The challenge of multispeaker lip-reading

Stephen Cox, Richard Harvey, Yuxuan Lan, Jacob Newman, Barry-John Theobald

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK.

{s.j.cox,r.w.harvey,y.lan,jacob.newman,b.theobald}@uea.ac.uk

Abstract

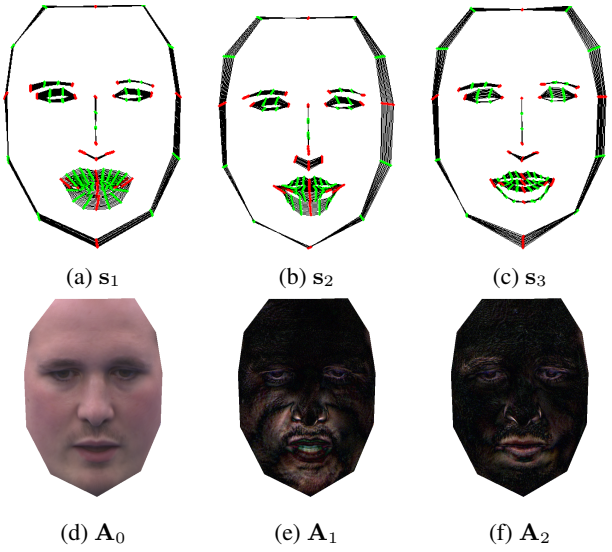
In speech recognition, the problem of speaker variability has been well studied. Common approaches to dealing with it include normalising for a speaker's vocal tract length and learning a linear transform that moves the speaker-independent models closer to to a new speaker. In pure lip-reading (no audio) the problem has been less well studied. Results are often presented that are based on speaker-dependent (single speaker) or multi-speaker (speakers in the test-set are also in the training-set) data, situations that are of limited use in real applications. This paper shows the danger of not using different speakers in the training- and test-sets. Firstly, we present classification results on a new single-word database AVletters 2 which is a high-definition version of the well known AVletters database. By careful choice of features, we show that it is possible for the performance of visual-only lip-reading to be very close to that of audio-only recognition for the single speaker and multi-speaker configurations. However, in the speaker independent configuration, the performance of the visual-only channel degrades dramatically. By applying multidimensional scaling (MDS) to both the audio features and visual features, we demonstrate that lip-reading visual features, when compared with the MFCCs commonly used for audio speech recognition, have inherently small variation within a single speaker across all classes spoken. However, visual features are highly sensitive to the identity of the speaker, whereas audio features are relatively invariant.

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

1. Introduction

Automatic speech recognition (ASR) systems are known to benefit from the inclusion of visual cues in the recognition of degraded auditory speech [1, 2, 3, 4, 5, 6, 7]. Acoustic speech features are augmented with visual features extracted from the mouth region in a video sequence to supplement the information available to the recognizer. Typically these features are low-level, image-based representations, or higher-level, model-based representations of the visible articulators. Both forms of feature have been shown to work equally well in audiovisual speech recognition applications [3] and can also be used to aid speech coding [8].

However, while audiovisual recognition has received much attention, pure lip-reading (video only) has been perceived as capable of producing results of such poor quality that they are only useful as an adjunct to the more reliable results from audio-only recognition. In [3] for example, on a twenty-six class problem (the letters of the alphabet) the video-only accuracy was 44.6% compared to the audio-only result of around 85%. Even in trained human lip-readers the poor performance of lip-reading has caused some comment, [9] for example, and to improve lip-reading performance it is commonplace to use a "lip speaker" who is a person trained to visually articulate speech so that it is easier to lip-read. This paper attempts to quantify the poor performance of visual-only speech recognition and examines its causes.



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

(d) $\mathbf{A}_0$ (e) $\mathbf{A}_1$ (f) $\mathbf{A}_2$

Figure 1: The shape and appearance variation captured by an AAM of a single speaker. The first three shape vectors (a—c) are overlaid on the mean shape, while the first two appearance vectors (e–f) have been scaled for visualization.

