

Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications

Robert Kaucic, Barney Dalton, and Andrew Blake

Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK

Abstract.

Developments in dynamic contour tracking permit sparse representation of the outlines of moving contours. Given the increasing computing power of general-purpose workstations it is now possible to track human faces and parts of faces in real-time without special hardware. This paper describes a real-time lip tracker that uses a Kalman filter based dynamic contour to track the outline of the lips. Two alternative lip trackers, one that tracks lips from a profile view and the other from a frontal view, were developed to extract visual speech recognition features from the lip contour. In both cases, visual features have been incorporated into an acoustic automatic speech recogniser. Tests on small isolated-word vocabularies using a dynamic time warping based audio-visual recogniser demonstrate that real-time, contour-based lip tracking can be used to supplement acoustic-only speech recognisers enabling robust recognition of speech in the presence of acoustic noise.

1 Introduction

Since verbal communication is the easiest and most natural method of conveying information, the possibility of communicating with computers through spoken language presents an opportunity to change profoundly the way humans interact with machines. Voice interactive systems could relieve users of the burden of entering commands via keyboards and mice and prove indispensable in situations where the operator's hands are occupied such as when driving a car or operating machinery. Much research has focused on the development of spoken language systems and rapid advances in the field of automatic speech recognition (ASR) have been made in recent years [7, 23]. Although progress has been impressive, researchers have yet to overcome the inherent limitations of purely acoustic-based systems, particularly their susceptibility to environmental noise. Such systems readily degrade when exposed to non-stationary or unpredictable noise as might be encountered in a typical office environment with ringing telephones, background radio music, and disruptive conversations. Acoustic solutions typically employ noise compensation methods during preprocessing or recognition to reduce the effect of the noise. The preprocessing approaches often use spectral subtraction or adaptive filtering techniques to remove the additive noise from the signal [14]. Hidden Markov Model (HMM) decomposition, where separate models are used for the clean speech and noise, is a common method used to provide compensation during recognition [21, 11]. While these approaches have

proven to be effective, they ignore a basic tenet, that is, the multi-modal nature of human communication. Here we attempt to exploit this by using visual information in the form of the outline of the lips to improve upon acoustic speech recognition performance.

The majority of automatic lipreading research to date has focused primarily on establishing that visual information can be used to supplement acoustic speech recognition on small isolated-word vocabularies. The extraction of features in real-time has been largely ignored by lipreading researchers—deferring that complication to the future. Real-time feature extraction is obviously required for practical audio-visual language understanding systems. However, to track in real-time, it is often necessary to reduce the dimensionality of the image data through parameterisation which could result in the loss of important recognition information. This work demonstrates that, despite this loss of information, visual features obtained from tracking the lips in real-time can supplement automatic acoustic speech recognisers.

Two lip trackers, one that tracks lips from a profile view and the other from a frontal view, have been developed. Both are capable of locating, tracking, and compactly representing the lip outline in real-time at full video field rate (50/60 Hz). The ‘profile lip tracker’ follows the outline of the upper and lower lips and needs no cosmetic assistance. Tracking from the frontal view is more difficult as the lips are set against flesh-tones with consequently weak contrast. Therefore, when the frontal view was used, the speaker wore lipstick to enhance the contrast around the lips. The tracker framework is identical for tracking assisted and unassisted lips and thus for the frontal view, assisted lips were used to demonstrate the feasibility of using real-time dynamic contour-based lip trackers in audio-visual speech recognition applications. Preliminary work has begun on tracking natural lips from the frontal view and a tracking sequence using this tracker is presented as well, although, to date, no recognition experiments have been conducted using it.

Visual features extracted from the lip trackers are incorporated into a dynamic time warping (DTW) based isolated-word recogniser. Recognition performance is evaluated using acoustic only, visual only, and audio-visual information with and without added artificial acoustic noise. The experiments demonstrate that visual information obtained from tracking the lip contour from either view can improve upon acoustic speech recognition, especially in speech degraded by acoustic noise.

