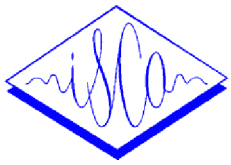
ISCA Archive
http://www.isca-speech.org/archive



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

Limitations of Visual Speech Recognition

Jacob L. Newman, Barry-John Theobald and Stephen J. Cox

School of Computer Sciences
University of East Anglia, Norwich, UK
jacob.newman@uea.ac.uk, b.theobald@uea.ac.uk, s.j.cox@uea.ac.uk

Abstract

In this paper we investigate the limits of automated lip-reading systems and we consider the improvement that could be gained were additional information from other (non-visible) speech articulators available to the recogniser. Hidden Markov model (HMM) speech recognisers are trained using electromagnetic articulography (EMA) data drawn from the MOCHA-TIMIT data set. Articulatory information is systematically withheld from the recogniser and the performance is tested and compared with that of a typical state of the art lip-reading system. We find that, as expected, the performance of the recogniser degrades as articulatory information is lost, and that a typical lip-reading system achieves a level of performance similar to an EMA-based recogniser that uses information from only the front of the tongue forwards. Our results show that there is significant information in the articulator positions towards the back of the mouth that could be exploited were it available, but even this is insufficient to achieve the same level of performance as can be achieved by an acoustic speech recogniser.

Index Terms: automated lip-reading, visual speech recognition, articulatory analysis

1. Introduction

Visual-only speech recognition is known to be a difficult problem [1]. Automated systems generally extract visual features from image regions that contain the mouth and train recognition systems to identify classes of sounds based on the temporal pattern apparent in these features — a process known as automated *lip-reading* [2, 3, 4]. Conversely, humans tend to use more information than is available at just the lips. For example, humans also will use head movements, facial expressions, body gestures and, more importantly, language structure and context to help them identify what was spoken — a process referred to as *speech-reading* [5].

Most of the focus on constructing automated speech recognition systems that utilise visual information has been with respect to *audiovisual* speech recognition [6, 7, 8]. That is, augmenting an acoustic speech recogniser with visual features to improve the robustness of the recogniser to acoustic noise. Few studies focus on developing pure automated *lip-reading* systems, and fewer still focus on developing *speech-reading* systems. The complexity of building an automated system that integrates all of the modalities used by human speech-readers is far beyond the current state of the art. Thus a major limitation of automated lip-reading systems is that not all of the speech articulators are visible, and so the information available to the system is somewhat limited. This means that the differences in articulation for many sounds are where they cannot be seen, and so many sounds appear visually similar on the lips. For example, the place of articulation for the phonemes /b/, /m/ and /p/ is at the lips (these are bilabial stops and require a closure of the lips). The differences between these sounds are in the voicing and the nasality. Similarly /f/ and /v/ are labiodental fricatives and require the lower lip and upper teeth to come into close contact. Again, one of the main differences between these sounds is that /f/ is voiceless whilst /v/ is voiced, a difference that generally cannot be seen at the lips.

It is customary to divide the set of phonemes into visually contrastive groups, referred to visemes (or visual phonemes) [9]. Visual speech recognition then involves constructing models for visemic classes rather than phonetic classes, and using these models to recognise speech from only visual data. There are many problems with this. Firstly, the number of unique words is significantly fewer using a visemic rather than a phonetic transcription. Secondly, there is no standardised set of visemes — many possible mappings of phonemes to visemes have been proposed. Thirdly, the same "viseme" can appear very different because of coarticulation effects (e.g., the /l/ and /n/ in "lean" and "loon" are very different visually yet supposedly have the same visemic label and so the same visual meaning).

In this work we are interested in understanding the limitation of lip-reading systems. That is, *how accurate can a visual speech recognition system be given that not all of the articulators can be seen*? To this end, we construct automatic speech recognisers trained on various articulatory information, and measure how the performance of these systems degrade as information is withheld from the recognisers. We compare the results with a typical automated lip-reading system.