2. Methods

High-definition uncompressed video (1920×1080) of five subjects each reciting the 26 letters of the alphabet seven times was recorded using a tri-chip Thomson Viper FilmStream high-definition camera. All video was captured in full-frontal pose and in a single sitting to constrain lighting variation as far as possible. Subjects were asked to maintain a (reasonably) constant head pose and speak in a neutral style (i.e. no emotion). All speakers were fluent in English. The acoustic speech was captured using a boom microphone positioned close to the speaker, but so as not to occlude the face in the video. The video was stored as 50Hz progressive scan and the audio as 16-bit 48-kHz mono. Speakers were asked to begin and end each utterance with closed lips and the video was segmented manually to these points. The segmentation procedure attempts to eliminate visual silence that is commonly much less stable and more difficult to model compared to audio silence. For example, anticipatory coarticulation effects cause the mouth to open or the lips to begin rounding far in advance of the onset of speech.

As in [3] two types of visual feature are tested in our lip-reading system: Active Appearance Model (AAM) parameters, and Sieve features. The *shape*, $\mathbf{s}$, of an AAM is defined by the concatenation of the x and y-coordinates of n vertices that form a two-dimensional triangulated mesh: $\mathbf{s} = (x_1, y_1, \ldots, x_n, y_n)^T$. A compact model that allows a linear variation in the shape is given by,

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

where the coefficients p_i are the shape parameters. Such a model is usually computed by applying principal component analysis (PCA) to a set of shapes hand-labelled in a corresponding set of images. The base shape $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are the (reshaped) eigenvectors corresponding to the m largest eigenvalues. An example shape model is shown in the top row of Figure 1. The *appearance*, $A(\mathbf{x})$, of an AAM is defined by the pixels that lie inside the base mesh, $\mathbf{x} = (x, y)^{\mathrm{T}} \in \mathbf{s}_0$. AAMs allow linear appearance variation, so $A(\mathbf{x})$ can be expressed as a base appearance $A_0(\mathbf{x})$ plus a linear combination of l appearance images $A_i(\mathbf{x})$:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbf{s}_0, \tag{2}$$

where the coefficients 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 [10]. A_0 is the mean shape normalised image and the vectors A_i are the (reshaped) eigenvectors corresponding to the l largest eigenvalues. An example appearance model is shown in the bottom row of Figure 1. The AAM feature vector used to encode each frame is formed by concatenating the appearance parameters after the shape parameters: $(p_1, ..., p_m, \lambda_1, ..., \lambda_l)^T$.

A set of images for each speaker is hand-labelled by placing 73 landmarks on the face, 34 of which model the inner and outer lip contours. There are between and 20 to 36 training frames per speaker. From these images and labels, two AAMs for each individual are built. The first is a low-resolution model built by scaling the base shape, $\mathbf{s}_0$, to contain ≈ 6000 pixels. The second is a full-resolution model, where the base shape retains its natural scale. The low-resolution model is then used to label the entire video sequence automatically using the *project-out inverse compositional AAM search* [11]. The down-sampling of the base shape benefits the fitting by significantly reducing the search space: the model is less likely to become trapped in a local minimum. The output of this initial fit is the approximate landmark locations in all video frames. These positions are then refined by performing a further fit using the full resolution model, where the landmark locations of the coarse fit are used as the starting location in each frame. In practice we have found this two-step procedure provides a more robust solution to a conventional fit, even when a standard multi-resolution search is used.

The second type of feature derives from *sieves*, [12], which are a class of scale-space filters. The one-dimensional variants can be described as a cascade of filters such that the signal at scale s is $x_s = f_s(x_{s-1})$ where x_0 is the original signal and $f_s()$ is a scale-dependent operator and is one of the greyscale opening $\mathcal{O}_s$, closing $\mathcal{C}_s$, $\mathcal{M}_s$, or $\mathcal{N}_s$ operators where $\mathcal{M}_s = \mathcal{O}_s\mathcal{C}_s$, $\mathcal{N}_s = \mathcal{C}_s\mathcal{O}_s$, $\mathcal{O}_s = \psi_s\gamma_s$ and $\mathcal{C}_s = \gamma_s\psi_s$. ψ_s is defined as:

$$\psi_s(x_{s-1}(n)) = \min_{p\in[-s,s]} z_{s-1}(n+p) \tag{3}$$

$$z_s(n) = \max_{p\in[-s,s]} x_{s-1}(n+p) \tag{4}$$

with γ_s *mutatis mutandis* with $\max$ and $\min$ swapped. An important property of sieves, and one which gives them their order-N complexity [13], is that the small scales are processed before the larger ones – they are a cascade with the output from the small scale feeding into the larger scale. In the original literature the morphological operator was replaced with a recursive median filter (the so called m-sieve) but nowadays the variants given above are more common.