2 Lipreading

It is well known that human speech perception is enhanced by seeing the speaker’s face and lips—even in normal hearing adults [9, 17]. Several researchers [9] have demonstrated that the primary visible articulators (teeth, tongue, and lips) provide useful information with regard to the place of articulation and Summerfield [20] concluded that such information conveyed knowledge of the mid- to high-frequency part of the speech spectrum—a region readily masked by background noise.

Motivated by this complementary contribution of visual information, researchers have recently developed audio-visual speech recognisers which have proven to be robust to acoustic noise [15, 22, 19, 5, 6]. These systems can be classified by the visual features they extract into three categories—pixel-based systems, lip-velocity systems, and lip-outline/measurement systems. The pixel-based systems [15, 22, 5] maximise retention of information about the visible articulators by using directly or indirectly the grey-level pixel data around the mouth region. Unfortunately, these systems tend to be highly susceptible to changes in lighting, viewing angle, and speaker head movements. They also usually employ computationally expensive processing algorithms to locate the mouth and/or extract relevant recognition features. While these systems serve as excellent research platforms, the extensive processing required limits their use in real-time or near real-time applications. The lip-velocity systems [12] assume that it is the *motion* of the lips that contains the relevant recognition information especially with respect to determining syllable or word boundaries and thus extract the velocities of different portions of the lips. A similar limitation exists for this approach where computationally expensive procedures like optical flow analysis and morphological operations are used to extract the lip velocities which prevents their use in near real-time applications. The lip outline/measurement systems [10, 19] extract geometrical features from the lip outline or oral cavity. Typical features include the height, width, area, and spreading (width/height) of the mouth. These systems are able to extract visual features in real-time, although they avoid many of the complications of tracking in real-world images by tracking strategically placed reflective dots on the face.

The recognition systems presented here fall into this last category, however, real-time feature extraction is achieved without the need for markers by parameterising the lip outline and learning the dynamics of moving lips.

3 Lip tracker

The lip trackers resulted from the tailoring of Blake et al.’s [2, 3] general purpose dynamic contour tracker to the specific task of tracking lips. The 2D outline of the lips is parameterised by quadratic B-splines which permits sparse representation of the image data. Motion of the lips is represented by the x and y coordinates of B-Spline control points, $(\mathbf{X}(t), \mathbf{Y}(t))$, varying over time. Stability of the lip tracker is obtained by constraining the lip movements to deformations of a lip template, $(\overline{\mathbf{X}}, \overline{\mathbf{Y}})$. Lip motion is modelled as a second order process driven by noise with dynamics that imitate typical lip motions found in speech. These dynamics are learned using a Maximum Likelihood Estimation (MLE) algorithm [4] from representative sequences of connected speech. Temporal continuity is provided by a Kalman filter which blends predicted lip position with measurement/observation features taken from the image. To enable real-time tracking, the search for image features is confined to one dimensional lines along normals to the lip curve. The profile lip tracker uses high contrast edges for image features while the frontal lip tracker uses a combination of edges and intensity valleys.

4 Tracking

The profile view is favourable for tracking because the mouth appears sharply silhouetted against the background whereas, in a frontal view, the lips are set against flesh-tones with consequently weak contrast—a problem for visual trackers. However there is, of course, a potential loss of information in profile viewing in that the tongue and teeth are no longer visible. There may also be a loss of shape information in the lip contour itself, since its width is no longer directly observable in profile, and our experiments suggest that lip width is significant for audio-visual speech analysis. Figure 1 shows that the tracker can follow the lips

