SCALE BASED FEATURES FOR AUDIOVISUAL SPEECH RECOGNITION

I A Matthews, J A Bangham and S J Cox

Abstract

This paper demonstrates the use of nonlinear image decomposition, in the form of a sieve, applied to the task of audiovisual speech recognition of a database of the letters A–Z for ten talkers. A scale based feature vector is formed directly from the grayscale pixels of an image containing the talkers mouth on a per frame basis. This is independent of image amplitude and position information and neither accurate tracking or special markers are required. Results are presented for audio only, visual only and for early and late integrated audiovisual cases.

1 Introduction

Previous work has shown [1, 7, 9, 12, 14, 16, 17, 19] that the incorporation of visual information with acoustic speech recognition leads to a more robust recogniser. While the visual cues of speech alone are unable to discriminate between all phonemes (e.g. [b] [p]) they do represent a useful separate channel that can be used to derive speech information. Degradation of one modality, for example interfering noise or cross-talk for audio, or occlusion for video, may be compensated to some extent by information from the other modality. In some cases information in each modality is complementary, e.g. the phonemes [m] [n] vary only by place of articulation and are acoustically similar but visually dissimilar.

There are two fundamental problems in audiovisual speech recognition: visual feature extraction and audio-visual integration. The former can be approached using either model based or data driven methods. Models reduce the dimensionality of the problem to that of the model and allow the direct incorporation of any *a priori* knowledge of visual speech features. However, it is not known which visual features are most useful e.g. lip position/rounding/protrusion, position/presence of teeth and tongue etc. and these features can be difficult to extract reliably. Data driven methods do not require us to explicitly define the visual features as they are automatically learnt by the classifier.

The main problem working directly with images is in how to reduce the dimensionality while retaining as much essential information as possible. This problem has been addressed for example by [7, 8, 17]. In this paper we use a nonlinear image decomposition method, a sieve [2–6], to transform an image into the granularity domain. This completely describes the image in terms of granules that have the attributes position, amplitude and scale. A visual feature vector is formed using only the scale information in an attempt to define a feature that is relatively intensity and position invariant and yet entirely derived from the image itself. This scale-based visual feature vector is used in a Hidden Markov Model (HMM) based recogniser for visual only, early integrated and late integrated audiovisual cases.

2 Sieve Decomposition

A *sieve* or *datasieve* is an algorithm that uses rank or morphological filters to simplify signals over multiple scales, that preserves scale-space causality, and can reversibly transform a signal to a granularity domain. Sieves represent the development of mathematical morphology to form an alternative to wavelet decomposition.

School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK

A sieve is defined by,

$$\Phi_m(X) = \phi(\Phi_{m-1}(X)) \quad \text{where} \quad \Phi_0(X) = X$$

The operator ϕ_m may be an open/close (M sieve) or close/open (N sieve) or a recursive equivalent. In the one-dimensional case, $\Phi_m : Z \to Z$ is based on a series of increasing scale operations. Defining $\phi_m, m = 1, 2, \ldots, m$ as recursive median,

$$\rho_m f(x) = med(\rho_m f(x-m+1), \ldots, \rho_m f(x-1), f(x), \ldots, f(x+m-1))$$

gives the recursive median or R sieve,

$$R_m(X) = \rho_m(R_{m-1}(X)),\ R_0(X) = X$$

the granularity of which is obtained from,

$$Gran_R(X)(m) = (R_m(X)) - (R_{m+1}(X))$$

The granularity consists of the set of *granules G* that represent the non-zero intervals in the granule functions and are characterised by the triplet {position, amplitude, scale}. The sieve transform maps the signal into a set of granules,

$$S : Z \to G$$

The inverse, S^{-1}, may be obtained by summing the re-expanded granules.

