NONLINEAR SCALE DECOMPOSITION BASED FEATURES FOR VISUAL SPEECH RECOGNITION

Iain Matthews, J. Andrew Bangham, Richard Harvey and *Stephen Cox*
School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK
e-mail: {iam,ab,rwh,sjc}@sys.uea.ac.uk

ABSTRACT

A mathematical morphology based filter structure called a sieve is used to process mouth image sequences of a talker's mouth and form visual speech features. The effects of varying the type of filter, the post-processing and hidden Markov model (HMM) parameters on recognition accuracy are investigated using two audio-visual speech databases.

1 INTRODUCTION

It is well known that visual speech information is used to improve the intelligibility of speech, especially in noisy conditions [10, 20]. Audio-visual blend-illusions [15] where the perceiver 'hears' something other than either of the, deliberately different, audio or visual speech signals demonstrate how fundamentally bimodal our speech perception is. The addition of visual speech features to computer speech recognition has also demonstrated improved accuracy, for example [1, 8, 18]. A recent overview of the field can be found in [13].

In this paper we address the problem of visual speech feature extraction. At the extremes, the methods for extracting visual features can be classified as either model-based or low-level. An example of the model-based approach is to use some form of 'snake', for example [14]. Low-level methods, such as 'eigenlips' [8], do not assume any prior information and operate directly on the pixel values.

Our low-level approach is to use a robust nonlinear multiscale morphological filter structure called a sieve [3–5]. This allows us to decompose an N-D signal by scale. In 1-D scale is length, in 2-D scale is area and so on. We apply a 1-D sieve to a 2-D image by scanning the image vertically and so measure the vertical lengths that form the image. This information is further processed to form scale-histograms that represent a measure of the overall structure of the image independent of position information. Finally, principal component analysis (PCA) is used to form a reduced vector in the directions of most variance. We assess the effects on visual speech recognition accuracy of varying the type of processing used to build these features.

2 DATABASES

In the absence of a standard database each research group has collected their own. Two easily obtained are the *Tulips* database of isolated digits recorded by Javier Movellan at UCSC [17] and our own *AVletters* database of isolated letters [9, 11].

The AVletters database consists of three repetitions by each of ten talkers, five male (two with moustaches) and five female, of the letters A–Z. Each utterance was digitised at quarter frame PAL resolution (376 × 288 at 25fps) using a Macintosh Quadra 600AV in ITU-R BT.601 8-bit headroom greyscale. Audio was simultaneously recorded at 22.05kHz, 16-bit resolution [1]. The mouth images were further cropped to 80×60 pixels after locating the centre of the mouth in the middle frame of each utterance.

The Tulips database contains two repetitions of the digits 1–4 by each of 12 talkers, 9 male and 3 female. The database was digitised at 100 × 75 resolution at 30fps using a Macintosh Quadra 840AV in ITU-R BT.601 8-bit headroom greyscale. Audio was simultaneously recorded at 11kHz, 8-bit resolution. This database is available from `http://cogsci.ucsd.edu/~movellan/`.

Table 1 compares both databases.

	AVletters	Tulips
Task	'A'–'Z'	'1'–'4'
No. talkers	10	12
Repetitions	3	2
Utterances	780	96
Frames	18,562	934
Image size	80 × 60	100 × 75
Lighting	ceiling	ceiling + side

Table 1: Comparison of databases.

3 MULTISCALE SPATIAL ANALYSIS

The method we use has its theoretical roots in mathematical morphology and is similar to granulometry. A

[1] This database is available on CDROM by contacting the authors.

sieve is related to alternating sequential filters (formed from openings and closings) and recursive median filters. Sieves preserve scale-space causality [3–5] and they can transform the signal to another domain, called granularity, and such a transformation is invertible [2].

