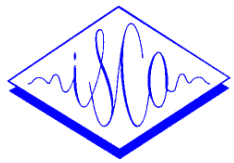

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

TWO- AND THREE-DIMENSIONAL AUDIO-VISUAL SPEECH SYNTHESIS

N M Brooke (1) and S D Scott (2)

1. Media Technology Research Centre, Department of Mathematical Sciences,
University of Bath, BATH BA2 7AY, United Kingdom
2. Now at The Asia Pacific Institute of Information Technology, 50490 Kuala Lumpur, Malaysia

ABSTRACT

An audio-visual speech synthesiser has been built that will generate animated computer-graphics displays of high-resolution, colour images of a speaker's mouth area. The visual displays can simulate the movements of the lower face of a talker for any spoken sentence of British English, given a text input. The synthesiser is based on a data-driven technique. It uses encoded, video-recorded images and sounds of a real speaker to find optimal parameter values for, or 'train', hidden Markov models (or HMMs) that capture both the sounds and facial gestures for each of the speech sounds of British English. To synthesise an utterance, the trained HMMs associated with the speech sounds are invoked in the appropriate sequence to produce outputs which can be decoded into an image and sound sequence. Whilst the basic image syntheses are two-dimensional, they can be pasted onto a three-dimensional wireframe model of the lower part of a head, which, when the jaw outline is adjusted, produces a plausible three-dimensional visual speech animation.

1. INTRODUCTION

The benefits that seeing a talker's face can bring to speech perception, especially in noise or where there is hearing-impairment, are now well rehearsed and have been quantified as being equivalent to a gain in signal-to-noise ratio of 8-10 dB [1]. This relatively modest gain can, however, represent a large improvement in word recognition rates, particularly at higher acoustic noise levels. The visual cues, which are mostly cues to place of articulation, essentially complement the acoustic cues to manner of articulation. As a result, access to visual as well as acoustic speech data can be useful, even in low levels of acoustic noise. More specifically, accurate and rapid visual syntheses of the facial speech movements have a number of potentially important applications ranging, for example, from producing training aids for language learning or speech-reading to high-speed, low-cost cartoon film-making.

Conventionally, facial speech animations have been based on three-dimensional wireframe models of the head whose conformation can be adjusted over time to simulate the facial movements. These are then rendered to produce fully shaded and textured colour images of a talking head [2]. This approach tends to exploit highly specialised graphics work-stations to attain real-time or near real-time performance. Furthermore, the derivation of suitable control data to govern the varying conformations of the wireframe models is usually difficult. For example, if phonetically-specified target conformations are specified for key frames, these have to be bonded and intermediate frames have to be interpolated, which can require hand-working and explicit models of coarticulation to account for context-sensitive articulatory variations [e.g. 3]. Alternatively there have been attempts to model the detailed anatomy of the head including the muscles and soft tissues [4]. The measurement of muscle activity is difficult and there is in any case no well-established mapping between muscle activity and phonetic descriptions of speech utterances.

2. DATA-DRIVEN SYNTHESIS

Consequently, a data-driven approach is attractive and its main principles and advantages have been set out earlier [e.g. 5,6,7,8]. In the method described in this paper, HMMs have been employed to encapsulate a statistically-based description of both the visible movements of a talker's mouth region and of the acoustic speech signals, for all the individual speech sounds. The set of visual models is trained from frontal images of a real speaker enunciating sentences of continuous speech and the set of acoustic models from the simultaneously recorded soundtrack.

By invoking the appropriate sequences of trained visual and acoustic HMMs, decoding the outputs produced and displaying them together simultaneously, audio-visual syntheses can be created. In the case of the visual outputs, no artificial model of the head is required and, more importantly, all the perceptually significant natural features of the original images, such as skin texture and shading, plus intermittently or partially visible articulators such as the teeth and tongue reappear automatically in the syntheses. There is no need specifically to measure or analyse these difficult articulators. In addition, both visual and acoustic

outputs can re-create the natural variability of real speech productions that are captured in the statistically-based descriptions of the HMMs and it is possible to generate syntheses in close to real-time using the trained HMMs.