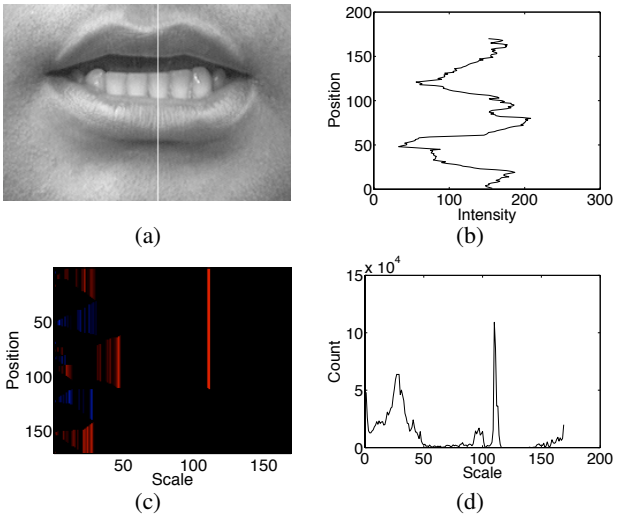


Figure 2: A vertical scan-line from a greyscale version of the mouth sub-image (a) is shown as an intensity plot (b). The granularity spectrum from an m-sieve with positive/negative granules shown in red/blue (c). These granules are then counted, or summed, over all scan-lines to produce the scale-histogram (d).

When applied to lip-reading outputs at successive scales can be differenced to obtain *granule functions* which identify regional extrema in the signal by scale. These difference signals form a scale signature which should change as the mouth opens. The feature extraction system follows that used in [3] and is illustrated in Figure 2. The mouth sub-image is 250×170 pixels and its position inside the full frontal image is determined using the landmarks that delineate the mouth identified by the AAM search. The sub-image is converted to greyscale (Figure 2)(a). Each vertical scan-line (b) is passed through a sieve to create a granularity spectrum (c). These are then summed over all vertical scan-lines to obtain the spectrum (d). Because granules can be positive and negative and have varying scale, there are several options for summing (see [3] for details). Here we test them all and select the best. PCA reduces the dimensionality of the sieve features and experiments have shown that retaining more than the first 20 components does not significantly improve classification performance. Thus, sieve features are transformed by applying PCA to the covariance matrix and retaining the top 20 coefficients.

The audio features are 13 Mel-frequency cepstral coefficients (MFCCs) [14] including the energy. Time derivatives: delta, Δ, and acceleration, $\Delta\Delta$, coefficients are appended to the static features to give a 39-dimensional feature vector. Each

feature vector is calculated from a 20ms window with a 50% overlap (a frame rate of 100 Hz).

3. Results

For classification 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 [3, 1]. The standard HMM toolkit, HTK [15], is applied here for building and manipulating HMMs. All HMMs are models of complete words, where in this context a word is a letter of the alphabet. The models are trained from complete utterances of the word, including any silence before and after the word: there are no separate silence models.

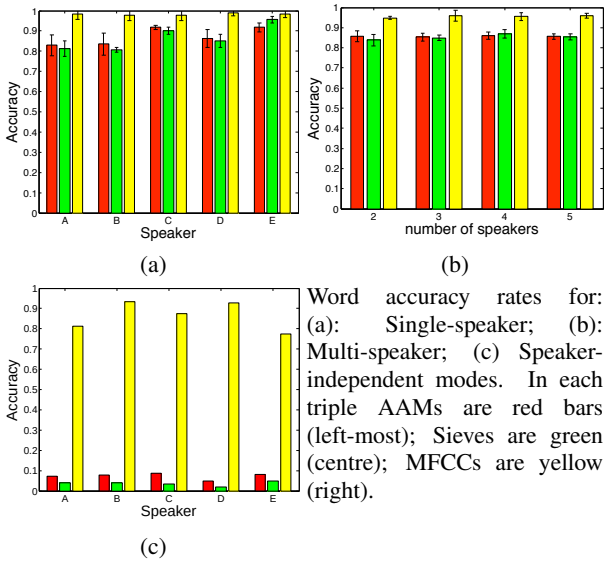
Left-right HMMs with a Gaussian Mixture Model (GMM) associated with each state are used. Because of the small size of the data set for both the acoustic and visual modalities, a diagonal covariance matrix is used for each component of the GMMs. HMMs are initialised using the Viterbi algorithm, via HTK module `HInit`, with a maximum of 20 iterations. Baum-Welch re-estimation is then used (via `HRest`) to refine the HMMs. The number of HMM states and the number of Gaussian mixture components are varied systematically to find the best classifier: the number of states varies from $\{1, 3, 5, \ldots, 15\}$ and the number of mixture components $\{1, 3, 5, 7, 9\}$, giving a total of 40 combinations. Table 1 lists the types of features and HMM parameters used throughout the experiments.