3 Visual Feature Extraction

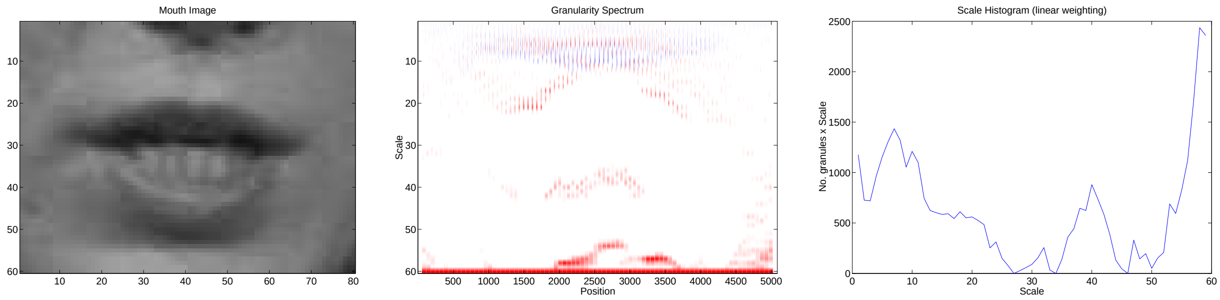
The goal of visual feature extraction is to obtain information from the current image frame that is robust across varying intensity, scale, translation, rotation, viewing angle and talkers. This is an even more difficult task for pixel based systems, but the use of e.g. active shape models [13], deformable templates [11] or dynamic contours [12] to track the lip outline removes the possibility of learning any other visual cues that may be significant.

The one dimensional recursive median sieve defined above may be used to decompose a two dimensional image by scanning in a given direction, e.g. over each column of the image for the vertical case—the direction in which most lip motion occurs during speech. In practice a sieve transform is applied to an entire image in a single pass. The resulting granularity spectrum contains the position, amplitude and scale information of the set of granules that describe the image. As a transform of the original image it contains all the information.

We discard the amplitude attribute of the granules as this largely codes the intensity of the image and we require the visual feature to be independent of intensity variation. The position variation of a granule is dependent on inter-frame differences of the image feature to which it belongs. To use position information would require the identification and tracking of 'interesting' granules (image features), which is counter to the data driven paradigm. The scale parameter is relatively insensitive to intensity variations (until quantisation effects become large) and translation. As the image scene varies between frames (e.g. mouth opens, teeth become visible) the number of granules at any scale in the image will change.

We can now collapse the granularity spectrum, by ignoring position and amplitude information, and form a *scale histogram*. This is the number of granules of each scale summed across the entire image. This also substantially reduces the dimensionality, from the raw image data to that of the maximum scale used in the sieve transform, 60 pixels for this work. Figure 1 demonstrates this process.

Example scale histograms are shown in Figure 2 for the utterance sequence "D-G-M". The top panel shows typical frames from the image sequence, a neutral frame and the centre frame from each utterance. The mouth region was roughly located and the scale histogram of the region obtained. This is plotted as intensity, white represents a large number of granules and scale increases down the y-axis. The number of granules found at a scale clearly change whenever the mouth moves and remains stationary at other times. The bottom panel shows the corresponding audio signal. The utterances are isolated letters and, as expected, the visual cues can be seen to begin before and end after the acoustic signal.



Mouth image Granularity spectrum Scale histogram

Figure 1: Forming a scale histogram. The mouth image is sieve transformed to form a granularity spectrum. Amplitude and position information is removed to form a scale histogram.

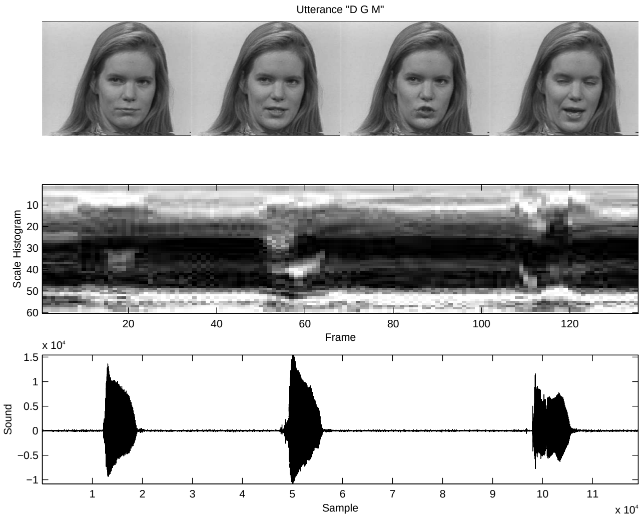


Figure 2: Example scale histogram and audio waveform for the utterance sequence “D-G-M”.

4 Database

An audiovisual database was recorded for ten talkers, five male (two with moustaches) and five female. Each talker repeated each of the letters A to Z three times, a total of 780 utterances. Recording took place in the University TV studio under normal studio lighting conditions. Three cameras simultaneously recorded different views of the talker: full face, mouth only and a side view. All recording was done to tape, the full face to SVHS quality. The output of a high quality tie clip microphone was adjusted for each talker through a sound mixing desk and fed to all video recorders.

