AUDIOVISUAL SPEECH RECOGNITION USING MULTISCALE NONLINEAR IMAGE DECOMPOSITION

Iain Matthews, J. Andrew Bangham and Stephen Cox

School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK
{iam,ab,sjc}@sys.uea.ac.uk

ABSTRACT

There has recently been increasing interest in the idea of enhancing speech recognition by the use of visual information derived from the face of the talker. This paper demonstrates the use of nonlinear image decomposition, in the form of a 'sieve', applied to the task of visual speech recognition. Information derived from the mouth region is used in visual and audiovisual speech recognition of a database of the letters A-Z for four talkers. A scale histogram is generated directly from the grayscale pixels of a window containing the talkers mouth on a per frame basis. Results are presented for visual-only, audio-only and in a simple audiovisual case.

1. INTRODUCTION

It has already been shown [1, 6, 8, 10, 13, 15, 16, 17] 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, e.g. 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 the 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.

A major problem in audiovisual speech recognition is that of visual feature extraction. There are two general methods for extracting these features; model based or data driven. The model based approach immediately reduces the dimensionality of the problem to that of the model and allows the direct incorporation of any *a priori* knowledge of visual speech features. The problem arises that we do not know exactly which visual features to use, e.g. lip position/rounding/protrusion, presence and position of teeth, tongue etc. and these features are difficult to extract because they require the use of complex tracking algorithms. Using data driven methods we do not have to explicitly define the visual features as they are automatically learnt by the classifier. The main problem with images lies in how to reduce the dimensionality while retaining as much essential information as possible. This problem has been addressed for example by [6, 7, 15].

In this paper we use a nonlinear image decomposition method, a sieve [2, 3, 4, 5], to decompose an image into the granularity domain. This completely describes the image in terms of granules that have the attributes scale, position and amplitude. The 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 Hidden Markov Model recognition for the visual only case (computer speechreading) and in a simple implementation of an audiovisual speech recogniser (AVSR).

2. SIEVE DECOMPOSTION

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.

A sieve is defined by,

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

The operator ϕ_m may be an open/close (*M sieve*) or close/open (*N sieve*) or a recursive equivalent.

For the one dimensional case that will used in this paper, $\Phi_m : Z \to Z$ is based on a sequence of increasing scale operations ϕ_m, $m = 1,2,...,m$. Defining ϕ_m as a recursive median,

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

gives the recursive median, or *R sieve* used here,

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

The *granularity* of a signal is obtained from,

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

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 {scale, amplitude, position}. 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 about 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 [11] or deformable templates [9, 10] to track the lip contours removes the possibility of learning any other visual cues that may be significant.

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

We must 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 of a given scale in the image will change.

One way of reducing the dimensionality of the transformed image is by forming a *scale histogram*. The number of granules of each scale is summed across the entire image. This provides a simple method of substantially reducing the dimensionality of the raw image data to that of the maximum scale used in the sieve, namely 60 pixels.

Example scale histograms are shown in Figure 1. The top panel shows typical frames from the image sequence. 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. The number of small scale granules at the top of the panel clearly change whenever the mouth moves. The bottom panel shows the corresponding sound signal. Figure 2 shows a single utterance by a different talker in more detail. The utterances are isolated letters and, as expected, the visual cues can be seen to begin before and end after the acoustic signal.

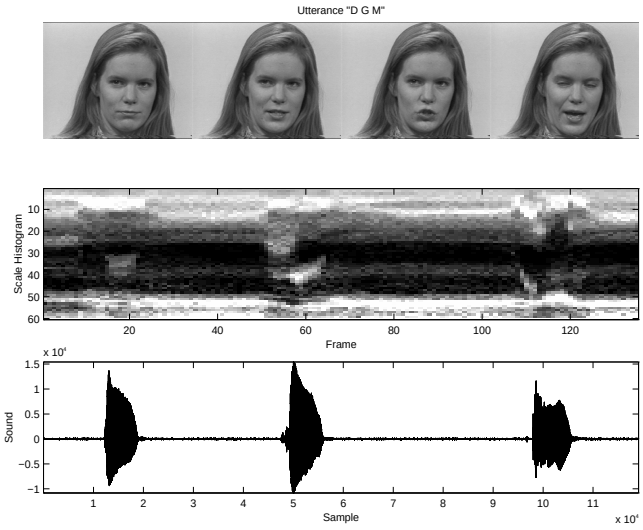

Figure 1: Top: Four images of the utterance "D G M". Middle: The scale histogram. Bottom: The audio signal.