Table 1: Combinations of parameters tested. A total of 160 feature/classifier combinations are tested for AAMs, and 1,440 feature/classifier combinations are tested for Sieve features.

AAM	Sieve	MFCC
mouth model face model	m-, o-, c-sieve sh, a, $\lvert a\rvert$, a^2 histogram	20ms MFCC window 100Hz frame rate
Number of HMM states: 1, 3, …, 15.		
Number of HMM mixture components: 1, 3, …, 9.		
HMM features: f(static feature), $f\Delta$, $f\Delta\Delta$.		

Speaker dependent refers to a mode in which the classifier is trained on speech from a single speaker and tested on different speech from the same speaker. As there are seven repetitions of each letter from each speaker, a seven-fold cross-validation is used with a different example held-out in each fold to allow the computation of the mean accuracy and the standard error. Both an AAM for the complete face and an AAM for the mouth only have been tested. Here only results for the mouth-only AAMs are presented since these always lead to a higher classification performance than full-face models. An explanation for this could be that the full-face model is coding irrelevant information not directly related to speech production, thus introducing ambiguity. All AAM features are normalized to zero mean and unit variance.

Figure 3 (a) shows the mean word accuracy rates, $\bar{a} = 1 - \bar{e}$ where $\bar{e}$ is the mean word error rate with ± 1 standard error for the best performing classifiers using AAMs, Sieves and audio MFCCs for each individual speaker. For AAM features the best configurations used feature vectors of the form $[f, \Delta, \Delta\Delta]$. For speakers C and D seven states are best. All classifiers have 1 mixture component per state. The sieve features show greater classifier variability, with two speakers, C and D, having shortened feature vectors of the form $[f, \Delta]$. The number of HMM states varies from 5 to 9 across speakers with 1 mixture component per state. For the MFCCs the $[f, \Delta]$ form of the feature vector is chosen for speakers A and B whereas the remainder have the $[f, \Delta, \Delta\Delta]$ form. All classifiers have five states and five mixtures apart from Speaker A which has three mixtures per state.

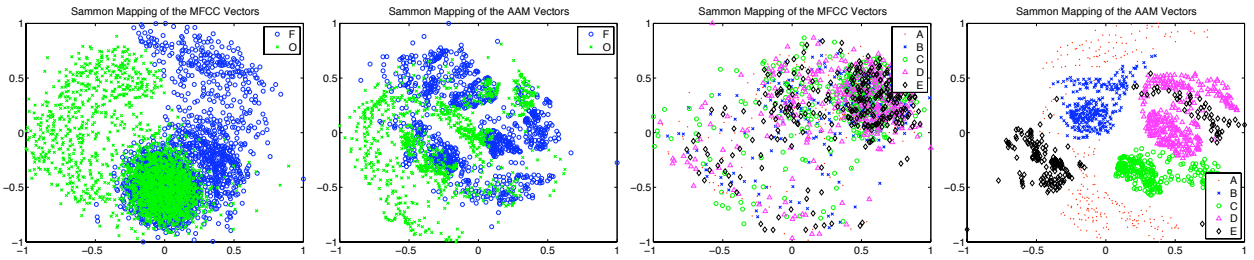


Word accuracy rates for: (a): Single-speaker; (b): Multi-speaker; (c) Speaker-independent modes. In each triple AAMs are red bars (left-most); Sieves are green (centre); MFCCs are yellow (right).

Figure 3: Best accuracy using AAMs, Sieves, and audio-only-MFCCs for (a) each individual speaker respectively and (b) for groups 2, 3, 4 and 5 speakers (the multi-speaker configuration) and for (c) the speaker independent configuration. Where showm the error bars are ± 1 standard error.