3. DATA CAPTURE AND COMPRESSION

It is, nonetheless, difficult to manage in a short time-span the volumes of data involved in the generation of sequences of facial images unless the image data can be compressed. A reduction in the spatial resolution of images of the lower face down to about 16 x 12 pixels has been shown to retain the important visual cues to the perception of vowels. Further compression was also shown to be feasible, for example, by using neural networks [9]. However, an alternative data-driven technique known as principal component analysis (or PCA) offers a number of advantages over neural networks for image encoding and data compression and was successfully applied to monochrome images of the oral area [5]. Figure 1 shows typical tracks for PCA coefficients 1-4 from PCA-encoded oral image sequences recorded during speech production and illustrates their smooth variation with time.

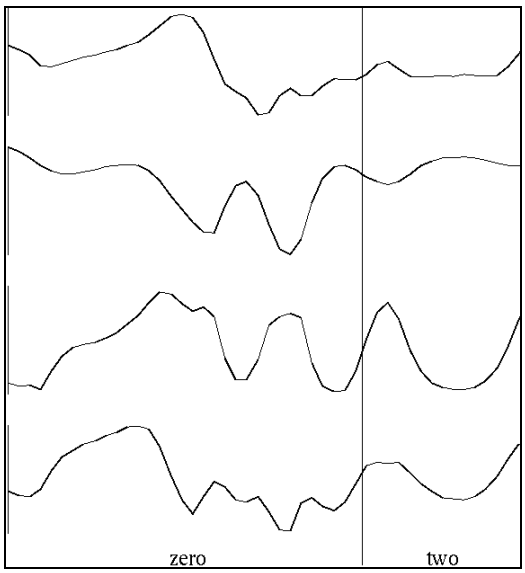

Figure 1: Time-variations of PCA coefficients 1-4 from PCA-encoded images of a speaker's oral region during the enunciation of number word sequences, as shown. The tracks move smoothly.

3.1. Visual data

Images of the lower face of a British English speaker enunciating 200 exemplars from the SCRIBE sentence list (about 20 minutes of speech) were videorecorded in colour on a BETAcam system. A simultaneous audio recording was made using the camera's condenser microphone. Each sentence was phonetically-transcribed and then manually-processed, using an Abekas A66 digital video disk, into phonetic segments labelled by the starting and finishing frame addresses.

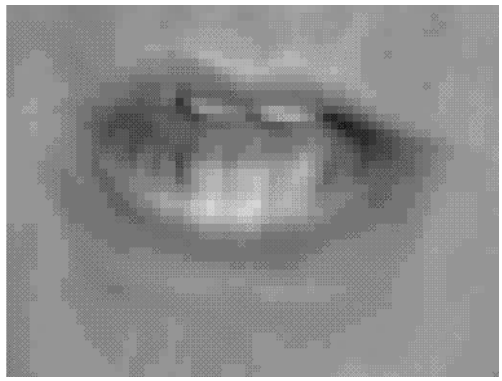

Figure 2: Example of a frame of 64 x 48 pixels spatial resolution reconstructed from an MPCA encoding (see text), as used in the visual synthesiser. Note the presence of teeth, tongue and skin texturing features.

An extended, multi-level variant of the PCA method (MPCA) was applied to the digitised colour images of the oral region having reduced this area to a spatial resolution of 64 x 48 pixels, as shown in the specimen frame in Figure 2. This is well above the lower resolution limit for retaining the essential visual cues to speech. In the MPCA scheme, the RGB-coded colour images were sub-divided into 16 sub-units of 16 x 12 colour pixels and each sub-unit was subjected to PCA and encoded so that 85-90% of the variance was accounted for, which required 30-50 components. The components from all of the sub-units were then combined and subjected once more to PCA. The first forty components from the second application of PCA were used to encode the colour images.