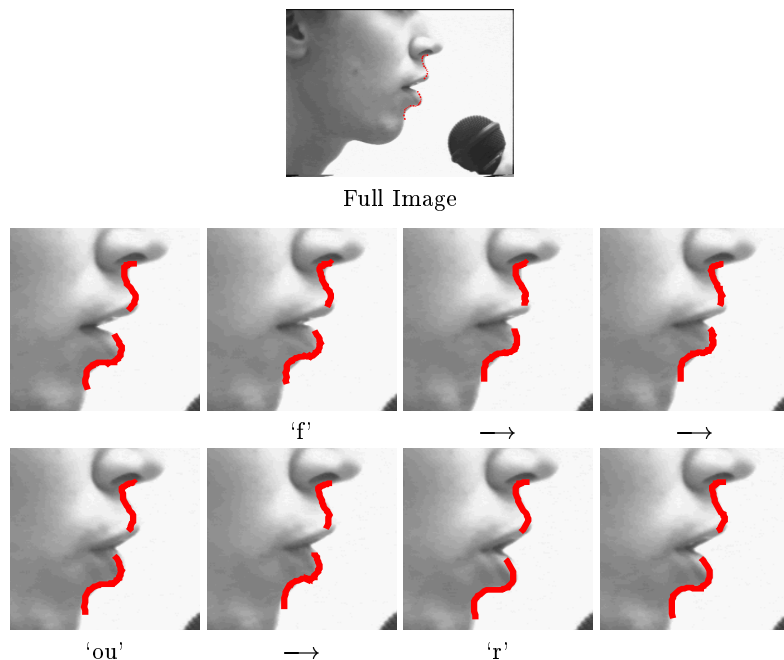


Fig. 1.: *Tracking the word “four”. Snapshots taken approximately every 40 ms. The tracker accurately follows the lower lip during the f-tuck (curling of the lip to form the ‘fa’ sound) in tracked frames 3 and 4 and continues tracking through the lowering of the jaw necessary for the ‘our’ sound.*

even during subtle lip movements such as the f-tuck in the word ‘four’. Similar tracking results were obtained using a frontal view when lipstick was worn to enhance the contrast around the lips [8].

Natural (un-aided) lips can also be tracked from the frontal view using the dynamic contour framework; however, instead of using edges for image features, the intensity valley between the lips is used to locate the corners of the mouth and upper lip. This valley has been shown to be robust to variations in lighting, viewpoint, identity and expression [13] and proves to be a reliable feature for

lip tracking. Figure 2 shows a tracked sequence of the word ‘five’ using valley features for the upper lip and edge features for the lower lip. This mode of

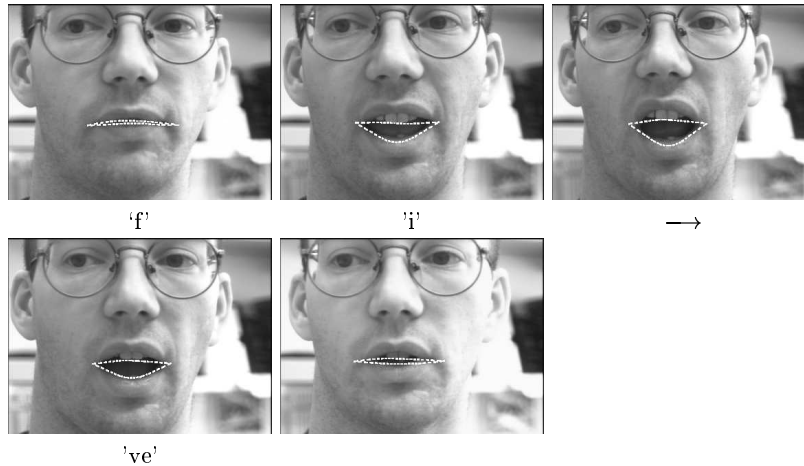


Fig. 2.: *Tracking the word “five” using the valley tracker. Snapshots taken approximately every 60 ms. Whilst tracking is stable and the outline closely approximates the inner mouth region, the upper lip contour becomes confused by the presence of the teeth mistaking them for the inner lip and continues to track them throughout the sequence.*

tracking, being frontal and free of the need for cosmetics, is attractive as a basis for audio-visual analysis. However, there are problems with the system as developed thus far. First, the upper tracked contour has an affinity both for the inside lip and for the teeth when visible, whereas clear differentiation of lips and teeth is a requirement for the application. Secondly, it is difficult to pinpoint mouth corners accurately—the dark visual feature (valley) tends to extend beyond the mouth, resulting in the slightly elongated contour. We know from visual speech recognition experiments (detailed later) that the width of the mouth (oral cavity) contains important recognition information for word discrimination tasks, so further work is needed before this tracker is entirely adequate for speech recognition applications.