Some care is needed when interpreting this variation in classifier parameters across speakers. The performance differences as the parameters alter are not very large, but it is noticeable that there is more variation in the visual classifiers than the audio classifiers. There are noticeable variations across speakers (some individuals are more difficult to lip-read than others) but the visual results are surprisingly good. Many sounds of speech differ only in their voicing and nasality, which generally cannot be seen. In this respect we would expect the visual appearance of the letters “B” and “P” to be almost identical, yet the classifier is disambiguating these letters. This suggests that clustering phonemes into a visually contrastive set of *visemes* using static descriptions of speech gestures, as is commonplace, is incorrect. Whilst the phonemes /b/ and /p/ are both bilablials, so in principle look the same, the dynamics of the articulations are distinct. Hence a static description of a viseme is not informative.

Multi-speaker refers to a mode in which the classifier is trained on speech from several speakers and tested on different speech from the same set of speakers. This approach is widely adopted for evaluating visual speech recognition systems [3]. Figure 3 (b) shows the performance of classifiers trained and tested using data from two (A, C), three (A, C, D), four (A, C, D, E), and five (A, C, D, E, B) speakers. Note that AAMs are built only from the speakers used to train a particular classifier. For example, in the case of three speakers, hand-labelled images of speakers A, C and D are used to construct the AAM. As with the speaker dependent results, the accuracy on the visual-only results is surprisingly good compared to those reported in, for example, [3], although the number of mixture parameters has increased implying that the HMM is modelling the identity of



(a) (b) (c) (d)

Figure 4: Sammon mappings for audio MFCC features (a) and visual AAM features (b) across all speakers. The utterance "F" is shown as blue circles; the utterance "O" is shown as green crosses. The variation by speaker is shown in (c) (audio) and (d) (video) for the single utterance "F". The colours and marker shape vary with the identity of the speaker (A to E)

the speakers.

Speaker independent refers to a mode in which the classifier is trained on speech from several speakers and tested on speech from different speakers. The data-set used in this work contains five speakers, so a five-fold cross-validation is used to test performance, in which a different speaker held out in each fold. As with the previous two experiments, we find that $m-$, $o-$ and $c-$sieves are more or less similar in terms of performance, so the results for only $m-$sieves are reported. A separate AAM is built for each fold, where the images for the test speaker are not included in constructing the model. In this configuration the AAM classifier has nine states, one mixture and an augmented feature vector, $[f, \Delta, \Delta\Delta]$, the Sieve classifier has five states, one mixture and an augmented feature vector in the $[f, \Delta]$ configuration and the MFCC classifier has the $[f, \Delta, \Delta\Delta]$ feature vector with five states and five mixtures per state. Figure 3 (c) shows the test accuracy of each fold for the best performing classifier settings. The maximum mean accuracy rate and corresponding standard error, $(\bar{a}, s_e)$, when using AAM parameters, Sieve features and MFCCs is $(0.21, 0.05)$, $(0.06, 0.01)$ and $(0.87, 0.05)$ respectively. In this configuration the AAM is the best performing visual method. This is probably because the shape component of the AAM allows for scale normalization.

4. Commentary

The drop in performance as one moves to speaker independent recognition is interesting since it implies that previous results, which focussed on single-speaker or multi-speaker systems, have over-estimated the performance of lip-reading systems. Of course, it is always possible to better tune the HMMs to different speakers via feature mean normalisation, Maximum Likellihood Linear Regression (MLLR) adjustement or Maximum a priori (MAP) adjustment as in [16].

The results of a typical MLLR adaptation are shown in Table 2. The conclusions from Table 2 are that the performance on unseen speakers can be quite poor; that the state-of-the-art adaptation algorithm, MLLR, is often not sufficient to compensate for the drop in performance; and that drop in performance seems vary a lot by speaker. It is also worth noting that in Table 2 the best accuracy was reached after using 160, 145, 160, 115 and 160 utterances to adapt the HMM models, which are substantial fractions of the data available.

