Proceedings of the Institute of Acoustics

AUTOMATIC SPEECH RECOGNITION THAT INCLUDES VISUAL SPEECH CUES

N.M. Brooke (1), M.J. Tomlinson (2), and R.K. Moore (2)

(1) School of Mathematical Sciences, University of Bath, Bath
(2) Speech Research Unit, DRA, Malvern

1. INTRODUCTION

The movements of speakers' faces can convey visual cues to speech which tend to complement the acoustic cues; for example, acoustic cues to the place of articulation which are easily destroyed by noise are often visually robust [1]. The hearing-impaired have long been aware of the potential benefits of augmenting their restricted acoustic inputs with visual speech signals in order to lip-read. However, automatic speech recognizers should also be capable of enhanced performance, especially in noisy conditions, if their conventional acoustic inputs can be augmented with visual data. This approach is attractive because it promises a reasonable level of noise immunity at a potentially low computational cost. Whilst there are techniques for processing acoustic signals to provide compensation for a noisy environment, they are commonly computationally expensive or only partially effective [2].

The principal problems in implementing audio-visual speech recognizers are a) managing the large volumes of data that sequences of images generate and b) integrating visual and acoustic data so that the best use can be made of both together. One early visual recognition system [3] computed the differences in pixel intensity between successive monochrome oral images in order to build a time-varying template that could be used for pattern matching with templates for a set of reference words. Another prototypical, isolated word, speaker-dependent visual recognition system [4] achieved data compression by windowing video images so as to retain only the mouth region of speakers, reducing these oral images to black and white and then contour-coding to capture the boundaries of the dark areas which represented the oral cavities. These codes were used to construct time-varying templates of parameters such as the width, height, area and perimeter of the oral cavity areas, which were used for comparison with corresponding templates from a reference vocabulary of spoken digit words. In this system, acoustic word recognition was performed independently and in parallel. Audio-visual recognition was achieved by combining heuristically the word recognition results from the two modalities and it was therefore difficult to assess accurately the gains arising from the addition of the visual speech component, which is an early objective of audio-visual speech recognition experiments.

The present paper describes the development of a speech recogniser based upon hidden Markov models, or HMMs [5], that use *composite*, audio-visual feature vectors. The visual component of the HMMs is monochrome oral image data, compressed by means of the method of principal component analysis. By bringing together these two techniques, this approach integrates both visual and acoustic data and processes them together throughout. It contrasts with hierarchical approaches in which categorical decisions may be taken at a number of different points within independent processes, after which possibly relevant data are lost. In order to assess the performance of the system, a speaker-dependent, small-vocabulary task was chosen, namely, the recognition of connected digit triples (for example, 'six one seven'), using a continuous recognition system with the ten digit words as its effective vocabulary. Even though this is a linguistically trivial recognition task, it is nonetheless one with a range of potentially useful applications. A range of levels of simulated noise was superimposed on the acoustic input to the system in order to investigate the benefits gained by using the visual component of the speech signal.

2. AUDIO AND VIDEO DATA COLLECTION AND PROCESSING

2.1. The database

A database was constructed from recordings of a single, native English speaker enunciating utterances from the NATO RSG-10 standard digit triple lists 3A, 3B and 3C. Each list consisted of 50 triples and was recorded twice. Simultaneous, synchronised video and audio recordings were made.

2.2. Recording arrangements

The speaker was seated in a sound-proof booth in a chair with a head-rest to which were attached guide posts. With this arrangement a) the head movements could be largely eliminated and b) the head could be brought to a fixed orientation against the guide posts. A videocamera was located approximately one metre away from the speaker's face at the same height as the speaker's mouth and a Shure SM48 microphone was mounted on a boom at about 0.2 metre from the speaker's mouth and somewhat below it, out of line of sight of the camera. Frontal lighting was provided by two festoon lamps arranged on either side of the camera and as close to it as possible. A TV monitor was placed on a table in front of the speaker and just out of line of sight of the camera so that the speaker could see the facial images as they were being recorded. The horizontal limits of the lips in their rest position was marked on the screen. The speaker was instructed to maintain this head alignment during 'takes'. Prompting sheets were also placed on the table where they could be read without requiring head movement. The recordings were monitored by an operator outside the booth and who could talk to the subject over an intercom if errors were noted which necessitated re-recording.