3.2. Acoustic data

The acoustic signal was digitised at 16 kHz and 8 ms Hamming-windowed segments were subjected to FFT. The resulting short-term spectra (40 values in the frequency range 50 Hz to 5 kHz) were smoothed and subjected to PCA, to generate a set of 20-element encoded audio vectors, together with pitch estimates. A clapper board was used to synchronise the image and sound recordings.

4. CONSTRUCTION OF TRIPHONE HMMs

Simple, left-to-right HMMs with state re-entry and sequential state transitions were used to model both visual and acoustic data. One model was generated

for each triphone (i.e. for each phone within every preceding and following phone context). Each model had three states, unless it modelled a phone following or preceding a silence, in which cases extra states were added to model visual artefacts, such as the drawing of breath. Each state of the HMMs was modelled by a single multivariate Gaussian distribution of the PCA- or MPCA-encoded acoustic or visual data, with a diagonal covariance matrix. Where sufficient examples of a triphone were available from the recorded SCRIBE database (about 700), the HMMs were refined using a standard Baum-Welch re-estimation. The remaining triphone models were trained by using the appropriate diphones from the database to estimate the parameters of a) first pair of states and b) the final state. Where diphone data was unavailable, a simple monophone model was estimated. This illustrates very clearly the volume of training data needed for a comprehensive set of HMMs.

5. GENERATION OF SYNTHESES

Shareware available on the internet for research use was used to generate phonetic transcriptions of typed input sentences. The shareware included a text-to-phoneme algorithm and a word-to-phoneme dictionary. Additional, simple phonological rules were devised and implemented to refine the word-by-word phonetic transcription, for example, by eliminating repeated phones at word boundaries. Grammatical (word class) tags and lexical stress marks were also passed forward with the phonetic transcription. The resultant, phonetically transcribed text input was used to define the appropriate sequences of HMMs that it was necessary to call up to generate outputs.

5.1 Smoothing of HMM outputs

The visual and acoustic HMMs were run independently to generate the outputs in the two modalities. Running them in this way resulted in a small and variable degree of audio-visul asynchrony, which is observed in real speech and has been found to have a significant effect in processing audio-visual signals for automatic speech recognition [10].

Using HMMs to generate unconstrained outputs does not, however, lead to smoothly changing MPCA- or PCA-tracks. These correspond to smoothly changing facial images and formant movements and are the result of anatomical and physiological constraints. It is therefore necessary to constrain the time-varying tracks of the encoded outputs and this has been described in more detail previously [8]. The method i) permitted the HMMs to be used to generate repetitions of a single utterance that could still simulate the variability of human performance; ii) generated smoothly varying MPCA and PCA tracks; and iii) suppressed over-rapid local changes in the code coefficients that in effect represented articulator velocities not achievable by a human speaker.

5.2 Duration modelling

The duration modelling within the triphone HMMs was based on the durations of the training material in the SCRIBE database. In order to be able to synthesise unrestricted sentences in which triphones could appear in contexts other than those occuring in the database, some form of additional duration modelling was necessary. This has also been described elsewhere [8] and was based on observations drawn from the sentences of the SCRIBE database. It was observed that: a) phone durations shortened in longer sentences; b) phone duration in unstressed and stressed words like nouns, adjectives,verbs, adverbs and pronouns (as opposed to those in reduced words like articles and conjunctions) was shortened in longer words; and c) phone duration of stressed vowels was lengthened. It was possible to calculate a overall speed-up factor by multiplying the speed-up factors from the three sources described above. The overall speed-up was then applied to compute the change in the expected duration of a phoneme, as represented within the transition probabilities of the trained HMMs.