2. Data set and Features

The articulatory features used in this work are drawn from the MOCHA-TIMIT data set [10], which consists of a series of electromagnetic articulography (EMA) measurements for a male and a female speaker. The EMA data are captured at 500Hz and represent the x- and y-positions of eight points on the mid-sagittal plane at the upper and lower lips, upper and lower incisors, tongue tip, blade and dorsum, and the velum.

The video data for lip-reading used in this work are drawn from a custom made data set of 25 speakers reading the United Nations declaration of human rights. The video was recorded using a Sanyo Xacti FH1 camera at 1080p resolution and 60 frames per second. Note, the MOCHA-TIMIT data set does have corresponding video data, but the videos were not captured under 'good' conditions. For lip-reading we require some constraints to be imposed on the speaker(s) — the speaker must be facing the camera so the articulators always are visible, and the head pose should ideally be reasonably constant.

To extract visual speech features for recognition, an active appearance model (AAM) [11] is trained manually from

a few tens of images for each speaker. Our choice of visual feature is motivated by earlier work comparing visual features for lip-reading [12]. We have found that model-based features significantly outperform image-based features (such as eigenlips or discrete cosine transform (DCT) features). To construct an AAM, each image is marked with a number, k, of feature points that identify the features of interest on the face — the inner and outer lip contours in this case. The feature points are normalised for pose (translation, rotation and scale) and are subject to a principal components analysis (PCA) to give a compact model of shape of the form:

$$\mathbf{s} = \overline{\mathbf{s}} + \mathbf{S}\mathbf{b}_s \tag{1}$$

where $\mathbf{s}$ represents a vector of concatenated feature points and $\overline{\mathbf{s}}$ is the mean shape. The columns of $\mathbf{S}$ are the n leading eigenvectors of the covariance matrix defining the modes of variation of the shape, and the shape parameters, $\mathbf{b}_s$, define the contribution of each mode of variation in the representation of $\mathbf{s}$. An example shape model is shown in Figure 1.

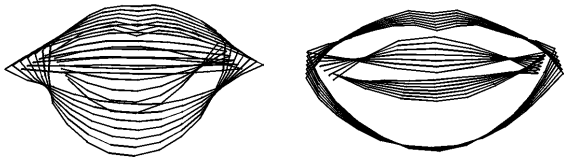


Figure 1: The first two modes of variation of the shape component of an AAM varying between ± 3 standard deviations from the mean. The first mode appears to capture variation due to mouth opening and closing, and the second appears to capture variation due to lip-rounding.

AAMs also allow for appearance variation, where each image is shape normalised by warping from the labelled feature points, $\mathbf{s}$, to the mean shape, $\overline{\mathbf{s}}$. The pixel intensities within the mean shape are concatenated and the resultant vectors are subject to a PCA. A compact model of the appearance variation is given by:

$$\mathbf{a} = \overline{\mathbf{a}} + \mathbf{A}\mathbf{b}_a \tag{2}$$

where $\mathbf{a}$ is a shape-normalised image (i.e. warped onto the mean shape) and $\overline{\mathbf{a}}$ is the mean appearance image. The columns of $\mathbf{A}$ are the m leading eigenvectors of the covariance matrix defining the modes of variation of the appearance, and the appearance parameters, $\mathbf{b}_a$, define the contribution of each mode of variation in the representation of $\mathbf{a}$. An example appearance model is shown in Figure 2.

Figure 2: The mean and first three modes of variation of the appearance component of an AAM. The appearance images have been suitably scaled for visualisation.