An autocue presented the letters A to Z three times in a non-sequential order. Each talker was asked to return their mouth to the neutral position after each utterance and allowed to simply watch the autocue. No attempt at restraining was made but talkers were asked not to move their mouth out of frame of a mouth close up camera.

For this work only the full face data has been used. All 780 utterances have been digitised at quarter frame PAL (376×288) resolution and full frame rate (25Hz) using the standard frame grabber hardware of a Macintosh Quadra 660AV. All video was digitised to 8-bit grayscale. Audio was simultaneously digitised using 16-bit resolution at 22.05kHz to ensure audiovisual time alignment.

Each utterance movie was hand segmented using the video channel so that each image sequence began and ended with the talkers mouth in the neutral position. The audio data within this window was then hand labelled as silence–letter–silence.

5 Unimodal Recognition

The oral region was manually extracted from each of the utterance movies of the database. This was done by positioning a window of 80×60 pixels centrally on the mouth image of the middle frame of each image sequence. Although there is some head motion of the talkers the mouth always stays within this region.

Scale histograms were generated for all utterance movies by applying a vertical one dimensional recursive median sieve to each frame. The vertical dimension of all images is 60 pixels so this is the dimensionality of each scale histogram. All recognition experiments were performed using the first two utterances from each of the ten talkers as a training set (20 training examples per utterance) and the third utterance from each talker as a test set (10 test examples per utterance). Classification was done using 10 state, left to right HMMs with each state associated with a single Gaussian density with a diagonal covariance matrix. All HMMs were implemented using the HTK Hidden Markov Model Toolkit V1.4.

To further reduce the size of the visual feature vector principle component analysis (PCA) was applied to the entire data set. Experiments have used 10 (accounting for 95% of variance) and 20 (99%) coefficients. Further experiments have compared simple averaging of adjacent values and using scale histograms generated from zero padded images (this has the effect of preserving the largest scale, i.e. 'dc' values) and performing PCA using both covariance and correlation matrices (to ignore or imply equal importance of all variables). All visual results were obtained using visual features interpolated from 40ms to 20ms frame rates.

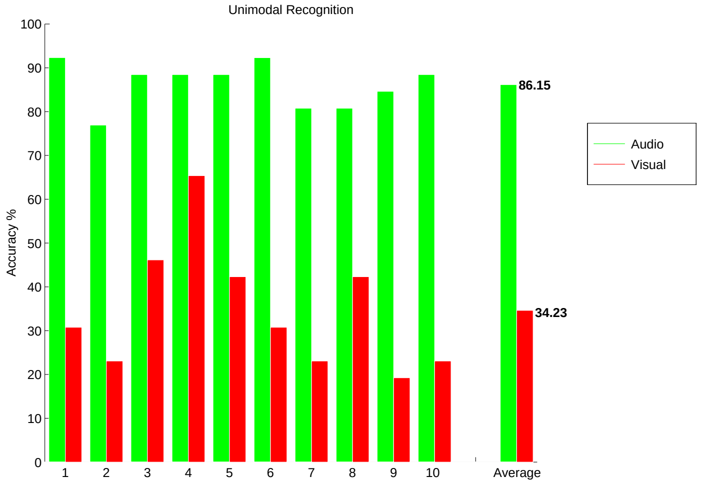


Figure 3: Unimodal results for each talker. Overall accuracy: visual only 34.23%, audio only 86.15%.

Figure 3 shows the results for the best visual case (20 PCA coefficients calculated using the covariance matrix of non-dc preserved scale histograms) on a per-talker basis with audio only results for comparison. The audio features consisted of 12 MFCC coefficients plus an energy term and delta coefficients, i.e. 26 coefficients, calculated at a frame rate of 20ms. The same HMM topology and training/test data was used for audio only as visual only tasks.

The variation in visual only performance (20–65% correct) is clearly much greater than that for audio only (77–92% correct), and there is little correlation between poor audio and poor visual performance.

6 Bimodal Recognition

The two extremes for bimodal integration are the early and late methods. For early integration a single classifier is used with composite audiovisual features. In contrast late integration combines the result of two independent classifiers, one audio, one visual. Further methods have been proposed by [15] and implemented by [1] which extend both these methods by adding a recoding step. We have implemented an example of both early and late integration methods.