2.2.1. Video recordings. The monochrome camera was fitted with a 8mm focal length, f1.8 aperture lens incorporating an automatic iris system. The camera was adjusted so that the speaker's face from just above the nostrils to just below the chin occupied the whole frame. The images of the oral area were digitally captured in 64 x 64 pixels (1 byte of intensity per pixel) at 25 frames per second, using a specially developed audio-visual collection unit based on Transputer technology. At this data rate, images could be captured and stored on a PC disc in real-time.

2.2.2. Audio recordings. The audio-visual collection unit (see Section 2.2.1 above) was designed and programmed to perform a 26-channel filterbank analysis of the signals from the microphone, which were sampled at 20 kHz. No pre-emphasis was used in these experiments. The centre frequencies and bandwidths of the channels were set at values used in the JSRU channel vocoder [6], with extensions to cover the range 60 Hz to 10kHz. The filterbank outputs, in logarithmic form, were sampled 100 times per second. These samples represented the acoustic speech data. The filterbank representation is a convenient and economical one for prototypical experiments and is unlikely to be significantly worse than more sophisticated representations for small-vocabulary, speaker-dependent experiments involving noise contamination [7]. A simultaneous DAT recording was also made for archiving and reference purposes; it was not synchronised with the video recordings and was not therefore used for the recognition experiments.

2.2.3. Audio and video signal synchronisation. A single crystal clock was used to drive both a video sync. generator and a sampling clock generator. The videocamera was locked to the video sync. and the audio-visual collection unit was controlled by the sampling clock generator. In this way, the 25 video frames per second were synchronised with the 100 filterbank outputs per second in a strict 1:4 ratio. The video and audio data were multiplexed for transmission from the audio-visual collection unit to the PC and de-multiplexed for separate storage of the audio and video on the PC's disc.

2.3. Video data processing and PCA encoding

To reduce the video data to manageable proportions, the 64 x 64 pixel facial images were first reduced to 16 x 16 pixel images by simple pixel averaging. A fixed window of 6 x 10 pixels was then selected, completely covering the oral region. The window's position was fixed relative to the full image and the spatial resolution was chosen to lie near the limit below which experiments [8,9] have suggested that visual speech cues are lost. Further video compression was achieved by using principal component analysis, or PCA [10]. PCA is a data-driven method that requires no *a priori* information about the structure of the data. Although it is only one of a number of possible methods, it has a number of attractions [11]. Image sequences from 200 of the recorded digit triples were used were used to build a PCA encoder. The 'mean image' of the set was computed and subtracted from each of the images prior to PCA. Each resultant image represented a point in 60-dimensional 'pixel-intensity' space which PCA transformed into a set of orthogonal axes such that each new axis accounted for as much as possible of the remaining variance. The training showed that only a small number of axes was needed to account for the greater part of the variance; the first axis accounted for 13.0%; the first three for 32.4% and the first ten for 62.1% of the variance. An *n*-channel PCA encoder used the co-ordinates in the first *n* axes derived from the training process to code an image. Since the oral images are both highly structured and dynamically constrained by the underlying anatomy, images outside the training set can be coded accurately in this way, as illustrated in Figure 1. The PCA encodings have a generally consistent relationship with specific articulatory gestures. The values of the individual PCA coefficients that represent the encodings of each image show variations with time over a sequence of images that are broadly smooth and compatible with articulatory rates of change.

2.4. Addition of simulated noise to acoustic data

Simulated wide-band, stationary noise was added to the acoustic channel in order to investigate the benefits of adding a visual channel to the recognition system. The use of a filterbank acoustic representation (see Section 2.2.2) precluded the direct superimposition of digital speech and noise waveforms. Instead, a filterbank representation of simulated noise was added to the filterbank representation of the speech signal using a 'max' operation, which, in the log space of the filterbank representation, is an approximation to addition in linear space. Filterbank representations of the speech data from the training set and noise signals at different levels were used to estimate signal power and hence to calibrate the noise-contaminated samples in terms of their signal to noise ratio in dB. The simulated noise was effectively spectrally flat to within 1 dB across the full frequency range and could be added 'on-line' as an optional stage in the recognition process (see Section 3).

