

USING THE VISUAL COMPONENT IN AUTOMATIC SPEECH RECOGNITION

N Michael Brooke

Media Technology Research Centre,
School of Mathematical Sciences,
University of Bath,
BATH, BA2 7AY
United Kingdom

ABSTRACT

The movements of talkers' faces are known to convey visual cues that can improve speech intelligibility, especially where there is noise or hearing-impairment. This suggests that visible facial gestures could be exploited to enhance speech intelligibility in automatic systems. Handling the volume of data represented by images of talkers' faces implies some form of data compression. Rather than using conventional feature extraction approaches, image coding and compression can be achieved using data-driven, statistically-oriented techniques such as artificial neural-networks (ANNs) or principal component analysis (PCA). A major issue is the combination of the audio and visual data so that the best use can be made of the two modalities together. Perceptual experiments may offer guidance on suitable machine architectures, many of which currently use hidden Markov models (HMMs).

1. INTRODUCTION

The visibility of a talkers's face has long been known to improve speech intelligibility, especially where the acoustic speech signal is degraded by noise, or where there is hearing-impairment. The benefit gained from the visual, facial cues has been quantitatively estimated to be equivalent to an increase of 8-10 dB in the signal-to-noise ratio when speech sentences are presented in a noise background [9]. This observation suggests that, if the acoustic inputs to conventional speech recognition systems could be augmented by data about the visible speech gestures, an enhanced-performance, audio-visual recognition system should be possible. Indeed, one of the challenges of speech technology is to be able to provide robust and accurate automatic systems capable of operating successfully in a wide range of environments, including those where high levels of noise and vibration may be encountered. Aircraft cockpits are one example of a demanding environment in which reliable automatic speech recognition is becoming an important requirement.

Techniques are available to provide some compensation for adverse acoustic environments [12], but in general they provide either a limited improvement only, or require excessive computational power. The incorporation of visual data into speech recognition potentially offers a significant degree of acoustic noise immunity at a reasonable computational cost. The

two major problems are (a) how to deal with the volume of data that moving images represent, and (b) how to combine the acoustic and the visual data in order to make the best use of both, taken together. Many current implementations of audio-visual speech recognizers use hidden Markov models (HMMs) and different architectures for combining the modalities are possible. Cognitive studies of speech perception may offer guidance to direct the choice of a suitable architecture.

2. VISIBLE FEATURES VERSUS IMAGES

Management of image data can most simply be achieved by identifying and extracting features from the images and using these instead. The problem then becomes the identification of the features that are relevant to recognition. To avoid this problem, an alternative approach is to use the images themselves as the source of data and to adopt a statistically-oriented, data-driven approach essentially to reduce the redundancy in the information contained in the pictures and hence to compress the image data. The advantage of using the images is that they encapsulate not only the lips, but other perceptually significant features such as the teeth, tongue and skin texture.

2.1. Feature Extraction

In one of the earliest, visual speech recognition systems [13], binary, black and white images of a talker's lower face were captured in real-time using special-purpose hardware. Measurements of specific facial features such as the oral cavity area, oral cavity perimeter, mouth width and mouth opening were extracted from the black areas of the image sequences and used to form time-varying templates that could be time-warped and matched against stored reference templates from the recognition vocabulary. In speaker-dependent, isolated word recognition experiments on a digit vocabulary, this relatively simple system achieved visual digit recognition rates that were well in excess of 90% correct. A later version used a simple metric to compare the complete outline shapes of the oral cavity [14]. Bimodal, audio-visual recognition was implemented in both systems by categorising words independently in each of the visual and auditory domains and then making a final classification using a heuristic which examined the two outputs for compatibility. Whilst mouth width and opening are known to be perceptually significant, the catalogue of relevant features is in general

difficult to define completely [6] and some of the relevant cues such as the tongue, which may appear only as an indistinct and ill-defined object, cannot easily be represented in simple parametric terms [11]. Nonetheless, a number of visual speech recognition systems continue to employ a range of features of this type with some success [e.g. 1, 18].