6. SYNTHESISER PERFORMANCE

Once the HMMs have been trained, the selection of appropriate sequences of HMMs, their invocation and the generation of the MPCA- or PCA-encoded image sequences and speech sounds can be accomplished very quickly indeed. On a specialised graphics workstation these stages are likely to be close to real-time, even without extensive code optimisation. The current rate-limiting stage is the reconstruction of the images and speech sounds from the encoded versions, especially the decoding of the MPCA-encoded images. In current synthesisers, the entire image sequence is reconstructed and stored before initiating the animated display. On a reasonably powerful PC system using a 166mhz. Pentium II processor, audio-visual syntheses can be achieved in between 5 and 8 times real-time. The SGI Indy R4000 graphics workstations on which the synthesis development was mainly carried out are somewhat faster, but not yet real-time.

It is possible to speed up the audio-visual syntheses by using a codebook of stored images, together with their corresponding MPCA coefficients and then extracting in sequence those that are closest to the generated, MPCA-encoded outputs. It is also

possible to build synthesisers that lie in the continuum between the extremes at which: a) a codebook only is used; or b) reconstructions from generated MPCA-encoded outputs only are used. The former are very fast indeed (on an SGI Indy R4000 they are almost real-time), but tend to lack smoothness; the latter are more smoothly articulated, but are slower. The position along the continuum can in principle be fixed by setting a tolerance such that, if a codebook entry is sufficiently similar to the computed image, as expressed in MPCA space, then the codebook image is used directly; otherwise computed images are decoded and used. This approach implies that the synthesiser could be adapted to match the computing power available, which is important where wide-scale deployment is anticipated.

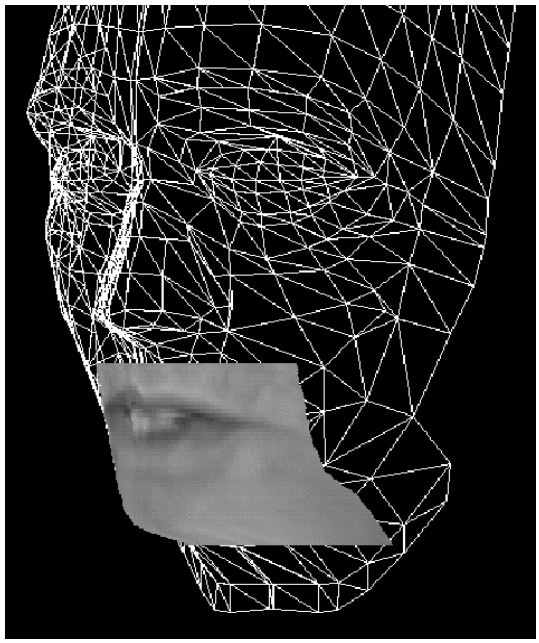

Figure 3: A three-dimensional head model, with a two-dimensional synthesised image of the oral region pasted, or projected, onto a smooth mask of the lower face.

Whilst the synthesiser can generate acoustic output as well as images of the talker's mouth, the former is not yet of a very high quality and will demand further development. The main objective thus far has been to generate high quality visual syntheses.

The acoustic element of the syntheses, however, successfully demonstrates that HMM-based methods are in principal feasible and that combined, audio-visual synthesis can be achieved via parallel and independent processing of the two modalities. Further development of the acoustic ouput system is envisaged.

7. TWO- AND THREE-DIMENSIONAL DISPLAYS

The synthesiser has been designed to show that accurate and rapid, two-dimensional animated displays of a talker's oral region are possible and the displays consist essentially of animated colour images of the kind shown in Figure 2.

However, the obvious longer-term goal, in the context of the rapid expansion of multimedia computing, is to augment computer agents such as virtual actors by providing accurate and efficient speech animation which they largely lack at present. The first step towards this goal is the synthesis of a plausible talking mouth that can be incorporated into a moveable head.