The sieve may be defined in any number of dimensions by defining the image as a set of connected pixels with their connectivity represented as a graph [12], $G = (V, E)$ where the set of vertices, V, are pixel labels and E, the set of edges, represent the adjacencies. Defining $C_r(G)$ as the set of connected subsets of G with r elements allows the definition of $C_r(G, x)$ as those elements of $C_r(G)$ that contain x.

$$C_r(G, x) = \{\xi \in C_r(G) | x \in \xi\} \tag{1}$$

Morphological openings and closings, over a graph, may be defined as

$$\psi_r f(x) = \max_{\xi \in C_r(G,x)} \min_{u \in \xi} f(u) \tag{2}$$

$$\gamma_r f(x) = \min_{\xi \in C_r(G,x)} \max_{u \in \xi} f(u) \tag{3}$$

The effect of an opening of size one, ψ_2, is to remove all *maxima* of area one when working in 2-D. In 1-D it would remove all maxima of length one. γ_2 would remove *minima* of scale one. Applying ψ_3 to $\psi_2 f(x)$ will now remove all maxima of scale two and so on. Sieves, and filters in their class such as alternating sequential filters with flat structuring elements, depend on repeated application of such operators at increasing scale. Each stage removes maxima and/or minima of a particular scale. The output at scale r is denoted by $f_r(x)$ with

$$f_1 = \mathcal{Q}^1 f = f \text{ and } f_{r+1} = \mathcal{Q}^{r+1} f_r \tag{4}$$

where $\mathcal{Q}$ is one of the γ or ψ operators. Illustrations of sieves and formal proofs of their properties appear elsewhere [3]. The differences between successive stages of a sieve, called *granule functions*, $d_r = f_r - f_{r+1}$, contain non-zero regions, called *granules*, of only that scale.

In one-dimension the graph, (1), becomes an interval

$$C_r(x) = \{[x, x + r - 1] \, | x \in \boldsymbol{Z}\} \tag{5}$$

where $\boldsymbol{Z}$ is the set of integers and C_r is the set of intervals in $\boldsymbol{Z}$ with r elements and the sieves so formed give decompositions by length. For lipreading a 1-D sieve is used to measure the lengths of features seen vertically down the face in the mouth region and these vary as the mouth opens and shuts.

The sieves used in this paper differ in the order in which they process extrema. In 1-D the effect of applying an opening of size one, ψ_2, is to remove all maxima of length one, an *o*-sieve. Likewise a γ_2 would remove minima of length one, a *c*-sieve. These are identical to granulometries.

For this work, we also use a variant in which the maxima and minima are removed in a single pass. This is equivalent to applying a recursive median filter at each scale [5]. The sieve so formed is called an *m*-sieve. It inherits the ability to robustly reject noise in the manner of medians [7] and is much quicker to compute than conventional scale-space preserving schemes.

A granularity is obtained for each mouth image of an utterance, in turn, by applying a one-dimensional sieve along each vertical line. A large number granules are obtained and the problem is how to reduce the number of values to manageable proportions. Here, we take the simple step of creating a histogram of granule scales. This is a rough measure of the overall shape of the mouth. It provides a simple method of substantially reducing the dimensionality from that of the raw image data to the maximum scale used in the sieve. In these examples between 60 and 100 scales are used. The observation vector for the HMM classification is formed by further processing each "scale-histogram".

The simplest form of scale-histogram is obtained by counting the number of granules found at each scale, from 1 to maximum scale and plotting this as a histogram, *sh*. An alternative is to calculate "granule energy" by summing the squared amplitudes, a^2. Other alternatives include summing the raw amplitudes, a and the absolute amplitudes, $|a|$, noting that granules can have negative amplitude. Examples of these are shown in Figure 1.

The changes in a scale-histogram can be followed over time in Figure 2 for a $|a|$ histogram. The scale-histogram is plotted as intensity, white represents a large number of granules. The top row is the smallest scale and the bottom the largest.

The dimensionality of the scale-histograms is further reduced to 5, 10, 15 or 20 features by principal component analysis.

