemic classes /f/ (red dots) and /u/ (green triangles) after applying a global z-score normalisation. The lack of separation in the features between the classes suggests poor discrimination ability between the classes.

The feature improvement techniques examined in this paper to overcome the above issue include: a z-score normalisation per speaker [9] and HiLDA [2].

4.1. Per-speaker z-score normalisation

Rather than applying a single, global z-score normalisation across the entire data set, it might be more sensible to minimise the distance between the feature spaces for different speakers. That is, to z-score normalise features in the scope of each individual speaker [9] rather than across speakers. Figure 4 illustrates the positive affect of the application of this normalisation, but Figure 5 shows it does little to improve the between-class variability. Of course, assumptions of utilising this technique are that there are adequate data to reliably estimate the mean and standard deviation of the features for each speaker, and that during testing the identity of the speaker is known.

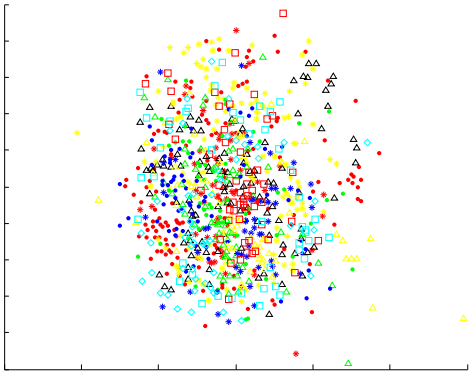


Figure 4: Sammon projection of the AAM features for 12 speakers representing the visemic class /f/ after z-score normalising the features per speaker. Different symbols and colour combinations represent different speakers. This normalisation provides an improvement over the features shown in Figure 2.

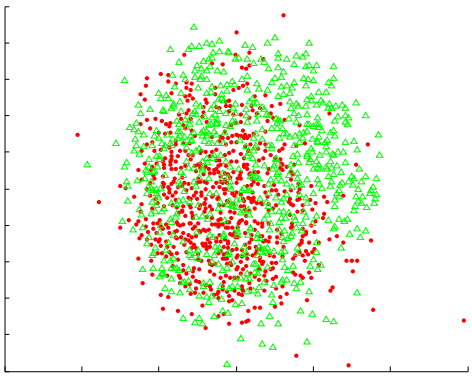


Figure 5: Sammon projection of the AAM features for 12 speakers representing the visemic classes /f/ (red dots) and /u/ (green triangles) after z-score normalising the features per speaker. This normalisation has done little to improve the between-class separation of the features shown in Figure 3.

4.2. Hi-LDA

Both AAM and DCT features are static. They encode information within the current frame and do not encompass dynamic information that may span several frames or across phones (or visemes). In [2], a new visual feature, named Hi-LDA, was introduced in which successive frames of features are considered

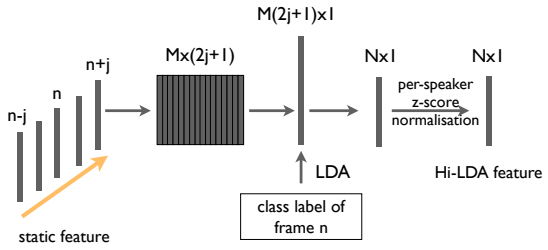


Figure 6: An illustration of constructing Hi-LDA visual features: $2j + 1$ frames centred on the current frame are stack to form hyper-vectors, which later are subject to LDA to learn the optimal projection that maximises the between class distances and, at the same time, minimises the within class distances.

as a single vector. These are used in conjunction with linear discriminant analysis (LDA) to learn the optimal projection that best describes the dynamics of speech, and thus enhance the discriminative power of the features. Figure 6 illustrates the construction of such features. For current frame n, a chunk of $2j + 1$ frames centred on frame n are reshaped to form a single "hyper-vector". During training, LDA learns from these hyper-vectors a set of orthogonal projections that maximise the between class distance and, at the same time, minimise the within class distance. The class label used in LDA can be anything with a discriminative meaning, e.g. phone (or viseme) labels from an audio transcription, or a HMM state sequence from a Viterbi decoding path. In this paper, the audio from the training data are force-aligned [10] to provide viseme labels for LDA training.