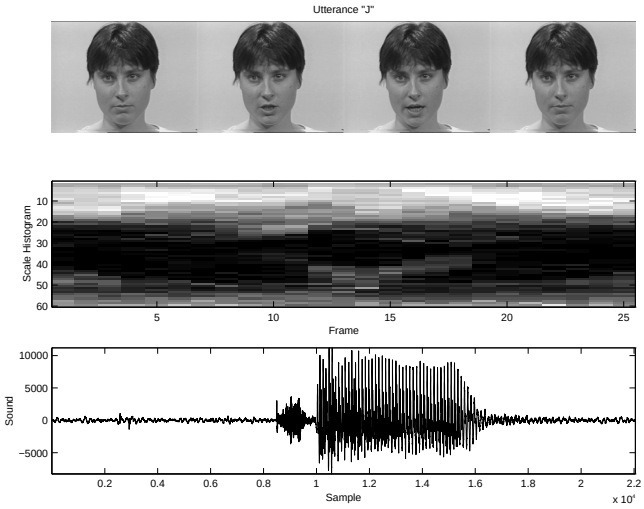

Figure 2: Top: Four images (first, last and two between) of the utterance "J". Middle: The scale histogram, showing variation as teeth appear, lips become rounded etc. Bottom: The audio signal.

4. DATABASE

An audiovisual database was recorded of four talkers, two male, two female, all final year undergraduate students and native British English speakers. Each talker repeated each of the letters A to Z three times, a total of 312 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 video tape, the full face recording 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 the head was made but subjects were asked not to move their mouth out of frame of the mouth close up camera.

For this work only the full face data has been used. All 312 utterances have been digitised at quarter frame PAL (376×288) resolution and full PAL frame rate (25Hz) using the standard frame grabber hardware of a Macintosh Quadra 660av. All video was digitised using 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 mouth in the neutral position. The audio data within this window was then hand-labelled as SILENCE-LETTER-SILENCE.

5. VISUAL RECOGNITION

The oral region was manually extracted from each of the utterance movies of the database. This was done simply 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. In other experiments this stage has been automated by using an area sieve to find the face.

Scale histograms were generated for all utterance movies by applying a vertical one dimensional sieve to each frame. The vertical dimension of all images is 60 pixels so this is the dimensionality of each scale histogram. This is still too large to use as a visual feature vector and two methods were applied to reduce this further. All recognition experiments were performed using the first two utterances from each of the four talkers as a training-set (8 training examples per utterance) and the third utterance from each talker as a test-set (4 test examples per utterance).

The first data reduction method used was simply to reduce the 60 coefficients to 30 by averaging adjacent coefficients. This was still too many to integrate into an audiovisual recogniser but provided a benchmark for visual-only recognition. The 30 coefficients formed the observation vector for a 10 state, left to right hidden Markov model (HMM). Each state of the HMM was associated with a single Gaussian density with a diagonal covariance matrix.

To further reduce the size of the visual features principle components analysis (PCA) was applied to the entire data set. The top ten eigenvectors were then used as the visual feature vector for the HMM topology used for the previous experiment. Results for the two experiments are given in Table 1.

Although there is some degradation in performance when a reduced set of PCA coefficients is used it was decided that dimensionality reduction / performance trade-off was adequate for an initial investigation. Improved extraction of visual features from scale histograms is currently under investigation.

Method	Recognition Accuracy
Original 30 values	50.00%
Top 10 PCA values	43.27%

Table 1: Visual-only recognition performance for isolated letters across four talkers.

6. AUDIOVISUAL RECOGNTION

A simple audiovisual recognition experiment was performed by forming concatenated audiovisual feature vectors. The audio features consisted of 12 MFCC coefficients plus an energy term plus the delta coefficients, i.e. 26 coefficients, calculated at a frame rate of 20ms. Because visual features are formed at a frame rate of 40ms, adjacent visual vectors were interpolated with an extra vector. The visual vectors were then concatenated with the audio vectors to form a composite audiovisual vector at a frame rate of 20ms.

The same HMM topology was used as described in section 5 and again the composite feature vectors of the first two utterances from each talker were used for training, the third for testing. Models were also tested with different weightings applied to the audio and visual parts of the feature vector. All models were trained on data with no added noise and tested on data with varying acoustic signal to noise ratios. The SNR of the audio utterances varied, the highest being about 30dB and the lowest 20dB. An appropriate amount of Gaussian white noise was added to make all audio signals have the same SNR for a given experiment.

To provide a benchmark, an audio-only recogniser was built using the HMM topology described in section 5. However, because the database had been segmented on the visual data, a silence model was used to model the noise before and after the utterance.

Figure 3 shows that when no extra audio noise is added, the audiovisual recognition performance is improved by about 10% over audio-only and about 40% over visual-only. However, for SNR's of 20dB–10dB, audiovisual performance is inferior to performance from both audio-only and visual-only. At very low SNR, (6dB and below), audiovisual performance is marginally superior to audio-only performance but remains well below the performance obtainable from visual-only. The fall off in performance of the audiovisual models when compared with audio-only may be attributed to the fact that when the audio and visual vectors are concatenated, sections of silence before and after the utterance cannot be modelled by a separate silence model (as they are in the audio-only model).