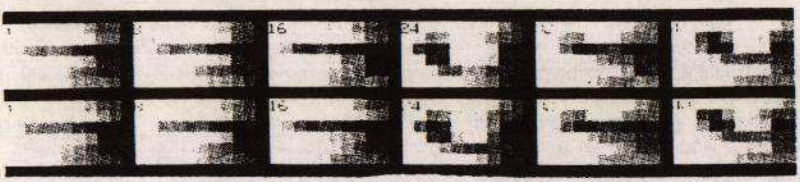


Figure 1: Oral images (6 x 10 pixels) recorded during speech production. The top row shows the original images and the bottom row, the reconstructions from a 6-channel PCA image encoding.

3. THE RECOGNITION SYSTEM AND THE TRAINING PROCEDURE

A continuous speech recognition system was used to identify the connected digits within the triples, using sub-word HMMs. These consisted of thirty 3-state HMMs to model the set of *triphones*, that is, the single phonemes in the context of each of their preceding and following phonemes, that can occur in the test vocabulary. The HMMs allowed transitions only to successor states or back to the same state. Four HMMs were also used to model various non-speech events such as silences, lip smacks and breath noise. Composite, audio-visual feature vectors were used in the models and were constructed by concatenating the *m* filterbank (audio) inputs with *n* PCA coder (visual) outputs. Four replications of the PCA codes for each recorded video frame were used to create composite, audio-visual feature vectors at 100 Hz (see Section 2.2.3). Each state of each HMM was identified with a single multivariate Gaussian density function of dimension (*m+n*) and a diagonal covariance matrix. The mean and standard deviation for each dimension of each state, as well as the state transition probability matrices, were re-estimated from initial values for each model during training, using the Baum-Welch algorithm [5,12], in the following way. The set of 200 digit triples that was used to build the PCA image encoder (see Section 2.3) was orthographically labelled and the triples were marked with their start and finish times, using a forced recognition alignment on the audio data and hand correcting where necessary. The various types of non-speech items were also marked. A pronunciation dictionary was used to define the phone sequences in the training data and allocate these in equal time slices among the states of a sequence of phonetically context-insensitive, or *monophone*, HMMs. The monophone HMMs were used to compute initial estimates of the means and standard deviations of the parameters of the feature vectors. These were then used to start the re-estimation procedure for the appropriate triphone HMMs.

4. RECOGNITION EXPERIMENTS AND RESULTS

Recognition experiments were carried out using 100 of the recorded digit triples that were not used either to build the PCA image encoder or for HMM training, with varying levels of added noise. Five conditions were examined in each experiment, according to the feature vector composition, as follows: a) the 26 filterbank parameters, or audio signal, only ('a26v0'); b) the same plus 1 channel of PCA image encoding ('a26v1'); c) the same plus 3 channels of PCA image encoding ('a26v3'); d) the same plus 10 channels of PCA image encoding ('a26v10'); and e) 10 channels of PCA image encoding, or visual signal, only ('a0v10'). The recognition results were scored as a percentage word accuracy, which takes insertion errors into account as well as substitution and deletion errors. A phone-mediated, dynamic programming method was used to find the optimal alignment between the recognised and test digits.

Two types of HMM were examined. The first type was computed with state-specific variances, that is, the means and variances of every parameter in the feature vectors were computed independently for each state of each HMM. The second type was computed with so-called grand variances, or, specifically, feature-specific variances. Each type of HMM was tested with and without the incorporation of a standard noise-tracking and masking algorithm [13] that compensated for some of the effects of noise in the acoustic signal and hence set a more realistic baseline for assessing the contribution of the visual component to speech recognition in noise. Separate model re-estimation was carried out for the SSV and the GV models in each of the experimental conditions, using acoustic speech signals without noise contamination. For practical reasons, re-estimation was not carried out separately for the noise-tracking and masking experiments. The visual-only condition was of course unaffected by the the addition of noise-tracking and masking to the recognizers.