Incidental head movements do not affect tracking performance as long as the lip tracker remains locked, however, rapid or large head movements may cause the tracker to lose lock and become unstable. Additionally, since the position of the head naturally influences the position of the lips, head movements may corrupt the recognition data. To compensate for this we are investigating the coupling of a head tracker to the lip tracker [18].

5 Feature Extraction

An essential part of any recognition system is the extraction of features that reliably represent the objects in the data set. The features must compactly represent

the data in a suitable form for recognition. For acoustic speech signals the features are typically the result of spectral analysis on the waveform [16]. Thus the acoustic pre-processing consisted of the extraction of 8 “mel-scale” filter-bank coefficients from overlapping 32ms windows and 20ms frames. The 20ms frame interval was chosen to coincide with the 50 Hz video rate to facilitate integration of the two modalities without additional sub-sampling or linear interpolation.

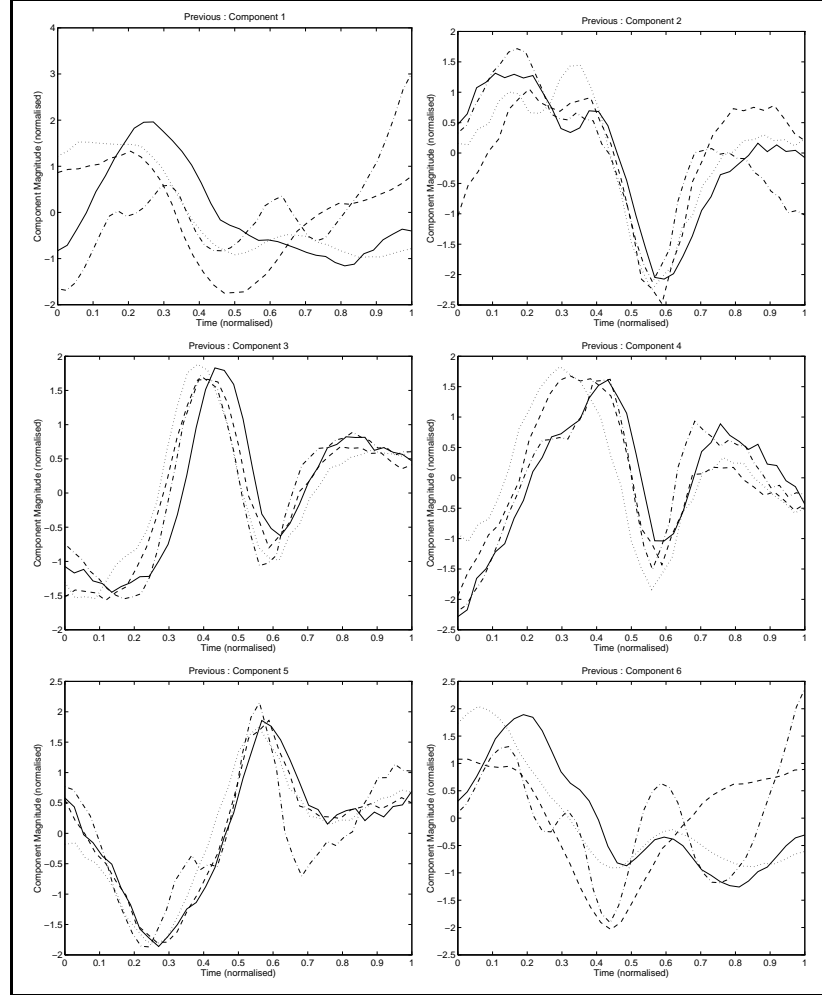


Fig. 3.: *Affine components 1 through 6 for four repetitions of the word ‘previous’. Each component has been thresholded, set to zero mean and unity variance and linear time normalised. Significant shape correlation exists in components 2 (Y translation), 3 (X scale), 4 (Y scale) and 5 (Y shear) across all four repetitions suggesting that they may contain useful recognition information.*