To solve automatically for the AAM parameters over an entire video sequence we typically use the inverse compositional project-out algorithm [13] to track the positions of the landmarks defining the AAM shape. The shape and appearance parameters can then be solved by computing the shape parameters using:

$$\mathbf{b}_s = \mathbf{S}^T \left(\mathbf{s} - \overline{\mathbf{s}}\right), \tag{3}$$

then warping from the shape $\mathbf{s}$ to $\overline{\mathbf{s}}$ and computing the appearance parameters using:

$$\mathbf{b}_a = \mathbf{A}^T \left(\mathbf{a} - \overline{\mathbf{a}}\right). \tag{4}$$

These feature vectors are concatenated to give the vector upon which the recognition experiments are based

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_s \\ \mathbf{b}_a \end{bmatrix}. \tag{5}$$

These vectors form a parameter trajectory though the AAM space corresponding to the words spoken by the speaker. To compensate for inter-speaker differences in the parameters, the visual features, which are highly speaker dependent, are z-score normalised per speaker. Inspection of these normalised features show their distribution is approximately Gaussian. We also normalise the EMA x and y points to compensate for physiological differences between the two speakers.

3. Recognition Experiments

In our visual-only recognisers, tied-state mixture triphone hidden Markov models (HMMs) are used to model coarticulation around a central phone, as is typical in state-of-the-art speech recognition systems. The phone level transcriptions required to construct these models for the EMA data are provided with the MOCHA-TIMIT data set, and for the AAM features the phonetic transcriptions were generated from automatically expanded word level transcriptions generated from the acoustic speech. A "flat start" [14] was then applied to the training (visual) data so that the segmentations were not influenced by the acoustic segmentation.

Each of the 44 phone-HMMs is a three-state single Gaussian model. These were replicated to form triphone models for all phone contexts present in the training data. At this stage, there are many triphones with too few examples to train accurate models. To overcome the problem of data sparsity, appropriate states are "tied" together in order to share data between infrequently occurring triphones. Deciding which states to tie is done using hierarchical tree clustering driven by left and right viseme context questions. In a phone recognition system, the most significant coarticulation rules are known from knowledge of the phonetic properties of speech. In this work we use the phone clustering rules provided by the Resource Management tutorial demo [15]. During clustering, rules that do not satisfy state occupancy and likelihood thresholds are ignored, leaving the most appropriate rules for the given parameters. The thresholds we specified retained between 6–7% of the total number of states after tying. Finally, the number of mixture components was increased sequentially from one to eight. To ensure the language model does not influence recognition, the grammar scale factor is set to 0, and to balance the insertion and deletion rates the insertion penalty is set to between -15 and -20 — see the HTK book [14] for more information on these properties of the recogniser.

3.1. Speaker Dependent Articulatory Features

We first establish the maximum accuracy that can be achieved using context-dependent phone recognisers trained on the EMA features from the MOCHA-TIMIT data set. Next, subsets of

these features are formed by removing systematically articulatory features. The degradation in performance of the recognisers can then be measured as a function of the loss of information that results from removing the articulator(s).

First, *speaker-dependent* recognisers are built for the two speakers in the MOCHA-TIMIT data set. A five fold cross validation set up is used, giving 92 test utterances per fold. The features initially consist of all 16 x–y coordinates from all 8 sensors, and then each sensor is removed in turn from the back to the front of the speech apparatus until only one articulator remains. This allows us to simulate the loss of information due to the limited visibility of the articulators toward the back of the mouth during lip-reading. To benchmark the performance against features used in traditional acoustic phone recognition, tied-state multiple mixture speaker-dependent HMMs also are trained and tested using MFCC features extracted from the acoustic speech signal.

3.1.1. Results

The mean phone recognition accuracy of the speaker-dependent recognisers is illustrated in Figure 3. A mean accuracy of 65% is obtained using MFCC features in a speaker dependent recogniser for both the male and the female subjects. This is in contrast to the mean accuracy of only 45% achieved using all articulatory features. This difference in performance is likely attributable to the lack of voicing in the articulatory features — these features measure only the *position* of the articulators. The removal of the EMA sensors gives an understandable reduction in the performance, and seems to show a significant decrease when no tongue back or velum information is present — these sensors may provide cues related to nasality, so when they are removed differentiating nasal from non-nasal sounds becomes difficult. Interestingly, using only the position of the lower lip as a feature for speech recognition provides a mean phone recognition accuracy of 12%, which still is significantly above chance (2.25%).