2.2. Direct Use of Images

Early, rather limited perceptual experiments on vowel identification of the five long vowels of British-English in an /hVd/ context [3] suggested that the important visual cues could essentially be retained in monochrome, dynamic recordings of the oral region of a single speaker's face in which the spatial resolution was as low as 16 x 16 pixels. An objective identification experiment was also carried out using a multi-layer perceptron (MLP) to classify single frames from the vowel nuclei, of oral images of one speaker uttering the eleven British-English vowels in a /bVb/ context, recorded at 16 x 12 pixels resolution. A three-layer artificial neural network (ANN) was used in which the intermediate, 'hidden' layer contained just 6 neural elements. Supervised learning was employed to train the network on labelled vowel tokens. When presented with unlabelled vowel tokens, the trained network had an average correct identification rate of 91% (84% correct for the worst case vowel). This suggested that (a) visual cues to vowel identity could be retained in images of relatively low spatial resolution and (b) the visual cues to vowel identity could be captured in a low-dimensionality intermediate representation of the images. The second observation has potential implications for data compression of oral images because it implies the possibility of an efficient image encoding scheme.

2.3. Data-driven Image Compression

The early experiments with ANNs assumed no *a priori* knowledge about the content, structure or significance of the data contained in an image and although ANNs proved to have potential drawbacks as data compression and coding devices [e.g. 4], they indicated that a data-driven, statistical view of images might result in a practicable image compression system. The method of principal component analysis (PCA) has been shown to be well-suited to the efficient encoding of oral images. Each of a 'training' set of image frames is represented as one point in an image space which is then transformed so that as much of the data variance of the set as possible is accounted for within as few of the axes of the transformed space, or principal components, as possible. In a recent study [2], approximately 15000 frames of a single speaker uttering digit triple words (e.g. 'six seven two') from the NATO RSG-10 database were videorecorded in monochrome. Following digitisation and reduction to 32 x 24 pixels resolution (i.e. well above the resolution at which visual cues may be lost), they were subjected to PCA and 82% of the variance was found to be accounted for by 15 principal components. This result is not wholly unexpected; the facial anatomy places constraints upon the range of attainable gestures, so that a set of facial images shows considerable internal patterning and structure. The principal components derived from

the training process were then used to define a PCA encoder that could be applied to the coding of unseen, 'test' images with sufficient accuracy to allow image reconstructions from the codes that embodied the essential visual speech cues, i.e. were essentially speech-readable. PCA encodings have been used as the visual component of speech pattern vectors [5].

3. AUDIO-VISUAL INTEGRATION

Whatever the final choice of representation of the visible speech gestures, the other major issue is how to integrate this with the information about the acoustic aspects of the speech signal so that the best use can be made of the two modalities together. The previous section indicated how perceptual experiments with human observers can guide the development of computer architectures and systems design, for example, by suggesting how well visual cues are captured at different image resolutions. Cognitive studies have also suggested different architectures for the combination of the auditory and visual modalities [15].

3.1. Models from cognitive psychology

Four models of audio-visual speech perception were described by Robert-Ribes *et al.* [15], as follows:

1. Direct identification (DI), in which acoustic and visual data are combined and transmitted directly to a single, bimodal classifier.
2. Separate identification (SI), in which two parallel, unimodal classification systems are employed and the results from each are fed forward for fusion and final decision making, for example, on a probabilistic basis.
3. Dominant recoding (DR), in which auditory processing is supposed to be dominant. Visual data is recoded into the dominant modality. Each modality thus generates a representation appropriate to the dominant modality, such as a tract transfer function. The two estimates are then fused and fed forward to a classifier.
4. Motor-space recoding (MR), in which both inputs are projected and recoded into an amodal common space, such as that of articulatory configurations, and the two representations are fused and passed to the classifier.

There is no general agreement about the model that most closely matches human perception, though Robert-Ribes *et al.* believe the evidence may favour the MR model. Evidence militates against the SI model and the general view is that fusion takes place somewhere above the peripheral input system, but at a pre-categorical stage [e.g. 10, 19].

3.2. Automatic recognition architectures

A wide range of methods using both neural networks and HMMs has now been deployed and these have been fully reviewed elsewhere [7]. In terms of the cognitive models of Section 3.1, no automatic audio-visual speech recognition system based on the MR model has yet been reported, though an early example

corresponding to a DR model was a system in which a multi-layer perceptron was used to code image data into a vocal tract transfer function [16]. Most current systems tend to lie in a continuum between the extreme integration strategies, which are either (a) to process the two modalities separately, using independent audio and visual recognisers whose outputs are subsequently combined, or (b) to combine the audio and visual data at the outset and process them throughout as a single, composite feature vector. The extremes essentially correspond to the SI and DI models of Section 3.1. The SI approach was typified by Petajan's prototypical recognizers [13, 14].

3.3. Architectures using HMMs

Following the success of their application to conventional, acoustic speech recognition, hidden Markov models are now being applied widely to bimodal, audio-visual speech recognition. Many of the contemporary studies have been concerned with exploring the benefits that may be gained by incorporating visible signals into the recognition process.