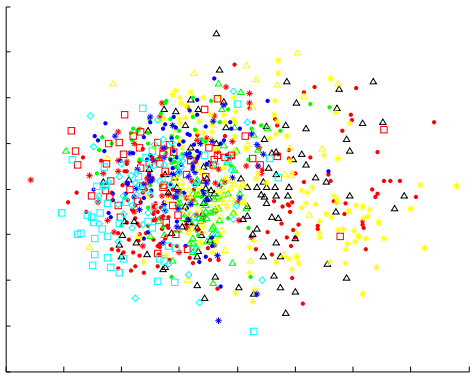


Figure 7: Sammon projection of the AAM features for 12 speakers representing the visemic class /f/ after performing a Hi-LDA speaker-normalisation. Different symbols and colour combinations represent different speakers. Hi-LDA provides an improvement over the features shown both in Figures 2 and 4.

Figures 7 and 8 show an apparent decrease in the within-class difference from different speakers, and a large improvement in the between class variability for features from two classes. For reference, a Sammon projection of the MFCCs for the same speakers and the same classes are shown in Figures 9 and 10. Comparing the Hi-LDA projected visual features and MFCCs, which are designed largely with speaker-independence in mind, we conclude that the Hi-LDA AAM features may have a discriminative power much closer to MFCC features compared with the a simple z-score normalisation of the visual fea-

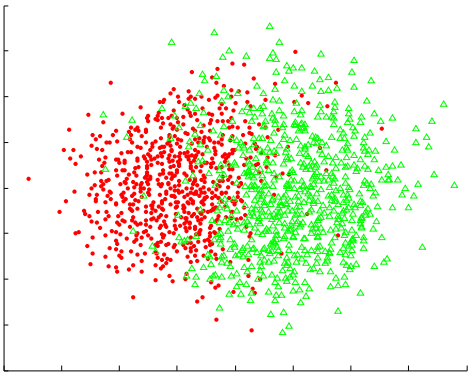


Figure 8: Sammon projection of the AAM features for 12 speakers representing the visemic classes /f/ (red dots) and /u/ (green triangles) after applying Hi-LDA. This approach provides a huge improvement in the between-class separation of the features shown both in Figures 3 and 5.

tures. This is tested by comparing the performance of an audio-only and a visual-only speech recognition system.

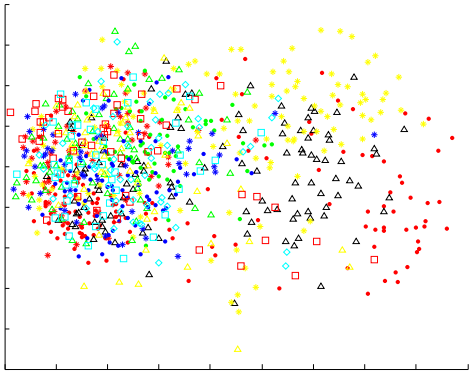


Figure 9: Sammon projection of MFCC features from 12 speakers representing the visemic class /f/. Different symbols and colour combinations represent different speakers.

5. Recognition Experiments

We use Hidden Markov Models (HMMs), which are the method of choice for speech recognition and have been shown to be successful for lip-reading [11, 12]. The HMM toolkit, HTK [13], is applied here for building and manipulating HMMs.

The acoustic speech from the dataset is force-aligned to generate phone-level transcriptions of the utterances, and a standard phoneme to viseme mapping [14] is applied to produce viseme-level transcriptions. These are then used to train and test both the visual-only and audio-only HMMs. Correspondingly, the pronunciation dictionary also is translated from phoneme-level pronunciations to viseme-level pronunications. 14 HMMs are trained on visual features: one for each viseme and one to m odel 'visual silence'. A 'short pause' model is tied to the middle state of the silence model. Left-right HMMs with three states and a diagonal covariance Gaussian Mixture Model (GMM) associated with each state are used.