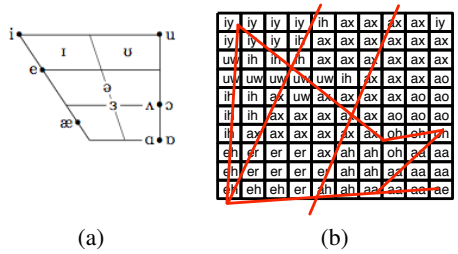
The reason for the significant degradation in performance of the visual-only speaker independent classifiers compared to the acoustic equivalent is apparent in a visualization of the features. Figure 4 (a) and (b) shows a Multidimensional Scaling or Sammon projection [17] of the multidimensional MFCC and AAM vectors into two dimensions for seven repetitions of the letter "F" for all speakers. The MFCCs (Figure 4(a)) do not show any separability by speaker, but for AAM features the separation of the clusters suggests the within-class (speaker) variation is dominating the between-class (letter) variation. Hence, new speakers projected onto the model of visual speech (the AAM or Sieve) will likely form new, distinct clusters so their speech is poorly recognized. Conversely in the acoustic modality the differences between speakers are not apparent. The parameters form two (relatively dispersed) clusters. One cluster is associated with the feature vectors representing silence, the other is associated with the feature vectors representing the utterance of the letter. Acoustic feature vectors from new speakers are likely to fall into the data cloud formed from previously seen speakers, and so are classified correctly, even when there is no training data from the new speaker.

Table 2: Variation in accuracy with speaker showing the mean accuracy over the test data for a multi-speaker system as each speaker is held-out with feature mean normalisation [16] (second column) and the accuracy measured on the held-out speaker (third column). The final column is the best effect of MLLR adjustment over a maximum of 160 utterances (out of 182).

Held-out speaker	Mean accuracy, $\bar{a}$		
	Multi-speaker	Held-out speaker	MLLR
1	0.711	0.0669	0.456
2	0.719	0.0495	0.894
3	0.721	0.132	0.818
4	0.691	0.044	0.324
5	0.740	0.121	0.636

5. Towards Visual-only Continuous Speech Recognition

The results in Section 3 represent *visual-only* isolated word recognition. For continuous recognition sub-word units of speech are required, which for acoustic speech are typically phonemes. In terms of visual speech, phonemes that are visually similar can be clustered into contrastive groups called *visemes*, a term first coined by Fisher [18] as an amalgam of "visual" and "phoneme". However, whilst the phoneme is a well defined and understood unit of speech, the viseme is both poorly defined and ambiguous. There is little agreement as to how many viseme groupings there are for a given language,



(a) (b)

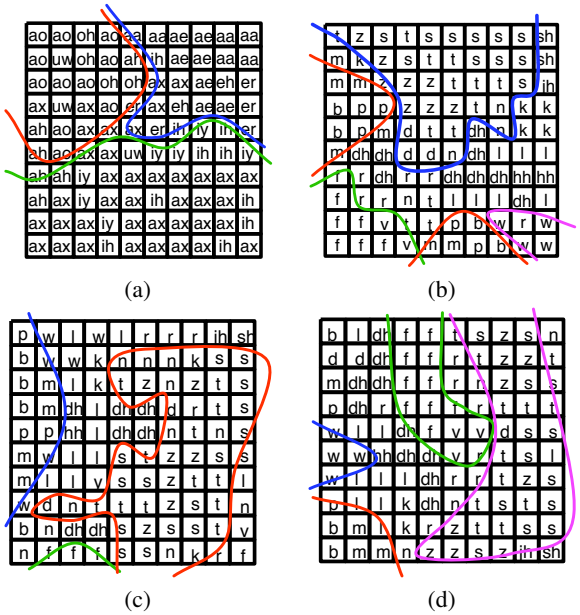
Figure 5: Cardinal vowel diagram (a) showing the position of the tongue in relation to degree of mouth opening for the vowels in RP English. The front of the mouth is to the left of the diagram and symbols appearing to the left of a dot are unrounded, while those to the right are rounded. A SOM (b) for vowels encoded as MFCCs from acoustic data encoding continuous speech. Overlaid onto the map is an approximation of the Cardinal vowel diagram. The labels are the phonetic symbols defined in the BEEP dictionary.

or how a set of phonemes maps to the set of visemes [19, 20, 18, 21, 22, 23, 24, 25, 26].