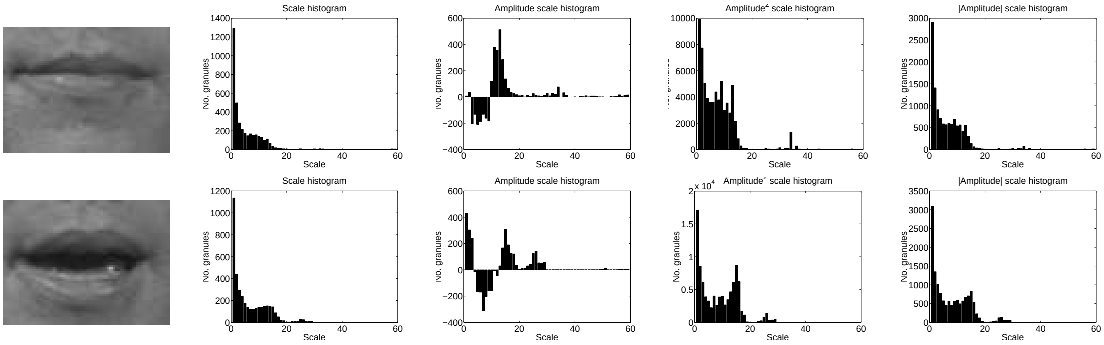
4 RESULTS

For the AVletters database 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). For the Tulips database recognition was performed using the first utterance from each of the twelve talkers as a training set (12 examples per utterance) and the second utterance from each talker as a test set (12 examples per utterance).

Classification was done using left to right HMM's, each state associated with a one or more Gaussian densities with a diagonal covariance matrix. All HMM's were implemented using the HMM Toolkit HTK V2.1 [21].

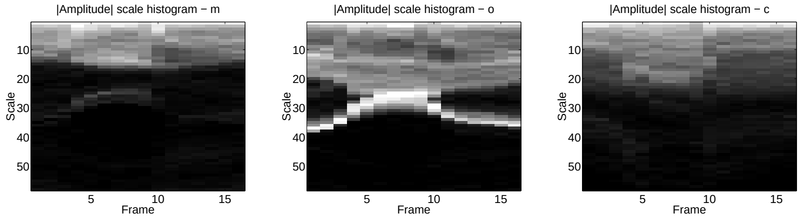
The first step is to find some ground rules on what best characterises the mouth movement. For example Figure 2 shows that the type of analysis could affect the result. We investigated the following options:

i. processing with the *m*-sieve, *o*-sieve or *c*-sieve;



mouth sh a a^2 $|a|$

Figure 1: Comparison of scale-histograms for closed, top panel and open, bottom panel, mouths. Abscissa runs from scale 0 to scale 60 and the ordinate shows the number of granules.



m-sieve o-sieve c-sieve

Figure 2: The changes in three different $|a|$ histograms over time observed for the utterance 'M'. Intensity is a function of absolute amplitude, the abscissa is time and the ordinate scale with small scale granules shown at the top. Left panel, m-sieve, middle panel o- and right panel c-sieve.

ii. preserving DC component of the image or ignore it;

iii. temporally interpolating the visual features or not;

iv. varying the number of principal components;

v. using the covariance or correlation matrix;

vi. varying the number of states in the HMM;

vii. varying the number of Gaussian modes per state.

We form features using PCA so all that needs to be determined are the eigenvectors of the covariance or correlation matrix. Exploring all the above variables was a lengthy computational task, however, the results show several clear trends that allow us to discount some options and present the only the interesting results. For example the experiments show that it is generally better to ignore the DC component and to use the covariance matrix when calculating the PCA.

It would be expected that most of the information would be associated with the boundary of the dark interior of the mouth. This is most effectively distinguished by a closing granulometry, and very badly characterised by an opening granulometry. The recognition results confirm this and we therefore concentrate on results from the c and m-sieve, which is bipolar and more robust [7].

The remaining results are summarised in Table 2. Using nine states and three Gaussian modes per state are preferred. There also seems to be a slight advantage in using interpolated data. The best results are obtained using the $|a|$ histograms from a c-sieve, followed closely by the m-sieve. The best results, 44.6% and 40.8%, are obtained with interpolated $|a|$ histograms for c and m-sieves respectively.

The trends in the results shown for the AVletters database are reflected in the results obtained with the Tulips database. Table 3 shows a direct comparison of results obtained using the best analysis options ($|a|$ histogram using a c-sieve, ignoring DC, PCA with covariance matrix).

States	5		7		9	
Modes	1	3	1	3	1	3
AVletters 10	16.5	30.8	25.4	37.7	30.0	37.3
AVletters 20	24.6	36.1	27.3	36.5	32.7	**44.6**
Tulips 10	66.7	54.2	**77.1**	58.3	75.0	72.9
Tulips 20	62.5	52.1	66.7	58.3	64.6	68.7

Table 3: Recognition accuracies, %, for Tulips and AVletters with variations in the HMM parameters: no. states and no. Gaussian modes per state. For 10 and 20 PCA coefficients.