Single Gaussian HMMs are initialised using flat start train-

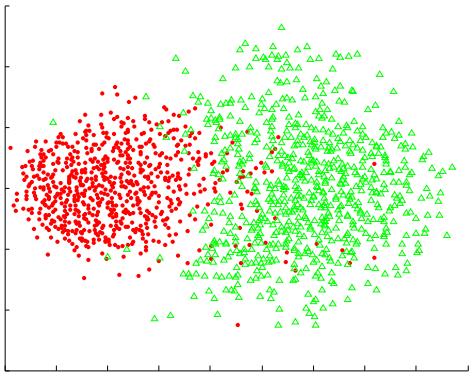


Figure 10: Sammon projection of the MFCC features for 12 speakers representing the visemic classes /f/ (red dots) and /u/ (green triangles). Note the similarity between the speaker-normalised visual features in Figure 8

ing via the HTK module `HCompV`. This is followed by a series of embedded training via `HERest`. Mixture incrementation is then applied to increase the number of Gaussian mixture components from 1 to 2, 5, and 9, all of which are evaluated during the experiments. A bigram word language model is constructed from the training data via `HLStats` and `HBuild` to give a language constraint during recognition. To test the effect of language model, when calling `HVite` during recognition, the word insertion penalty $p = \{-20, 0, 10\}$ and the grammar scale factor $s = \{0, 1, 5, 15\}$. Note that a large word insertion penalty tends to encourage sentences with more words, and a large grammar scale factor puts more emphasis on the language model during recognition.

A total of seven visual features are tested, described in Table 1. All are up-sampled to 100 fps to provide adequate training data demanded by HTK. We also test the performance of a standard audio-only recogniser using MFCC features with 39 dimensions (MFCCs augmented with Δ and $\Delta\Delta$ derivatives).

Feature Name	Explanation
aam	global z-score normalised AAM
aam_sp	speaker z-score normalised AAM
aam_hilda_sp	Hi-LDA projected AAM
dct_hsv_sp	DCT features from HSV images
dct_hsv_hilda_sp	Hi-LDA projected DCT from HSV images
dct_ph_sp	DCT features from pseudo-hue space
dct_ph_hilda_sp	Hi-LDA projected DCT features from pseudo-hue space

Table 1: The visual features used in the recognition experiments. All DCT features are speaker z-score normalised.

To test the robustness of the features across speakers, a 12-fold cross-validation was used, where for each fold, a different speaker is held-out for testing and the classifier is trained on the features from the remaining speakers. In addition, a different AAM is built for each fold where the data for test speaker are not included in constructing the model. The performance of the classifier is measured using the viseme accuracy rate acc, where

$$acc = \frac{H - I}{N}, \tag{7}$$

in which, N is the total number of viseme instances to be recognised, H is the number of correctly recognised viseme instances, and I is the number of insertion errors. The results for all recognisers are shown in Figure 11.

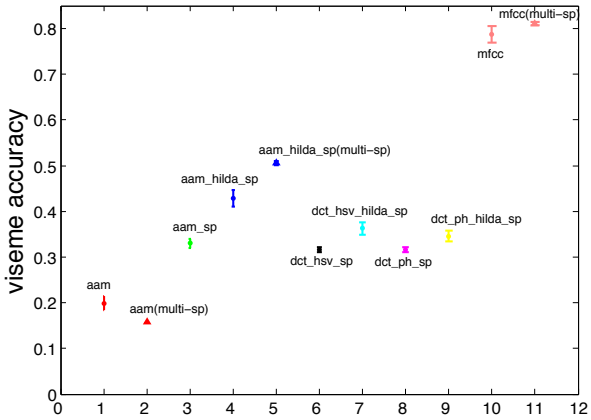


Figure 11: Results for both speaker-independent and multi-speaker recognition, evaluated using 12-fold and 5-fold cross-validation respectively. The mean viseme accuracy is plotted with the error bars showing ± 1 standard error. Note, results marked `(multi-sp)` are for multi-speaker recognisers, the others are speaker-independent. Only the highest accuracy rate is reported here for each feature type.