Traditionally clustering phonemes to visemes is achieved using subjective assessment, where viewers are asked to identify a consonant presented in /VCV/ nonsense syllables. Phonemes that are articulated close to the front of the speech apparatus form well defined groups. For example, /b,p,m/ (bilabials), /f,v/ (labiodentals), /θ,ð/ (interdentals), /t,d,s,z/ (alveolar) and /ʃ,ʒ,tʃ,dʒ/ (palato-aveolar) are often identified as being visually contrastive. However, not all studies agree that these are visemes, and there is little agreement as to how the remaining consonants should be grouped. The vowels themselves do not cluster in the same way consonants cluster, and the visual confusions tend to be arranged evenly across all vowels [22]. Here we briefly consider this in terms of clustering continuous visual speech represented as AAM parameters to determine if a data-driven mapping of visual speech parameters reflects previous studies that have used human subjects.

The self-organizing map (SOM) [27] can be used to cluster and visualise the high-dimensional AAM parameters in a two-dimensional space. Previously ([27] for example), SOMs have been used for visualising acoustic data as phonemes. Here, we use a SOM to form a map of speech parameters, learned from visual data. In particular we are interested in constructing maps for both acoustic and visual speech to compare the clustering in each modality. The map can be related to our knowledge of speech to identify how well defined the space of visual speech is compared with the acoustic space. The following maps were created using the SOM Toolbox for Matlab (http://www.cis.hut.fi/projects/somtoolbox).

Initially, to ensure the mapping is sensible, either perceptually or linguistically, we first extract MFCCs for the vowels in a continuous speech corpus [28] and construct a SOM from these data vectors. The map can then be related to the Cardinal vowel diagram, shown in Figure 5(a). The maps used throughout this work all used the same settings to allow a direct comparison. In particular, the map is a rectangular grid of size 10×10. The training used multi-pass batch training, where the rough training phase used an initial neighbourhood radius equal to the width of the map, which was updated linearly to half the map width. The fine-tuning training re-trained the map with the neighbourhood radius decreasing linearly from half the map width to a single node.



(a) (b) (c) (d)

Figure 6: SOMs trained on the shape and appearance components of an AAM encoding continuous speech. The maps show (a) vowels clustered as AAM parameters, and consonants clusters as (b) shape parameters, (c) appearance parameters, and (d) shape and appearance parameters. Superimposed onto each map is an approximate (manual) segmentation that suggest support for previous studies using human subjects.

The SOM constructed from vowels encoded as MFCCs is shown in Figure 5(b). Each box represents the output of a SOM unit and is labelled with the modal value of the phoneme label associated with the training data. Thus, in the top-left hand corner is a box labelled with /iy/ because this was the most common phonetic label associated with that mapping. The mapping does extraordinary well at capturing the psycho-linguisitc relationship between the vowels. To illustrate the relationship between the Cardinal vowel diagram and the map, an approximation to the vowel space of the cardinal vowels has been superimposed.

Figure 6 shows similar information, but for visual parameters. The map is labelled with the acoustic phoneme symbols corresponding that correspond to the visual frames in the corpus. Following [22], the visual clustering of vowels is much less clear than the audio clustering (Figure 5(a)). The relationship between the vowels appears to be lip-rounding (upper left segments (blue line)), mouth opening (upper right segments (green line)) and mouth closed (bottom of map (magenta curve)). However, in a psycho-linguistic sense, this map is much less structured than that in 5(b).

Generally visemes relate to the *place of articulation*, so are related more consonants than vowels. To test this using a SOM, we construct maps for shape parameters, appearance parameters, and both shape and appearance parameters, shown in Figures 6(b), (c) and (d). There are several notable points of interest in these figures. Most notably, only the most obvious phonemes are grouped as visemes — the bilabials, the labiodentals and the inter dentals and alveolars. Also that the rounded labials are not so well clustered in the appearance only data, which can be explained since the parameters represent shape-free data. Aside from these consonant groups, the remaining groups are ambigu-

ous. This tends to support findings using clusters generated via human subject assessment.