4.1. State-specific variance (SSV) models
Experiments A and B used SSV models, respectively without and with noise-tracking and masking.

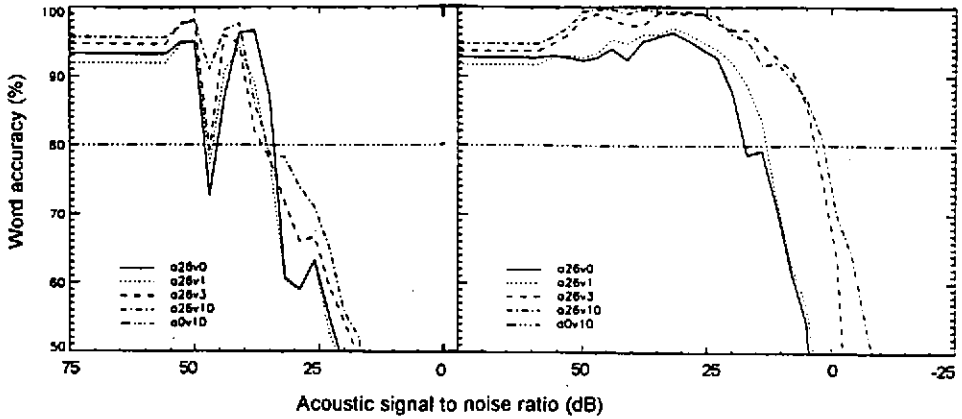
The results of experiment A are shown in Figure 2a. They showed that for an audio-only signal, low levels of noise down to a signal-to-noise ratio (SNR) of 60 dB did not significantly degrade the recognition performance. There was a notch at a SNR of 47 dB. This reflected the frequency with which the recognizer accounted for precisely this level and type of noise and the trailing end of a digit triple, as the digit 'two'. At SNRs below 37 dB, recognition performance fell off rapidly. Audio-visual recognition with only one channel of visual data produced no improvement outside the region of the notch. Three or ten channels of visual data in the composite speech signal both produced recognition gains in the low noise region down to a SNR of about 50 dB and again in the higher noise region at SNRs below about 33 dB. The gains were not maintained in the intervening region between 50 and 33 dB. Ten channels of visual data gave better performance than three at all noise levels. However, audio-visual performance at SNRs below 35 dB was consistently poorer than the 80% word accuracy rate of visual-only recognition. The visual data was pulled down by the corrupted audio data.

Use of noise-tracking and masking (Experiment B, shown in Figure 2b) produced the anticipated improvement in the audio-only recognition performance as well as consistent improvements in audio-visual recognition throughout virtually the whole range of SNRs. The notch of experiment A was also eliminated. One channel of visual data produced little improvement in performance, while ten channels produced the best improvement. Three channels of visual data were only slightly inferior to ten. The addition of noise-tracking and masking to the audio-visual recognition process produced useful gains at low and medium noise levels and substantial gains at high noise levels. It had therefore done its job of allowing the visual data to influence the recognition process when the acoustic signal was corrupted by noise. However, the combined mode recognition rates still fell below the visual-only recognition performance. For both the re-estimation and recognition processes, the probability of obtaining a particular observation vector in a given state of an HMM must be computed. For the uncorrelated Gaussian distributions of feature values assumed in the current HMMs, this calculation involves both a variance term and a term for the deviation from the mean value of each feature. When masking is invoked at high noise levels, the deviation term tends towards zero for the acoustic features. Consequently, when there is noise, the recognition process tends to be biassed towards the states of HMMs with small variances, hence pulling down the performance of the audio-visual recognition process despite the presence of an uncorrupted visual signal.

4.2. Grand variance (GV) models
Experiments C and D used GV models, respectively without and with noise-tracking and masking. GV models reduce the problems of bias towards states with small variances described above, by computing variances which are common across all states of all the models, but are still specific to each feature in the vector. The result would be that variance would play its normal part in the computations unless the means were masked. In this case, there would be no bias to particular states because the relevant variances would be equal across all states. Since GV models use fewer parameters than SSV models, recognition performance for GV models tends to be lower than for the corresponding SSV models.