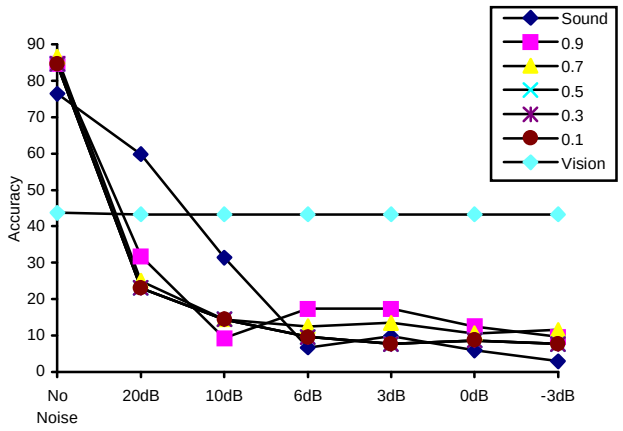

Figure 3: Audiovisual recognition results for various sound to vision weightings and signal to noise ratios.

7. CONCLUSIONS

These results indicate that the scale histogram visual speech feature vector can be successfully used in an audiovisual speech recognition system. A recogniser using only visual information derived from a scale histogram attained a performance of 50% accuracy on an isolated letter task. Using combined audio and visual features, recognition performance was improved by a maximum of 10% at high and low SNR's over an audio-only recogniser. Future work will focus on finding more effective ways of combining 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, 14, 16, 17] and in deriving more discriminative features from the scale histogram.

8. REFERENCES

1. Adjoudani, A. and Benoît, C. "On the Integration of Auditory and Visual Parameters in a HMM-based ASR", *Proc. NATO ASI Conference on Speechreading by Man and Machine, 1995, in press.*

2. Bangham, J. A., Chardaire, P., Pye, C. J. and Ling, P. "Multiscale Nonlinear Decomposition: The Sieve Decomposition Theorem", *IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 18, No. 5, 1996.*

3. Bangham, J. A., Ling, P. and Young, R. "Multiscale Recursive Medians, Scale-Space and Transforms with Applications to Image Processing", *IEEE Trans. Image Processing, Vol. 5, No. 6, 1996.*

4. Bangham, J. A., Ling, P. and Harvey, R. "Scale-Space From Nonlinear Filters", *IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 18, No. 4, 1996.*

5. Bangham, J. A., Harvey, R., Ling, P. and Aldridge, R. V. "Nonlinear Scale-Space from *n*-Dimensional Sieves", *Proc. European Conference on Computer Vision, Vol. 1, 1996, pp. 189-198.*

6. Bregler, C., Omohundro, S. M. and Konig, Y. "A Hybrid Approach to Bimodal Speech Recognition", *28th Asilomar Conference on Signals, Systems and Computers, Vol. 1, 1994, pp. 556-560.*

7. Brooke, N. M. and Scott, S. D. "PCA Image Coding Schemes and Visual Speech Intelligibility", *Proc. of the Institute of Acoustics, Vol. 16, No. 5, 1994, pp. 123-129.*

8. Brooke, N. M., Tomlinson, M. J. and Moore, R. K. "Automatic Speech Recognition That Includes Visual Speech Cues", *Proc. of the Institute of Acoustics, Vol. 16, No. 5, 1994, pp. 15-22.*

9. Hennecke, M. E., Prasad K. V., Stork, D. G. "Using Deformable Templates to Infer Visual Speech Dynamics", *28th Asilomar Conference on Signals, Systems and Computers, Vol. 1, 1994, pp. 578-582.*

10. Kaucic, R., Dalton, B. and Blake, A. "Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications", *Proc. European Conference on Computer Vision, Vol. II, 1996, pp. 376-387.*

11. Luettin, J, Thacker, N. A. and Beet, S. W. "Visual Speech Recognition Using Active Shape Models and Hidden Markov Models", *Proc. ICASSP, Vol. 2, 1996, pp. 817-821.*

12. Mase, K. and Pentland, A. "Automatic Lipreading by Optical-Flow Analysis", *Systems and Computers in Japan, Vol. 22, No. 6, 1991, pp. 67-75.*

13. Petajan, E. D., Brooke, N. M., Bischoff, B. J. and Bodoff, D. A. "An Improved Automatic Lipreading System to Enhance Speech Recognition", *Proc. Human Factors in Computing Systems, pp. 19-25, ACM, 1988.*

14. Robert-Ribes, J., Piquemal, M., Schwartz, J. and Escudier, P. "Exploiting Sensor Fusion Architectures and Stimuli Complementarity in AV Speech Recognition", *Proc. NATO ASI Conference on Speechreading by Man and Machine, 1995, in press.*

15. Silsbee, P. L. and Bovik, A. C. "Medium Vocabulary Audiovisual Speech Recognition", *Proc. NATO ASI New Advances and Trends in Speech Recognition and Coding, 1993, pp. 13-16.*

16. Silsbee, P. L. and Su, Q. "Audiovisual Sensory Integration Using Hidden Markov Models", *Proc. NATO ASI Conference on Speechreading by Man and Machine, 1995, in press.*

17. Tomlinson, M. J., Russell, M. J. and Brooke, N. M. "Integrating Audio and Visual Information to Provide Highly Robust Speech Recognition", *Proc. ICASSP, Vol. 2, 1996, pp. 821-824.*