6. Conclusions

This paper has examined the performance of automatic lip-reading using near ideal conditions: very high resolution images, noise-free audio and constrained talkers and vocabulary. In contrast to previous studies we find that the visual-only performance achievable from a single-speaker system is potentially excellent. However, unlike audio, when we attempt to classify video from of a speaker using a visual model trained on other speakers, the performance degrades very significantly. We see this effect on both types of visual model using the same machine learning technique that is known to be successful for audio. There are two explanations, by no means mutually exclusive, for these results. The first is that our current features are not good enough. Better features would encode the information in an utterance independently of the physiology of the speaker and what he or she does with their mouth when they speak. However, even speaker normalisation techniques that have been shown to work in a multi-speaker recognition system fail to bring the performance of visual only recognition to the level of audio recognition. The second is that there is more inherent variability in lip-reading than in speech recognition. This observation is certainly supported by anecdotal evidence from lip-readers, who often report that it takes them some time to "tune in" to a new talker. By contrast, adaptation to a new speaker in speech recognition is almost instantaneous, unless he or she speaks with a very severe accent. We have also briefly presented preliminary findings of data-driven clustering of visual clustering of continuous visual speech data in an attempt to establish a *unit* of visual speech for recognition. Reassuringly, our data-driven finds support previous studies using human subjective assessment. Our main finding is that speaker-dependent or multi-speaker recognition are a red herring since they are unlikely to extrapolate well to reality.

7. References

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

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

[3] 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.

[4] E. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D. dissertation, University of Illinois, 1984.

[5] G. Potamianos, C. Neti, and S. Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *Proceedings of Auditory-Visual Speech Processing*, St. Jorioz, 2003, pp. 95–104.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent developments in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[7] D. Stork and M. Hennecke, Eds., *Speechreading by Humans and Machines: Models, Systems and Applications*, ser. NATO ASI Series F: Computer and Systems Sciences. Berlin: Springer-Verlag, 1996, vol. 150.

[8] I. Almajai, B. Milner, and J. Darch, "Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise," in *Proceedings of INTERSPEECH-2006*, September 2006.

[9] B. Theobald, R. Harvey, S. Cox, G. Owen, and C. Lewis, "Lip-reading enhancement for law enforcement," in *SPIE conference on Optics and Photonics for Counterterrorism and Crime Fighting*, G. Owen and C. Lewis, Eds., vol. 6402, September 2006, pp. 640 205–1–640 205–9.

[10] 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.

[11] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, November 2004.

[12] J. A. Bangham, N. Bragg, and R. W. Young, "Data processing method and apparatus," GB Patent 9512459, June 1995.

[13] J. A. Bangham, S. Impey, and F. Woodhams, "A fast 1D sieve transform for multiscale signal decomposition," in *Signal Processing VII, 'Theories and applications*, G. Holt, Cowan and Sandham, Eds., vol. 7E.9, 1994, pp. 1621–1624.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[15] 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.

[16] G. Potamianos and A. Potamianos, "Speaker adaptation for audio-visual speech recognition," in *Proceedings of Eurospeech*, vol. 3, 1999, pp. 1291–1294.

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

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

[19] C. Binnie, P. Jackson, and A. Montgomery, "Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation," *Journal of Speech and Hearing Disorders*, vol. 41, pp. 530–539, 1976.

[20] K. Finn and A. Montgomery, "Automatic optically-based recognition of speech," *Pattern Recognition Letters*, vol. 8, no. 3, pp. 159–164, 1988.

[21] J. Franks and J. Kimble, "The confusion of English consonant clusters in lipreading." *Journal of Speech and Hearing Research*, vol. 15, pp. 474–482, 1972.

[22] F. Heider and G. Heider, "An experimental investigation of lipreading," *Psychological Monographs*, vol. 52, pp. 124–153, 1940.

[23] P. Kricos and S. Lesner, "Differences in visual intelligibility across talkers," *Volta Review*, vol. 84, pp. 219–225, 1982.

[24] E. Owens and B. Blazek, "Visemes observed by the hearing-impaired and normal-hearing adult viewers," *Journal of Speech and Hearing Research*, vol. 28, pp. 381–393, 1986.

[25] B. Walden, R. Prosek, and A. Montgomery, "Effects of training on the visual recognition of consonants," *Journal of Speech and Hearing Research*, vol. 20, pp. 130–145, 1977.

[26] M. Woodward and C. Barber, "Phoneme perception in lipreading," *Journal of Speech and Hearing Research*, vol. 3, pp. 212–222, 1960.

[27] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[28] B. Theobald, "Visual speech synthesis using shape and appearance models," Ph.D. dissertation, University of East Anglia, Norwich, UK, 2003.

[29] B. Dodd and R. Campbell, Eds., *Hearing by Eye: The Psychology of Lip-reading*. London: Lawrence Erlbaum Associates Ltd., 1987.