Among the visual features there is an obvious trend that Hi-LDA features outperform the other types of feature. This is explained by diagrams in Figures 2–7 and Figures 3–8, where Hi-LDA features shows a much smaller within-class variability and a larger between-class variability compared to the other visual features. Figure 11 indicates some improvement in performance after applying a per-speaker z-score normalisation, as demonstrated by the results of the aam and aam_sp features. It is also worth pointing out that all AAM features outperform DCT features of the same type, i.e. aam_sp vs. dct_hsv_sp, and aam_hilda_sp vs. $dct\_hsv\_hilda\_sp$.

For comparison, multi-speaker experiments are also conducted on selected features, i.e. aam, aam_hilda_sp, and $mfcc$, where a 5-fold cross-validation is used. For each fold, the classifier is trained on speech from all 12 speakers, and tested on held-out speech from the same set of speakers. The held-out speech is altered for each fold.

Normally, one expects to see higher recognition results using a multi-speaker recogniser compared with a speaker-independent recogniser. For the aam features, this is not the case. A possible cause is the smaller number of folds used in the multi-speaker experiment (5-folds for the multi-speaker instead of 12 for the speaker-independent test), which results in a smaller training set. As pointed out in Section 4.1, aam features of the same class tend to cluster by the identity of speaker, therefore a classifier will benefit from a larger training set which provides fuller coverage of the feature space. On the other hand, aam_hilda_sp features are much more speaker-independent, so that a classifier can cope with a smaller training set. However, the features still carry some speaker identity information, hence we observe a higher recognition performance for the multi-speaker recognisers compared to the speaker-independent.

5.1. Comparing audio-only and visual-only performance

Figure 8 shows a Sammon projection of the Hi-LDA AAM features. These appear to be as well separated between the classes as the corresponding MFCCs (Figure 10), yet the performance of the visual-only recogniser is significantly worse than the corresponding audio-only recogniser (0.44 vs. 0.78 viseme accuracy).

Figure 12 shows the confusion matrix for test data produced by the best performing visual-only classifier, which was trained using speaker-independent Hi-LDA AAM features. HMMs in the classifier have 9 mixture components and the language model parameters are $p = 10$ and $s = 15$. There are a total of 6747 viseme instances in the test data, and the viseme accuracy is 0.44. Reading the table from left to right, each entry is the number of times a viseme class is recognised as another viseme class. For example, row 1, column 2, indicates that there are 11 instances of 'ah' mis-recognised as 'eh'. The third column from the right is the number of times a viseme class is missed entirely by the classifier, i.e., the deletion error Del, and the last two columns are the correction rate $\%c$ and the deletion rate $\%d$ respectively. These are computed using:

$$c_i = \frac{N_{i,i}}{\sum_j N_{i,j}}; \quad d_i = \frac{Del_i}{\sum_j N_{i,j} + Del_i} \tag{8}$$

where $N_{i,j}$ denotes the number of times instances from class i are recognised as class j and $N_{i,i}$ is the number of correctly classified instance for class i. The entries on the bottom row are the overall correction rate and deletion rate, computed using:

$$c = \frac{\sum_i N_{i,i}}{\sum_{i,j} N_{i,j}}; \quad d = \frac{\sum_i Del_i}{\sum_{i,j} N_{i,j} + \sum_i Del_i} \tag{9}$$