Several different visual processing methods were examined in order to gain insight into which lip movements/deformations would be most beneficial for speech recognition. All visual features resulted from the projection of the lip outline, represented as a sequence of control points, onto a sub-space spanned by a reduced basis. The first basis chosen was the affine basis. Our experience in tracking had shown that the affine basis was insufficient to model all deformations of the lips, but given a successfully tracked lip sequence, we felt that projection onto the affine basis would still provide useful recognition information. Others [10, 19] have reported success using similar features obtained through tracking dots on the face. Additional visual feature representations were obtained through principal components analysis (PCA) where it was found that 99% of the deformations of the outer lip contour were accounted for by the first 6 principal components [8]. Furthermore, since we believed that global horizontal displacement of the lip centroid was not necessary for speech production and only a bi-product of spurious head movements (global vertical displacement is present as a result of the asymmetrical movement of the upper and lower lips), a third recognition basis was created by subtracting horizontal displacement (X translation) from each set of control points and then performing PCA on the remaining data. Similar to the original principal components analysis, it was found that 99% of the remaining lip motions were accounted for by just 6 components.

When choosing features to be used in recognition experiments it is important that the features chosen be repeatable across multiple repetitions of the same token (word), yet be sufficiently different between repetitions of dissimilar words. This was of special concern as Bregler et al. [5] had concluded that the outline of the lip was not sufficiently distinctive to give reliable recognition performance. However, several of the features in the affine basis do in fact satisfy these criteria. This can be seen in figure 3 where traces of the six affine features for multiple repetitions of the word ‘previous’ are shown. In the figure we see that components 2 (vertical translation), 3 (horizontal scale), 4 (vertical scale) and 5 (vertical shear) are consistent across all four repetitions which suggests that they may contain useful recognition information. Similarly, components 1 (horizontal translation) and 6 (horizontal shear) show little consistency which was expected as neither appears to play a role in the production of speech.

6 Recognition Experiments

Both the profile and the frontal lip tracker (with lipstick) were used to explore the extent to which lip contour information could aid speech recognition. Separate isolated-word, audio-visual recognition experiments were conducted using visual features extracted from each of the trackers. Raw visual and audio data were gathered simultaneously and in real-time (50 Hz) on a Sun IPX workstation with Datacell S2200 framestore. The visual data consisted of the mouth outline represented as (x, y) control points (10 for the side view and 13 for the frontal view) and the audio data 8-bit μ -law sampled at 8 KHz.

Recognition experiments were conducted using audio-only, visual only, and combined audio-visual DTW recognition. Composite feature vectors were created

by concatenating the acoustic and vision features, although it was possible to vary the relative weighting between the two modalities during recognition. Each feature was normalised to zero mean and unity variance over the entire frame sequence. The 20 repetitions of each word were partitioned into three sets. Two repetitions were used as exemplar patterns for matching, seven were used as a training set, and eleven as a test set.

6.1 Recognition using the profile view

Although the main experiments were done using the frontal view, it seemed important to run at least a pilot experiment using the side-view, given that tracking in profile is robust even without cosmetic aids. This was done to demonstrate that real-time (50Hz), unaided visual tracking for audio-visual speech analysis is indeed a possibility, albeit currently on a modest scale. A 10-word digit database was used.