6.1 Early Integration

For early integration a composite audiovisual feature vector was formed by concatenating audio and visual features to form 46 coefficient audiovisual features. The HMM topology described for unimodal recognition was used and Figure 4 shows the recognition accuracy results for audio only, visual only and early integrated audiovisual tasks over varying signal to noise ratios (SNR). The appropriate amount of Gaussian white noise was added to take the audio utterances from clean (25–30dB) to the desired SNR.

A problem with early integration is how to segment the utterances, our database is labeled for audio and visual data independently to allow the inclusion of the visual articulation that occurs before and after the acoustic utterance. Results using both methods are shown in Figure 4 and the longer visual segmentation gave improved results over audio segmentation. Both methods improve over audio only results but below 20dB accuracy falls below that of visual alone.

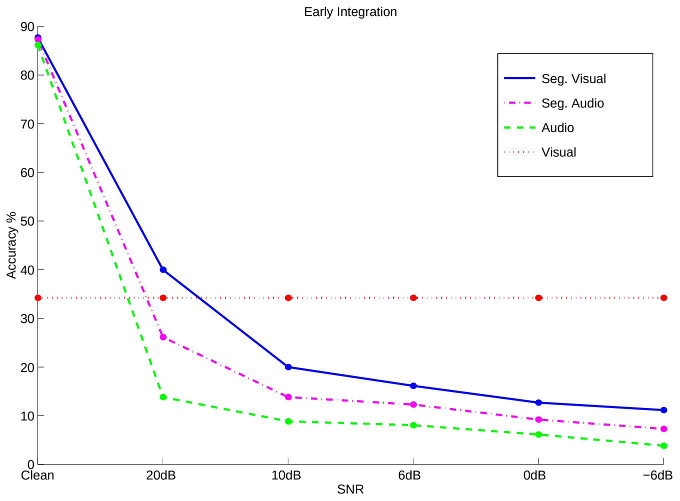


Figure 4: Early integration results.

6.2 Late Integration

For late integration a method of combining the output of two independent audio and visual classifiers must be found. One method for resolving disagreement is to assign the output to the classifier which is most ‘confident’. A confidence measure may be formed by examining the normalised log-likelihoods of each classifier. A simple method is to form the ratio of the maximum normalised log-likelihood and the next highest, if this ratio is high the next best candidate for classification was much lower than the best candidate. The output is then assigned to the classifier with the highest ratio.

$$output = \begin{cases} A & \text{if } A_1/A_2 > V_1/V_2 \\ V & \text{if } A_1/A_2 < V_1/V_2 \end{cases}$$

The results in Figure 5 show that for low SNR late integration gives poor results compared to audio alone, this is due to the visual classifier, by this confidence measure, being confidently wrong on several occasions. At higher SNR performance was improved over early integrated audiovisual, but falls below visual only at 10dB.

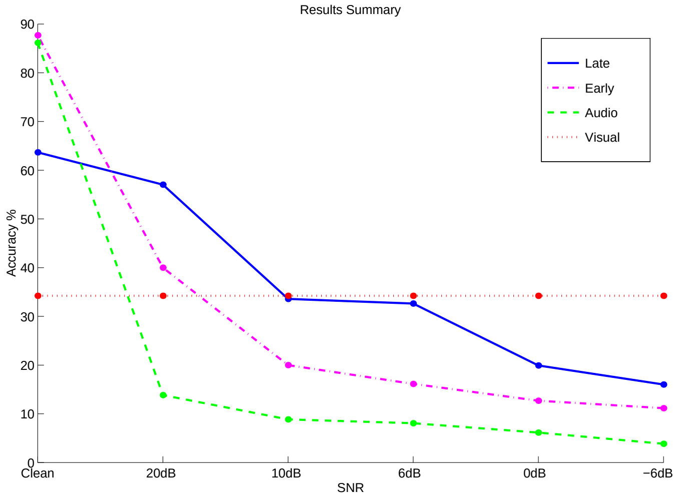


Figure 5: Late integration results.

7 Conclusions

These results indicate that the scale histogram visual speech feature vector can be successfully used in visual speech recognition. A recogniser using only visual information attained an average performance of 34% across ten talkers in a multi-talker, isolated letter recognition task. We have also demonstrated early and late integration methods of combining this information with standard audio features to improve recognition under noisy conditions.

Future work will focus on finding more effective ways of integrating the audio and visual information with the aim of ensuring that the combined performance is always at least as good as the performance using either modality [1, 15, 16, 19] and in deriving more discriminative features from a scale histogram to increase robustness across talkers.