prediction

ground truth	ah	eh	f	ao	t	uh	w	k	p	iy	aa	ch	oo	Del	%c	%d
ah	295	11	10	8	14	10	6	16	5	29	6	1	10	240	70.1	36.31
eh	28	330	2	9	11	6	4	9	3	19	3	2	3	189	76.9	30.58
f	2	1	205	2	4	1	9	4	0	1	0	1	0	66	89.1	22.3
ao	12	7	3	70	8	5	5	4	0	4	7	0	6	83	53.4	38.79
t	37	23	30	17	949	8	28	60	8	24	14	11	5	426	78.2	25.98
uh	2	3	3	3	1	30	2	1	0	3	3	3	3	49	52.6	46.23
w	2	1	6	2	7	0	227	17	1	3	0	4	1	132	83.8	32.75
k	23	12	13	16	28	1	25	638	7	18	9	7	13	408	78.8	33.5
p	9	8	10	0	11	2	29	12	241	2	3	1	5	129	72.4	27.92
iy	27	32	7	13	10	4	12	21	6	359	3	7	6	287	70.8	36.15
aa	1	3	0	0	1	2	1	3	1	1	20	0	1	25	58.8	42.37
ch	4	1	6	2	10	1	7	8	1	2	3	56	5	73	52.8	40.78
oo	2	4	1	1	5	1	2	0	0	0	0	0	44	37	73.3	38.14
Ins	62	45	30	22	69	13	55	73	8	50	27	10	24			
Total															75.26	31.78

Figure 12: Confusion matrix for a visual-only speaker-independent recogniser trained and tested using Hi-LDA AAM features.

Figure 13 shows the equivalent of Figure 12, but for the audio-only data. The language-model settings for both recognisers is the same. The main point to note is that although the features appear to be well separated in the visual domain, the performance of the visual-only recogniser is significantly lower than the audio-only recogniser. This drop in performance largely can be attributed to the number of deletions in the visual-only recogniser. In fact, the deletion rate for the visual-only recogniser is 32%. This accuracy is achieved with a recogniser that uses a large insertion penalty, reflecting the fact it encourages longer sentences. It is almost certain that the difference in performance is due to the coarticulation effects, where many of the visemes cannot be visually observed. It also indicates

prediction

ground truth	ah	eh	f	ao	t	uh	w	k	p	iy	aa	ch	oo	Del	%c	%d
ah	486	22	2	6	8	5	1	3	4	45	1	0	2	76	83.1	11.5
eh	37	488	6	5	5	5	2	2	0	11	6	2	3	46	85.3	7.443
f	1	0	256	1	9	1	0	3	0	1	1	0	0	23	93.8	7.77
ao	22	8	0	144	2	0	2	2	1	1	1	1	0	30	78.3	14.02
t	18	10	21	4	1374	4	1	29	7	3	1	11	1	156	92.6	9.512
uh	1	1	0	1	1	72	0	2	3	8	0	2	0	15	79.1	14.15
w	3	1	4	0	7	0	326	3	1	3	3	3	0	49	92.1	12.16
k	19	6	8	16	38	1	10	958	11	6	5	4	4	132	88.2	10.84
p	8	1	5	2	16	0	3	15	340	2	0	3	0	67	86.1	14.5
iy	16	17	2	10	8	13	3	11	4	629	1	1	5	74	87.4	9.32
aa	0	1	0	0	0	0	3	3	0	2	44	0	0	6	83	10.17
ch	0	0	0	0	4	0	2	0	0	1	0	166	0	6	96	3.352
oo	3	3	0	5	1	0	0	0	0	0	1	0	67	17	83.8	17.53
Ins	102	49	19	36	101	14	37	86	28	53	15	22	5			
Total															88.43	10.33

Figure 13: Confusion matrix for an audio-only speaker-independent recogniser trained and tested using $mfcc$ features.

that for visual speech recognition, a direct translation using a phoneme-to-viseme mapping may not be the best option. The result of a preliminary investigation into accessing subjectively what phonetic units can be (visually) observed during speech is shown in Figure 14. Viewers were asked to watch videos of people speaking without audio and mark where *interesting events* occur. The marked frames delimit short sequences, and we investigate the phonetic composition of these sequences. It is clear, from Figure 14 that the majority of the marked segments are composed of more than one phoneme — an action on the lips of a speaker does not relate directly to an individual phoneme, hence the high number of deletions in the visual-only recogniser.