Significant improvements in error rate were realised by incorporating the visual data-stream. Figure 4 shows the error rates for experiments conducted with

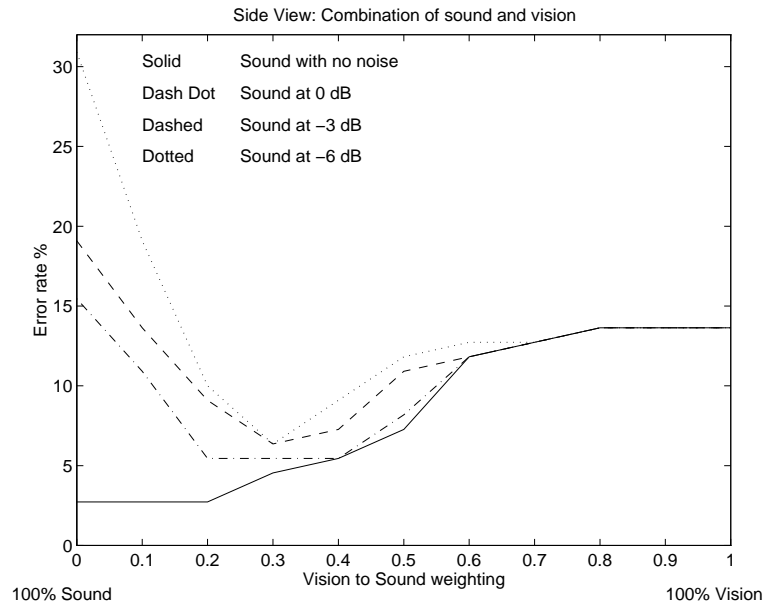


Fig. 4.: *Side View: Error rate variation on the test set as sound to vision weighting is varied. The incorporation of visual features extracted from the lip profile improves recognition performance at all noise levels. With a clean audio signal, vision is only marginally beneficial. However, as the audio signal becomes noisy, the contribution of vision is noticeably improved with a reduction in error rate from 15% to 5.5% for the 0 dB signal and from 19% to 6.5% for -3 dB. With the audio quality further degraded to -6 dB, the error rate drops from 31% to 6.5%.*

audio signals at various SNRs with different vision to sound weightings. Several points are evident from this graph. The first is that the audio-only recogniser performs better than the visual-only recogniser at high signal to noise ratios. This merely reflects the higher information content in audio data with respect to speech recognition in typical noise-free dialogue. Secondly, incorporation of the vision information improved performance at all noise levels—with the largest improvements occurring at the lowest signal to noise ratios—a key finding of this research. It is this increase in recognition performance due to the incorporation of visual information—a term we refer to as the *incremental vision rate*—that is a true measure of the added benefit of lip reading. As an example, one sees that at a SNR of -3 dB, an incremental vision rate (error rate reduction) of 12.5% (19% to 6.5%) is achieved for optimally combined vision/speech compared with speech alone.

6.2 Recognition using the frontal view

In the frontal view, vision data was represented in each of the three previously discussed bases—affine, pca, and pca minus X translation. Experiments were conducted on a 40-word database consisting of numbers and commands that might be used in an interactive voice system controlling a car phone, fax machine, or similar office equipment. Plots of error rates using the frontal lip tracker at various SNRs with different vision to sound weightings were similar to figure 4 in that incorporation of the vision information improved recognition performance at all noise levels [8]. The best error rates for each method of feature extraction are shown in figure 5 on sound at -3 dB SNR. All three bases provide a similar

Best error rates for each basis							
Basis	Acoustic		Visual		Combined		Incremental Vision
	training	test	training	test	training	test	Rate
affine	13.9%	16.6%	44%	52%	8.2%	9.3%	7.3%
PCA	13.9%	16.6%	42%	51%	9.6%	9.3%	7.3%
PCA no X	13.9%	16.6%	41%	49%	9.6%	9.8%	6.8%

Fig. 5.: *Frontal View: Best recognition error rates for the affine, PCA, and PCA without horizontal translation bases on sound at -3 dB SNR. The error rate of the acoustic-only recogniser is nearly twice that of the audio-visual recogniser demonstrating the benefit of incorporating visual information into the acoustic speech recogniser. All three bases provide a similar increase in recognition performance. This is encouraging as the geometrically derived affine basis presents an opportunity for speaker-independent recognition while the PCA bases are particular to a given speaker.*

increase in recognition performance. These results demonstrate that there is useful recognition information contained in the lip outline contrary to Bregler et al. [5] who claims that the outline of the lip is too coarse for accurate recognition. Furthermore, the comparable performance of the affine basis with respect to the derived bases suggests the possibility of developing a speaker independent

recognition system with the visual features represented as affine transformations of the lip template.