Work is in progress to determine if a scale histogram formed using a two dimensional sieve [3] can also be used to form visual features. The two dimensional sieve is implemented as described in section 2 but the granules now describe 2-D extrema, i.e. image patches. This may form more physically meaningful scale histograms as each granule can be readily interpreted as an image feature.

We are also investigating the use of active shape models [10, 13] as this will allow us to directly compare our pixel based method with a model based approach. The use of a model/pixel hybrid method would allow us to normalise a scale histogram with an estimate of the scale of the mouth region and so form features robust to scale (head motion to and from the camera). This would also provide a method of automatically locating the mouth region, although this method does not require accurate tracking of the mouth it must remain within the image frame.

Acknowledgements

The first author is funded by a research studentship from the EPSRC. The authors would like to thank Dr Richard Harvey for many useful discussions.

References

[1] A. Adjoudani and C. Benoît. On the integration of auditory and visual parameters in an HMM-based ASR. In Stork and Hennecke [18], pages 461–471.

[2] J. A. Bangham, P. Chardaire, C. J. Pye, and P. D. Ling. Mulitscale nonlinear decomposition: The sieve decomposition theorem. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(5):529–539, 1996.

[3] J. A. Bangham, R. Harvey, P. Ling, and R. V. Aldridge. Morphological scale-space preserving transforms in many dimensions. *Journal of Electronic Imaging*, 5(3):283–299, July 1996.

[4] J. A. Bangham, R. Harvey, P. Ling, and R. V. Aldridge. Nonlinear scale-space from *n*-dimensional sieves. *Proc. European Conference on Computer Vision*, 1:189–198, 1996.

[5] J. A. Bangham, P. Ling, and R. Harvey. Scale-space from nonlinear filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(5):520–528, 1996.

[6] J. A. Bangham, P. Ling, and R. Young. Mulitscale recursive medians, scale-space and transforms with applications to image processing. *IEEE Trans. Image Processing*, 5(6):1043–1048, 1996.

[7] C. Bregler, S. M. Omohundro, and Y. Konig. A hybrid approach to bimodal speech recognition. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 556–560, Pacific Grove, CA, Nov. 1994.

[8] N. M. Brooke and S. D. Scott. PCA image coding schemes and visual speech intelligibility. *Proc. Institute of Acoustics*, 16(5):123–129, 1994.

[9] N. M. Brooke, M. J. Tomlinson, and R. K. Moore. Automatic speech recognition that includes visual speech cues. *Proc. Institute of Acoustics*, 16(5):15–22, 1994.

[10] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. The use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12(6):355–366, 1994.

[11] M. E. Hennecke, D. G. Stork, and K. V. Prasad. Visionary speech: Looking ahead to practical speechreading systems. In Stork and Hennecke [18], pages 331–349.

[12] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In B. Buxton and R. Cipolla, editors, *Proc. European Conference on Computer Vision*, volume II of *Lecture Notes in Computer Science*, pages 376–387, Cambridge, Apr. 1996. Springer-Verlag.

[13] J. Luettin, N. A. Thacker, and S. W. Beet. Visual speech recognition using active shape models and hidden markov models. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 817–820, Atlanta, GA, May 1996. IEEE.

[14] E. D. Petajan, B. J. Bischoff, D. A. Bodoff, and N. M. Brooke. An improved automatic lipreading system to enhance speech recognition. Technical Report TM 11251-871012-11, AT&T Bell Labs, Oct. 1987.

[15] J. Robert-Ribes, M. Piquemal, J.-L. Schwartz, and P. Escudier. Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In Stork and Hennecke [18], pages 193–210.

[16] P. L. Silsbee. Audiovisual sensory integration using hidden markov models. In Stork and Hennecke [18], pages 489–496.

[17] P. L. Silsbee and A. C. Bovik. Medium vocabulary audiovisual speech recognition. In *New Advances and Trends in Speech Recogntion and Coding*, pages 13–16. NATO ASI, 1993.

[18] D. G. Stork and M. E. Hennecke, editors. *Speechreading by Humans and Machines: Models, Systems and Applications*, volume 150 of *NATO ASI Series F: Computer and Systems Sciences*. Springer-Verlag, Berlin, 1996.

[19] M. J. Tomlinson, M. J. Russell, and N. M. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 821–824, Atlanta, GA, May 1996. IEEE.