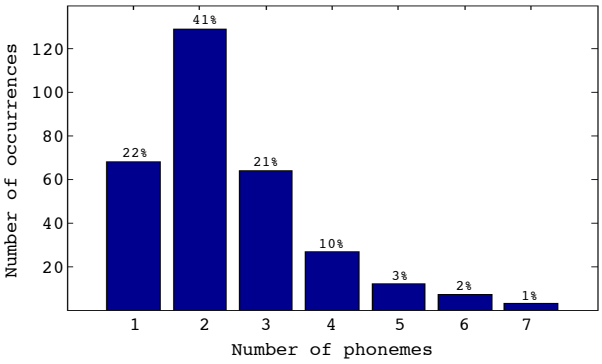


Figure 14: The number of phonemes forming visual events segmented manually in a video sequence. The percentages denote the approximate proportion of segments having that number of phonemes. This accounts for large number of deletions in a visual-only recogniser — an action on the lips of a speaker does not relate directly to an individual acoustic event.

6. Conclusions

In this work we have compared low-level, image-based features and high-level, model-based features for lip-reading. We find that the high-level AAM-based features significantly outperform the low-level features. We also have investigated two approaches for correcting the speaker-dependence of the visual features: namely per-speaker z-score normalisation and Hi-LDA. We have found that Hi-LDA gives a significant improvement in performance over z-score normalisation, and that visualising Hi-LDA AAM features shows the between class separation of the features is almost a good as the standard speaker-independent acoustic feature (MFCCs). However, testing these features in a recogniser suggest that the performance of the visual-only recogniser is significantly worse than a corresponding audio-only recogniser. Inspection of the confusion matrices suggests this degradation in performance is largely because approximately half of the events (visemes) to be recognised are deleted. This perhaps suggests that less future effort should be spent on the front-end of the recogniser (the feature extraction and normalisation) and more effort should focus on the back-end of the recogniser (integrating a better *visual* language model into the recogniser).

7. Acknowledgements

The authors gratefully acknowledge EPSRC (EP/E028047/1) for funding. We also are grateful to David Gibson for his assistance with Figure 14 and to Prof. Stephen Cox and Jacob Newman for discussions.

8. References

[1] S. Cox, R. Harvey, Y. Lan, J. Newman, and B. Theobald, "The callenge of multispeaker lip-reading," in *International Conference on Auditory-Visual Speech Processing 2008*, 2008, pp. 179–184.

[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*. MIT Press, 2004.

[3] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status." in *In Proceedings of the DARPA Speech Recognition Workshop.*, 1986.

[4] N. Eveno, A. Caplier, and P. Coulon, "Automatic and accurate lip tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 5, pp. 706–715, 2004.

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

[6] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.

[7] E. Ong, Y. Lan, B. Theobald, H. R., and R. Bowden, "Robust facial feature tracking using selected multi-resolution linear predictors," in *In Proceedings of the International Conference Computer Vision (ICCV)*, 2009.

[8] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 18, pp. 401–409, 1969.

[9] J. Newman and S. Cox, "Automatic visual-only language identification: A preliminary study." in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 4345–4348.

[10] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

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

[12] J. Luettin and N. Thacker, "Speechreading using probabilistic models," *Computer Vision and Image Understanding*, vol. 65, no. 2, pp. 163–178, 1997.

[13] S. Young, G. Evenmann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (version 3.2.1)*, 2002.

[14] C. G. Fisher, "Confusions among visually perceived consonantsnants," *Journal of Speech and Hearing Research*, vol. 11, pp. 796–804, 1968.