6.3 Evaluating visual shape components

Having determined the utility of lip shape information, the recognition performance of individual motion components was measured in order to determine which contribute most to recognition performance. It was hoped that a coherent picture would result yielding the lip movements most beneficial for speech recognition. Figure 6 shows the recognition performance achieved using vision







Best error rates using only a single vision component					
Basis component	Vision only		Combined		Incremental Vision Rate
	training	test	training	test	
Full affine	44%	52%	8.2%	9.3%	7.3%
X Trans	93%	93%	14%	17%	0.0%
Y Trans	76%	81%	13%	14%	3.0%
X Scale	59%	63%	9%	12%	5.0%
Y Scale	75%	79%	14%	16%	0.7%
Y Shear	77%	86%	14%	13%	3.2%
X Shear	86%	90%	14%	17%	0.0%
Full PCA	42%	51%	9.6%	9.3%	7.3%
1 	70%	74%	11%	12%	4.6%
2 	91%	91%	14%	17%	0.0%
3 	82%	88%	14%	17%	0.0%
4 	70%	75%	9%	11%	5.2%
5 	73%	82%	12%	12%	4.6%
6 	89%	94%	14%	14%	0.0%

Fig. 6.: Results of recognition performance using only one vision component from each of the bases. Recognition using sound alone at -3 dB was 14% for the training set and 17% for the test set. Full affine and Full PCA refer to overall recognition performance using all six components of each basis. The lip deformations represented by PCA components 1,4,5 and affine components Y Trans, X Scale, and Y Shear contribute the most to recognition performance implying that the recognition information of the lip outline can be expressed with just a few shape parameters.

components from each of the bases singly. Error rates are shown for the components used individually and in concert with the acoustic features. The tests were conducted on speech at a SNR of -3 dB. These results suggest that most of the recognition information is contained in only a few (2-4) shape parameters.

7 Conclusions and Future Work

Despite doubts expressed by other researchers [5], it has been shown that dynamic contours can be used to track lip outlines, with sufficient accuracy to be useful in visual and audio-visual speech recognition. Moreover, tracking can be performed at real-time video rates (50 Hz). Recognition experiments conducted on a 40-word database demonstrated that isolated words could be accurately recognised in speech severely degraded by artificial noise. Experiments reported here used Dynamic Time Warping as the recognition algorithm; however, given the state of the art in speech analysis [16], it is natural to try Hidden Markov Model recognition. Such experiments are in progress and initial indications are that vision similarly makes a significant contribution to lowering error-rates in accordance with results from others [1, 6].

It is known that human lip-readers rely on information about the presence/absence of the teeth and the tongue inside the lip contour [20]. For this reason it is likely that the best recognition results will ultimately be obtained from frontal views with this additional information extracted. Towards this end, we are developing a real-time un-assisted frontal lip tracker capable of extracting the lip contour as well as determining the presence/absence of the teeth and tongue; furthermore, we are investigating how the coupling of a head tracker to the lip tracker can be used to compensate for global head movements during tracking and recognition [18].

Acknowledgements: The authors wish to express special thanks to Dave Reynard, Michael Isard, Simon Rowe, and Andrew Wildenberg whose assistance and elegant software made this research possible. We are grateful for the financial support of the US Air Force and the EPSRC.