T	S	c-sieve				m-sieve			
		I		NI		I		NI	
		1	3	1	3	1	3	1	3
sh	5	26.2	36.2	24.6	34.6	21.9	38.1	18.5	36.2
	7	24.2	36.6	28.5	34.6	27.3	40.8	25.8	41.2
	9	30.8	37.7	30.8	39.2	27.3	38.9	28.5	40.8
a	5	18.9	35.8	20.0	33.5	21.2	31.5	20.4	32.3
	7	23.9	37.3	23.5	34.2	20.8	33.1	21.9	30.4
	9	27.3	38.5	28.5	36.2	27.8	34.6	22.7	33.9
$\|a\|$	5	24.6	36.2	22.3	40.0	19.6	36.2	20.8	35.8
	7	27.3	36.5	26.9	36.2	28.1	36.9	25.8	38.9
	9	32.7	**44.6**	30.0	41.5	30.0	40.8	28.1	39.6
a^2	5	17.7	34.2	13.1	31.9	18.1	31.9	19.6	29.6
	7	23.1	34.6	21.2	34.6	20.0	32.6	21.5	30.4
	9	21.5	37.3	27.7	28.4	23.4	31.5	21.2	30.8

Table 2: Shows how varying the HMM parameters: number of states, S, and Gaussian mixtures (1 or 3) affect recognition accuracy, %, for interpolated, I, and non-interpolated, NI, AVletters data for both c-sieve and m-sieve.

5 DISCUSSION

A major problem in this field, unlike acoustic speech recognition, is the lack of a standard task. Here we have attempted to overcome this by performing tests on two databases. However, there is an urgent need to compare this method with alternative visual feature extraction methods, particularly those using a model-based approach. We are currently addressing this.

A notable shortcoming of this system is that it is sensitive to scale variation. For example motion of the talker towards the camera introduces unwanted variation. This might be solved by using an automatic head tracker, such an approach has been implemented elsewhere [8, 16].

References

[1] A. Adjoudani and C. Benoît. *On the Integration of Auditory and Visual Parameters in an HMM-based ASR*, pages 461–471. In Stork and Hennecke [19], 1996.

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

[6] C. Benoît and R. Campbell, editors. *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, Rhodes, Sept. 1997.

[7] A. Bosson, R. Harvey, and J. A. Bangham. Robustness of scale space filters. In *BMVC*, volume 1, pages 11–21, 1997.

[8] C. Bregler and S. M. Omohundro. Learning visual models for lipreading. In M. Shah and R. Jain, editors, *Motion-Based Recognition*, volume 9 of *Computational Imaging and Vision*, chapter 13, pages 301–320. Kluwer Academic, 1997.

[9] S. Cox, I. Matthews, and A. Bangham. Combining noise compensation with visual information in speech recognition. In Benoît and Campbell [6], pages 53–56.

[10] N. P. Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, 12:423–425, 1969.

[11] R. Harvey, I. Matthews, J. A. Bangham, and S. Cox. Lip reading from scale-space measurements. In *Proc. Computer Vision and Pattern Recognition*, pages 582–587, Puerto Rico, June 1997. IEEE.

[12] H. J. A. M. Heijmans, P. Nacken, A. Toet, and L.Vincent. Graph morphology. *Journal of Visual Computing and Image Representation*, 3(1):24–38, March 1992.

[13] M. E. Hennecke, D. G. Stork, and K. V. Prasad. *Visionary Speech: Looking Ahead to Practical Speechreading Systems*, pages 331–349. In Stork and Hennecke [19], 1996.

[14] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In *Proc. European Conference on Computer Vision*, volume II, pages 376–387, 1996.

[15] H. McGurk and J. McDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, Dec. 1976.

[16] U. Meier, R. Stiefelhagen, and J. Yang. Preprocessing of visual speech under real world conditions. In Benoît and Campbell [6], pages 113–116.

[17] J. R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, 1995.

[18] E. D. Petajan. *Automatic Lipreading to Enhance Speech Recognition.* PhD thesis, University of Illinois, Urbana-Champaign, 1984.

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

[20] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2):212–215, Mar. 1954.

[21] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book.* Cambridge University, 1996.