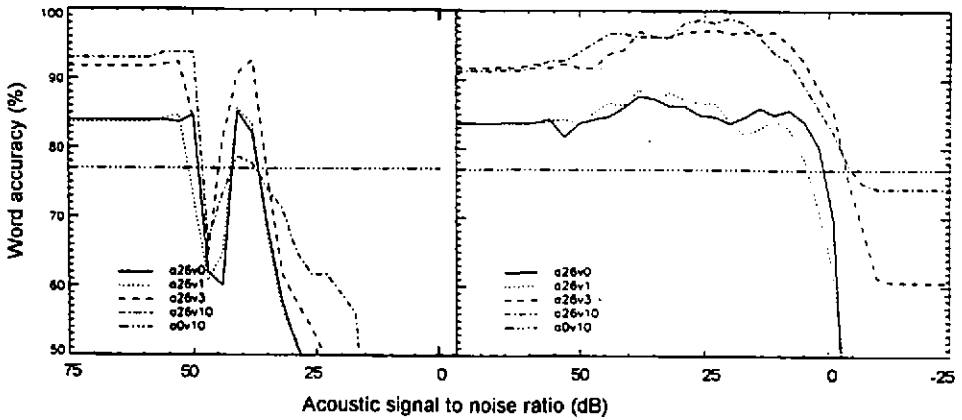
The results for experiment C, shown in Figure 2c, were broadly similar to those for experiment A, except for the expected overall reduction in word accuracy rates for the reasons described above.The addition of acoustic noise-compensation in experiment D (Figure 2d) produced the same anticipated gains as experiment B showed in comparison to experiment A for the SSV models. That is, whilst the recognition rates at low noise levels were slightly reduced, the noise tolerance at high noise levels was significantly increased. A single visual data channel shows no improvement over audio alone; three visual channels

are almost as good as ten and markedly improve performance. In particular, these curves flatten out at SNRs worse than -10 dB, indicating that the recognition system is relying almost entirely on the visual



a) Experiment A: SSV models, no acoustic noise compensation

b) Experiment B: SSV models with acoustic noise-tracking and masking



c) Experiment C: GV models, no acoustic noise compensation

d) Experiment D: GV models with acoustic noise-tracking and masking

Figure 2: Results of the audio-visual speech recognition experiments as acoustic noise is added to the speech signal. The experiments are described in Section 4 of the text.

data when the audio data is destroyed by noise. The 'a26v10' curve however flattens out below the visual-only performance level rather than trending towards this value. This may be due to incorrect synchronisation of the audio and visual data during the training phase, including the use of replicated video frames (see Section 3), so that the visual component of the optimised audio-visual models was not as accurate as the visual models optimised on visual data only.

5. DISCUSSION

The experimental results demonstrate that the use of non-acoustic data, in the form of visible lip information, can be incorporated into composite HMMs and can produce gains in performance for a speech recognition system applied to ten word digit vocabulary. Since the audio and visual data is integrated within the HMMs, there is no need to switch between different forms of recognition at different acoustic noise levels. The performance gains can be achieved over a large range of acoustic noise levels and are substantial. They can be expressed in a number of ways. For example, at the 90% word accuracy level, a recognition system based on SSV models with noise compensation and employing 3 channels of visual and 26 channels of audio information can tolerate 14 dB more audio noise than an audio-only system. If GV models are substituted for the SSV models, a gain of 13% in accuracy can be obtained at a SNR of 0dB when the 26 channels of acoustic data are augmented by 3 channels of visual data. The accuracy of the 26 audio plus 3 visual composite GV models using noise-tracking and masking, at a SNR ratio of 10 dB, is 97%. The highest accuracy achieved for the audio-only system at this noise level is 85%. Useful applications can be envisaged for audio-visual recognizers with this kind of vocabulary and performance. The results also indicate that the ideal audio-visual recognizer would use SSV models with appropriately large numbers of parameters, but in which the effect of the variances in noise-masked features can be removed from the calculations (see Section 4.2). It is possible that this might in future be achieved by using masking techniques such as that of Gales and Young [14] in order to achieve high accuracy at both low and high levels of noise.