References

1. A. Adjoudani and C. Benoit. On the integration of auditory and visual parameters in an HMM-based ASR. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*. NATO Scientific Affairs Division, Sep 1995.
2. A. Blake, R. Curwen, and A. Zisserman. A framework for spatio-temporal control in the tracking of visual contours. *Int. Journal of Computer Vision*, 11(2):127–145, 1993.
3. A. Blake and M.A. Isard. 3D position, attitude and shape input using video tracking of hands and lips. In *Proc. Siggraph*, pp. 185–192. ACM, 1994.
4. A. Blake, M.A. Isard, and D. Reynard. Learning to track the visual motion of contours. *Artificial Intelligence*, 78:101–134, 1995.
5. C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, pp. 669–672, Adelaide, 1994.
6. C. Bregler and S.M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proc. 5th Int. Conf. on Computer Vision*, pp. 494–499, Boston, Jun 1995.
7. R. Cole, L. Hirschmann, L. Atlas, et al. The challenge of spoken language systems: Research directions for the nineties. *IEEE Trans. on Speech and Audio Processing*, 3(1):1–20, 1995.

8. B. Dalton, R. Kaucic, and A. Blake. Automatic speechreading using dynamic contours. In *Proceedings NATO ASI Conference on Speechreading by Man and Machine: Models, Systems and Applications*. NATO Scientific Affairs Division, Sep 1995.
9. B. Dodd and R. Campbell. *Hearing By Eye : The Psychology of Lip Reading*. Erlbaum, 1987.
10. E. K. Finn and A. A. Montgomery. Automatic optically based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, 1988.
11. M.J.F. Gales and S. Young. An improved approach to the Hidden Markov Model decomposition of speech and noise. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, pp. 233–239, San Francisco, Mar 1992.
12. M.W. Mak and W.G. Allen. Lip-motion analysis for speech segmentation in noise. *Speech Communication*, 14(3):279–296, 1994.
13. Y. Moses, D. Reynard, and A. Blake. Determining facial expressions in real-time. In *Proc. 5th Int. Conf. on Computer Vision*, pp. 296–301, Boston, Jun 1995.
14. J.P. Openshaw and J.S. Mason. A review of robust techniques for the analysis of degraded speech. In *Proc. IEEE Region 10 Conf. on Comp., Control, and Power Engr.*, pp. 329–332, 1993.
15. E.D. Petajan, N.M. Brooke, B.J. Bischofy, and D.A. Bodoff. An improved automatic lipreading system to enhance speech recognition. In E. Soloway, D. Frye, and S.B. Sheppard, editors, *Proc. Human Factors in Computing Systems*, pp. 19–25. ACM, 1988.
16. L. Rabiner and J. Bing-Hwang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.
17. D. Reisberg, J. McLean, and A. Goldfield. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd and R. Campbell, editors, *Hearing By Eye : The Psychology of Lip Reading*, pp. 97–113. Erlbaum, 1987.
18. D. Reynard, A. Wildenberg, A. Blake, and J. Marchant. Learning dynamics of complex motions from image sequences. In *Proc. 4th European Conf. on Computer Vision*, pp. 357–368, Cambridge, England, Apr 1996.
19. D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. In *Proceedings International Joint Conference on Neural Networks*, volume 2, pp. 289–295, 1992.
20. Q. Summerfield, A. MacLeod, M. McGrath, and M. Brooke. Lips, teeth and the benefits of lipreading. In A.W. Young and H.D. Ellis, editors, *Handbook of Research on Face Processing*, pp. 223–233. Elsevier Science Publishers, 1989.
21. A.P. Varga and R.K. Moore. Hidden Markov Model decomposition of speech and noise. In *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, pp. 845–848, 1990.
22. B.P. Yuhas, M.H. Goldstein, T.J. Sejnowski, and R.E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, 78(10):1658–1668, 1990.
23. V. Zue, J. Glass, D. Goodine, L. Hirschman, H. Leung, M. Phillips, J. Polifroni, and S. Seneff. From speech recognition to spoken language understanding: The development of the MIT SUMMIT and VOYAGER systems. In R.P. Lippman, J.E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing 3*, pp. 255–261. Morgan Kaufman, 1991.