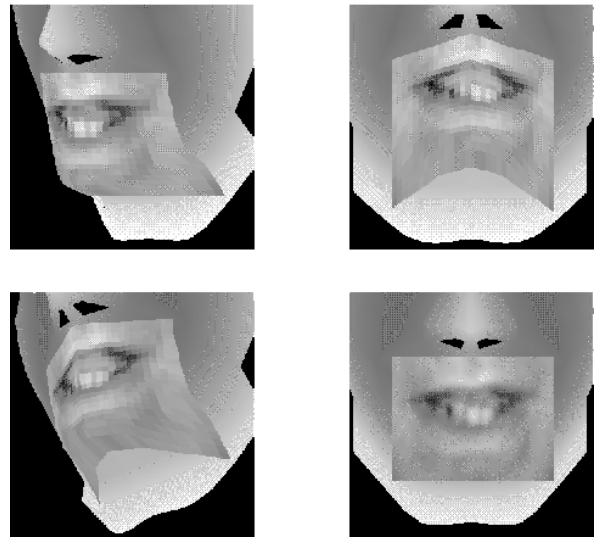

Figure 4: Illustrative frames from a prototypical talking head model, in which two-dimensional oral images generated by the audio-visual speech synthesiser have been pasted onto a three-dimensional head framework upon which they are carried as the head moves (taken from a final year undergraduate project report by J. Fisk, Dept. of Mathematical Sciences, University of Bath).

Early development work has been carried out in which the two-dimensional oral images have effectively been pasted (i.e. projected) onto a three-dimensional wireframe head which has a smooth and featureless, but otherwise accurately-shaped, oral region.

On a graphics workstation it is then possible to move the head under control of a joystick or trackball, while the animated sequence of images is displayed at the oral region. Figure 3 shows how the display is constructed, while Figure 4 gives examples of this kind of display at different facial orientations. The plausibility of the mouth images is maintained up to quite high degrees of head

movement away from a direct frontal view, even though the projection is no longer an accurate representation of what would be visible of the oral region on a real face. The extent of this illusion and the speech-reability of faces presented at varying orientations is an early candidate for further study. However, it is at least possible that this very simple method for modelling talking heads could prove adequate for most practical applications.

A later development of this idea is illustrated in Figure 5, which shows a three-dimensional model of the lower face and jaw. The oral image is now projected through an irregular aperture in the face mask from behind so that the composite image appears more coherent. The movement of the jaw was found to be very highly correlated with the value of the third MPCA coefficient and this code value was therefore used to control the position of the jaw region of the facial mask. The result is subjectively plausible.

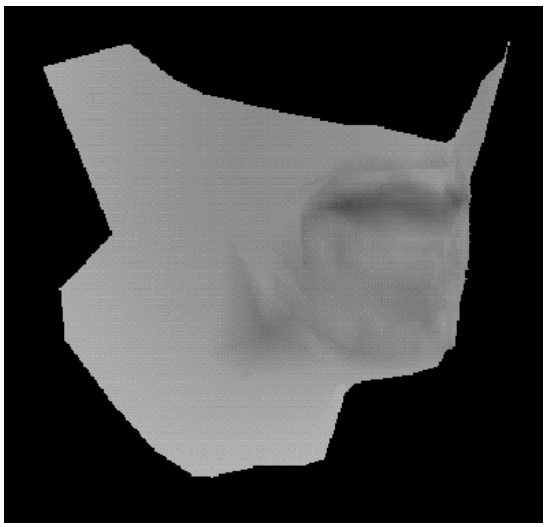

Figure 5: Development of a three-dimensional lower head model in which the oral image is projected through an aperture from behind the head mask and the jaw movement can also be controlled (see text).

8. REVIEW AND CURRENT ISSUES

Subjective tests of speech-readability were carried out on earlier versions of the visual synthesiser, using a vocabulary of digit words [6]. They suggested that a monochrome synthesiser, working at a lower spatial image resolution, was capable of being speech-read almost as well as a real face at the same resolution (32 x 24 pixels). The addition of colour made no significant performance difference, though it was found more natural by most subjects. It has also been found that there was a high correlation between the word accuracy rate in speech-reading tests and an objective measure of synthesis quality based on the difference between the MPCA tracks of real and synthesised utterances. Though it is now possible straightforwardly to set up speech-reading experiments using the current version of the synthesiser, they have still to be designed in detail and carried out. This is now an early objective. Since the synthesiser is capable of unrestricted syntheses, the experimental design is potentially complex.