The use of PCA coding may not provide the best possible visual representation. The lack of performance improvement when a single channel of visual data is added to form a composite audio-visual feature vector suggests that the first PCA coefficient contributes little in terms of discrimination. This may not be unreasonable, as it represents only a location along the axis in which the visual data showed the most variation. PCA has nothing to reveal about the way that classes are disposed along the dimension. Linear discriminant analysis [15] may, in the longer term, offer more useful insights, but has not yet been investigated. Alternatively, the contribution of the various PCA coefficients to discrimination might be subjectively explored by generating pair-wise scatter plots from these data. A further issue to be explored is the effect of using images of differing spatial resolution in the recognition process.

The gains of using visual data in automatic audio-visual speech recognition may also be assessed by comparison with the results of perceptual experiments equivalent to those described in this paper on human listeners, using the same test data. This general approach is already being used to assess the performance of computer graphics syntheses of visible facial speech articulations [16].

6. REFERENCES

[1] N M BROOKE, 'Visible speech signals: investigating their analysis, synthesis and perception', in 'The Structure of Multimodal Dialogue', edited by M.M. Taylor, F. Neel and D.G. Bouwhuis (North-Holland,Amsterdam) 249-258 (1989)

[2] B A MELLOR & A P VARGA, 'Noise masking in a transform domain', Proceedings of IEEE ICASSP, 2, 87-90 (1993)

[3] S NISHIDA, 'Speech recognition enhancement by lip information', Proceedings of CHI 86 (Association for Computing Machinery, New York), 198-204 (1986)

[4] E D PETAJAN, 'Automatic lipreading to enhance speech recognition', Proceedings of the Global Telecommunications Conference, Atlanta, Georgia (IEEE Communication Society), 265-272 (1984)

[5] L R RABINER, 'A tutorial on hidden Markov models and selected applications in speech recognition', Proceedings of the IEEE, 77, 257-286 (1989)

[6] J N HOLMES, 'The JSRU channel vocoder', Proceedings of the IEE, F127, 53-60 (1980)

[7] M KADIRKAMANATHAN, 'Hidden Markov model decomposition recognition of speech in noise: a comprehensive experimental study', Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, Nice, 187-190 (1992)

[8] N M BROOKE & P D TEMPLETON, 'Visual speech intelligibility of digitally processed facial images', Proceedings of the Institute of Acoustics Autumn Meeting, Windermere, 12(10), 483-490 (1990)

[9] N M BROOKE, 'Computer graphics synthesis of talking faces', in 'Talking Machines: Theories, Models and Designs', edited by G. Bailly, C. Benoit & T.R. Sawallis (Elsevier, Amsterdam), 505-522 (1992)

[10] B FLURRY, 'Common Principal Components and Related Multivariate Models' (Wiley, New York) (1988)

[11] N M BROOKE & M J TOMLINSON, 'Processing facial images to enhance speech communications', to appear in 'Proceedings of the Second Venaco Workshop on the Structure of Multi-modal Dialogue, Maratea, Italy, 1991', edited by M.M. Taylor, F. Neel and D.G. Bouwhuis

[12] K-F LEE, 'Large vocabulary speaker independent continuous speech recognition: the SPHINX system', Ph.D Thesis, Carnegie-Mellon University (1988)

[13] D H KLATT, 'A digital filterbank for spectral masking', Proceedings of IEEE ICASSP, 573-576 (1976)

[14] M J F GALES & S J YOUNG, 'An improved approach to hidden Markov model decomposition of speech and noise', Proceedings of IEEE ICASSP, 1, 233-236 (1992)

[15] C CHATFIELD & A J COLLINS, 'An Introduction to Multivariate Analysis' (Chapman & Hall, London), Chapter 7, 114-139 (1980)

[16]N M BROOKE & S D SCOTT, 'Computer graphics animations of talking faces based on stochastic models', Proceedings of ISIPNN '94, Hong Kong (IEEE), 73-76 (1994)