3.2. Speaker Independent Articulatory Features

To consider the problem in a more general sense, *speaker-independent* recognisers are employed, which are trained and tested using the same five-fold cross validation paradigm described in Section 3.1. That is, a recogniser is trained from the female EMA data, and tested only on the male EMA data, and vice versa. The performance is then averaged over both speakers. The performances of these recognisers are compared to a traditional speaker-independent (visual-only) AAM-based automated lip-reading system trained and tested on AAM features derived from the custom UN data set. The results are shown in Figure 4.

3.2.1. Results

The performance of the speaker-independent recogniser is significantly below that of the speaker-dependent recognisers. Using all sensors provides a mean phone recognition accuracy of only 29% (compared with 45% for speaker-dependent). However, we note the trend of the curves for both Figures 3 and 4 are very similar. A mean accuracy of 14.5% is obtained using the AAM features, with a standard deviation of 2.4. It is interesting to note where the line in Figure 4 marking the performance of the AAM-based recogniser intersects that of the line marking the performance of articulatory features. This appears to be somewhere between when the tongue blade and the tongue tip

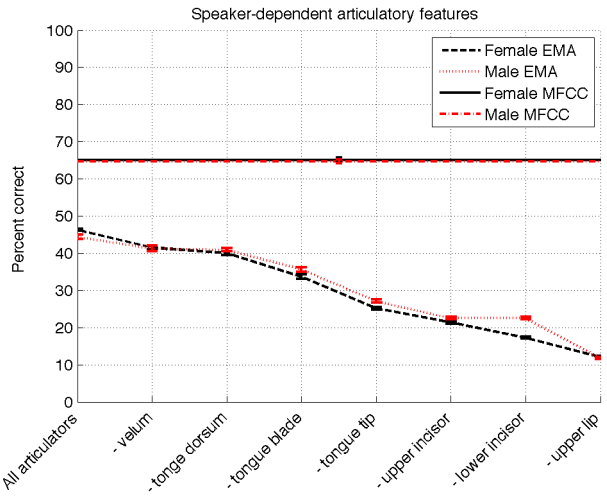


Figure 3: Speaker-dependent phone recognition using articulatory features. The left-most score along the x-axis denotes that all articulators were used by the recogniser. Each subsequent score denotes the removal of an articulator (named on the axis), and the removal is cumulative moving left-to right (e.g. -tongue dorsum indicates that both the velum and the tongue dorsum were not included). The plot shows the mean accuracy averaged over the five-folds, and the error bars denote ± 1 times the standard error. The accuracy (and the trend) for both speakers is very similar, and as would be expected the performance degrades as information is withheld from the recogniser.

are included (with the teeth and lips). This is a striking result that shows just how much information is contained in the sparse EMA features compared with the dense description provided by the AAM. This result also shows that similar information appears to be encoded in both feature sets, and possibly where the upper bound on the expected performance of a visual-only recogniser might be.

4. Summary and Future Work

In this paper we have investigated the limits of automated lip-reading. We first used EMA features to measure the maximum mean phone recognition accuracy that can be achieved using all eight of the sensors in the MOCHA-TIMIT data set. We then systematically removed sensor data to simulate the loss of information due to the rearmost articulators being not visible. The performance of these EMA-based recognisers was compared both with a traditional acoustic speech recogniser and a traditional AAM-based lip-reading system. We found that using all eight articulatory sensors, not surprisingly, achieved the best performance. However, this was significantly below the performance of an acoustic speech recogniser trained on the equivalent acoustic speech, but significantly better than an AAM-based system. We found that the performance of the EMA and AAM-based systems was approximately equal when the articulators in front of the tongue blade were available to the recogniser (i.e. information that can be *seen*). This perhaps suggests the upper bound on the expected performance of pure lip-reading using only visual features.