The SI architecture can be investigated by building and training and applying separate acoustic and visual HMMs, then combining the outputs. More commonly, HMMs have been built that represent a DI architecture, using a composite, acoustic and visual feature vector. One typical approach of this kind studied the speaker-dependent recognition of digit utterances from the NATO RSG-10 digit-triple lists [5]. The acoustic data vector consisted of the outputs of a 26-channel filter bank covering 60 Hz to 10 kHz at 100 frames per second. The visual data vector consisted of a 10-channel PCA encoding of 10 x 6 pixel monochrome images of the talker's oral area, recorded at 25 frames per second (10 principal components in this case accounted for 62.1% of the variance of the training images). Each image was replicated four times to match the acoustic frame rate. The audio-visual data consisted of a 36-element composite vector formed by concatenating the acoustic and visual vectors. Concatenated, 3-state triphone HMMs of the kind shown in Figure 1 were used. Each state of these models was associated with a single, multivariate continuous Gaussian distribution and a diagonal covariance matrix. The models were trained on 200 digit triples and tested on 100 digit triples. Simulated, spectrally-flat noise at different power levels was added to the acoustic channels of test tokens so that the benefits of adding a visual component could be studied.

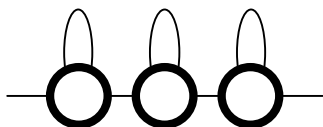


Figure 1: Conventional, composite-vector HMM with states traversed in a single left-to-right sequence.

The second and third columns of the Table below show the word error results at varying acoustic noise levels for conventional, acoustic and bimodal, audio-visual recognizers, respectively. To

simplify the table they are shown only for grand variance HMMs with a silence-tracking and noise-masking technique [8]. This represents an appropriate baseline for comparisons with the 'best' acoustic-only recognizers and was shown to be necessary to prevent the visual components being swamped by the errors caused by the noise in the acoustic channels. The models in this study were trained only with 'clean' data and used a single system. Other approaches using HMMs have attempted to compute an explicit weighting factor to bias the recognition in favour of the visual component when acoustic noise levels increase and thus to build an adaptive system [1, 17]. Although all of the systems have demonstrated the benefits of adding a visual component to the recognition process, especially when the acoustic noise levels are high, none has yet clearly demonstrated a bimodal performance that is uniformly better than the unimodal performance in either of the two domains. This ability, conversely, is common among humans.

3.4. A cross-product HMM architecture

Conventional HMM recognizers employing DI architectures assume synchrony of the visual and the acoustic data [e.g. 17]. An architecture using triphone HMMs that lies within the DI to SI continuum has recently been reported. It allows limited and variable asynchrony between the acoustic and visual signals within each triphone, although synchrony is re-asserted at the phone boundaries [20]. The structure of the HMMs is shown in Figure 2.

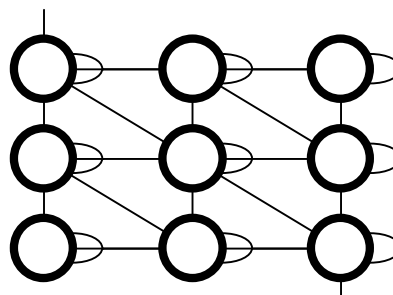


Figure 2: HMM in which the audio and visual data is decomposed into a two-dimensional array of states. Row i corresponds to state i of the audio model and column j to state j of the visual model. The parameters are computed from separately trained acoustic and visual models. Entry is via the top, left state and exit via the bottom, right state.

The speech recognition performance of this architecture (with grand variance) is shown in the fourth column of the Table below. It shows a marked improvement in performance over the 'standard', left-to-right HMMs described in Section 3.3. The improvement extends across the whole acoustic signal-to-noise range. Furthermore, the performance at virtually all acoustic noise levels is better than that obtained from either of the acoustic-only or the visual-only recognizers. Studies are now being started to examine further the effects of temporal asynchrony in the acoustic and visual signals, including their extension beyond the phone boundaries.

Acoustic SNR (dB)	Audio only	Audio-visual, standard model	Audio-visual, 'cross-product' model
23	15.3	1.7	1.3
14	14.0	3.7	4.3
5	16.0	10.7	10.7
4	67.7	22.3	19.0
-13	100.0	25.7	20.3
-22	100.0	25.7	20.3

Table: The % word error rates for grand variance HMM recognizers is shown as different levels of simulated noise are added to the acoustic signal. Visual-only recognition has a constant 23.0% word error.

The implementation of experimental systems for automatic speech recognition has already demonstrated the benefits of adding a visual component to the conventional, acoustic inputs. Cognitive models can provide a useful conceptual framework in the search for recognition architectures in which the acoustic and visual components are optimally integrated, as current work is showing.

4. REFERENCES

- Adjoudani, A. & Benoit, C. "On the integration of auditory and visual parameters in an HMM-based ASR", in *Speechreading by Humans and Machines* (Editors D.G. Stork & M.E. Hennecke), Springer-Verlag, in press.
- Brooke, N.M. & Scott, S.D. "PCA image coding schemes and visual speech intelligibility", *Proceedings of the Institute of Acoustics*, 16(5): 123-129, 1994.
- Brooke, N.M. & Templeton, P.D. "Visual speech intelligibility of digitally processed facial images", *Proceedings of the Institute of Acoustics*, 12(10): 483-490, 1990.
- Brooke, N.M. & Tomlinson, M.J. "Processing facial images to enhance speech communications: paper to the 2nd Venaco meeting, Maratea, 1991", *Bath Mathematics and Computer Science Technical Report 94-71*, pp. 18, 1994.
- Brooke, N.M., Tomlinson, M.J. & Moore, R.K. "Automatic speech recognition that includes visual speech cues", *Proceedings of the Institute of Acoustics*, 16(5): 15-22, 1994.
- Finn, K.I. "An investigation of visible lip information to be used in automatic speech recognition", Ph.D. Thesis, Georgetown University, Washington, D.C., 1986.
- Hennecke, M., Stork, D.G. & Prasad, K.V. "Visionary speech: looking ahead to practical speechreading systems", in *Speechreading by Humans and Machines* (Editors D.G. Stork & M.E. Hennecke), Springer-Verlag, in press.
- Klatt, D.H. "A digital filterbank for spectral masking", *Proceedings of ICASSP*, IEEE, 573-576, 1976.
- MacLeod, A. & Summerfield, A.Q. "Quantifying the contribution of vision to speech perception in noise", *British Journal of Audiology*, 21, 131-141, 1987.
- Massaro, D.W. "Bimodal speech perception: a progress report", in *Speechreading by Humans and Machines* (Editors D.G. Stork & M.E. Hennecke), Springer-Verlag, in press.
- McGrath, M., Summerfield, Q. & Brooke, M. "Roles of lips and teeth in lipreading vowels", *Proceedings of the Institute of Acoustics*, 6(4): 401-408, 1984.
- Mellor, B.A. & Varga, A.P. "Noise masking in a transform domain", *Proceedings of ICASSP*, IEEE, 2: 87-90, 1993.
- Petajan, E.D. "Automatic lipreading to enhance speech recognition", *Proceedings of the Global Telecommunications Conference (Atlanta, Georgia)*, IEEE Communication Society, 265-272, 1984.
- Petajan, E.D., Bischoff, B.J., Bodoff, D.A. & Brooke, N.M. "An improved automatic lipreading system to improve speech recognition", *Proceedings of CHI 88 (Washington, D.C.)*, ACM, 19-25, 1988.
- Robert-Ribes, J., Piquemal, M., Schwartz, J-L. & Escudier, P. "Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition", in *Speechreading by Humans and Machines* (Editors D.G. Stork & M.E. Hennecke), Springer-Verlag, in press.
- Sejnowski, T.J., Yuhas, B.P., Goldstein, M.H. & Jenkins, R.E. "Combining visual and acoustic speech signals with a neural network improves intelligibility", in *Advances in Neural Information Processing Systems* (Editor D.S. Touretzky), Morgan-Kaufmann Publishers, San Mateo, Ca., 2: 232-239, 1990.
- Silsbee, P.L. & Su, Q. "Audiovisual sensory integration using hidden Markov models", in *Speechreading by Humans and Machines* (Editors D.G. Stork & M.E. Hennecke), Springer-Verlag, in press.
- Stork, D.G., Wolff, G. & Levine, E. "Neural network lipreading system for improved speech recognition", *Proceedings of the International Joint Conference on Neural Networks (Baltimore, Md.)*, IEEE, 2: 285-295, 1992.
- Summerfield, A.Q. "Some preliminaries to a comprehensive account of audio-visual speech perception", in *Hearing by Eye: The Psychology of Lipreading* (Editors B. Dodd & R. Campbell), Lawrence Erlbaum Associates, Hillsdale, N.J., 3-51, 1987.
- Tomlinson, M.J., Russell, M.J. & Brooke, N.M. "Integrating audio and visual information to provide highly robust speech recognition", *Proceedings of ICASSP 96*, in press.