We note that only eight EMA sensors are included in the MOCHA-TIMIT data set, and that these measure the x–y position only along the mid-sagittal plane. This work would ben-

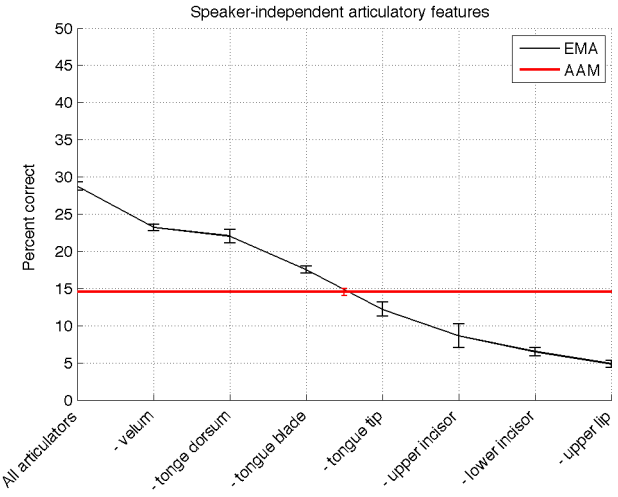


Figure 4: Comparing the performance of speaker-independent phone recognition using articulatory features and AAM features. Note the point of intersection between the two curves. AAM-based features capture the visible shape and appearance information, and perform as well as the EMA features when the articulators from tongue blade/tip forward are available.

efit by repeating the study using more articulatory data. For example, the performance of the recognisers using the frontmost articulators was surprisingly poor — this likely is because only the degree of mouth opening is being measured. Including other sensors would allow more complex mouth gestures to be captured. In removing the sensor information, we only considered removing the sensors one at a time from the back of the mouth forwards. Different combinations of sensors could be tested to determine the most useful set overall. In addition, future work will investigate a more detailed look at the ways the recognisers in the different modalities fail. The results presented here consider only the mean phone recognition accuracy. We might look, for example, at the particular classes of phone the different modalities are most able to accurately recognise. This would allow audiovisual recognisers to be coupled with EMA data, and the relative weight of the different modalities could be adapted according to the available information. This was not done here because we are interested only in investigating lip-reading (i.e. unimodal recognition). Ideally, we would also build a speaker-independent EMA recogniser using a far greater number of speakers than are available in the MOCHA-TIMIT data set. Additional speakers might improve the generalising power of our speaker-independent models as we would be providing the system with more information about the different variations of articulation across a number of people, rather than just one other person.

5. Acknowledgements

The authors gratefully acknowledge EPSRC (EP/E028047/1) for funding. Thanks also to Dr. Philip Jackson at the University of Surrey for his assistance with the MOCHA-TIMIT data set.

6. References

[1] 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.

[2] S. Hilder, R. Harvey, and B. Theobald, "Comparison of human and machine-based lip-reading," in *In the Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, 2009, pp. 86–89.

[3] Y. Nankaku, K. Tokuda, T. Kitamura, and T. Kobayashi, "Normalized training for HMM-based visual speech recognition," *Electronics and Communications in Japan, Part 3*, vol. 89, no. 11, pp. 40–50, 2006.

[4] J. Newman and S. Cox, "Speaker independent visual-only language identification," in *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[5] 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.

[6] I. Matthews, T. Cootes, J. Bangham, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1–16, 2002.

[7] G. Potamianos, C. Neti, G. Iyengar, and E. Helmuth, "Large-vocabulary audio-visual speech recognition by machines and humans," in *In Proceedings of Eurospeech*, 2001, pp. 1027–1030.

[8] 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.

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

[10] A. Wrench, "A new resource for production modelling in speech technology," in *In Proceedings of the Institute of Acoustics (WISP)*, vol. 23, no. 3, 2001, pp. 207–217.

[11] 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.

[12] Y. Lan, R. Harvey, B. Theobald, R. Bowden, and E. Ong, "Visual features for lip-reading," in *In Proceedings of the International Conference on Auditory-visual Speech Processing*, 2009, pp. 102–106.

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

[14] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge: Entropic Ltd., 1999.

[15] 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.