In addition, the synthesiser represents one distinct way forward and there are other types of synthesiser available also [e.g. 3]. An important issue is therefore to engage in comparative studies using testing procedures that are common and appropriate to the community as a whole.

A major issue that has not yet been examined systematically is the synthesis of visual speech gestures with a range of accompanying facial expressions, such as those related to emotional states like anger, sadness or fear. If facial speech synthesis is to be successfully applied to computer agents and virtual actors, this is an important requirement. The three-dimensional version of the synthesiser, described in Section 7, has been tested with deformation of the underlying wireframe head model. For example, the regions of the wireframe lying under the corners of the mouth area have been been shifted up and down to simulate, to a first approximation, speech plus happiness or sadness. However, the projections of the mouth image are not robust to large deformations of the underlying wireframe and this type of synthesis is not, in its present form, well-suited to emotional modelling. The analytical study of speech gestures in the presence of accompanying facial expressions is in any case very limited and needs extensive investigation. There is, for example, no clear description to indicate the degree and type of coupling that may exist between speech articulation activity and other facial muscle actions.

The performance of the synthesiser was discussed in Section 6. Although real-time performace has not yet been achieved, this does not seem to be a major issue. Given the rate of development of computer processing power, a reasonable, though non-real-time, performance at the present time suggests that real-time capability, even on non-specialised systems, is likely within the near future.

9. ACKNOWLEDGEMENT

The authors thank the Engineering and Physical Sciences Research Council of the United Kingdom (EPSRC) for the grant that supported the work reported in this paper.

10. REFERENCES

1. MacLeod, A. and Summerfield, A. Q. "Quantifying the contribution of vision to speech perception in noise," *British Journal of Audiology* 21: 131-141, 1987.

2. Parke, F. I. "Parametrized models for facial animation," *IEEE Computer Graphics and Applications* 2: 61-68, 1975.

3. Cohen, M. M. and Massaro, D. W. "Modelling coarticulation in synthetic visual speech," in Thalmann, N. M. & Thalmann, D. (editors), *Computer Animation 93 (Tokyo)*: 139-156, Springer-Verlag, Berlin, 1993.

4. Terzopoulos, D. and Waters, K. "Physically-based facial modelling, analysis and animation," *Journal of Visualisation and Computer Animation* 1: 73-80, 1990.

5. Brooke, N. M. and Scott, S. D. "Computer graphics animations of talking faces based on stochastic models," *Proceedings of the International Symposium on Speech, Image Processing and Neural Networks, IEEE, Hong Kong::* 73-76, 1994.

6. Brooke, N. M. "Talking heads and speech recognisers that can see: the computer processing of visual speech signals," in Stork, D. G. & Hennecke, M. E. (editors), *Speechreading by Humans and Machines*: 351-371, Springer, Berlin, 1996.

7. Brooke, N. M. "Computational aspects of visual speech: machines that can speechread and simulate talking faces," in Campbell, R., Dodd, B. & Burnham, D. (editors), *Hearing by Eye II*: 109-122, Pyschology Press, Hove, 1998.

8. Brooke, N. M. and Scott, S. D. "An audio-visual speech synthesiser," *Proceedings of the ESCA Workshop on Speech Technology in Language Learning (Marholmen)*: 147-150, 1998.

9. Brooke, N. M. and Templeton, P. D. "Visual speech intelligibility of digitally processed facial images," *Proceedings of the Institute of Acoustics* 16(5): 15-22, 1990.

10. Tomlinson, M. J., Russell, M. J. and Brooke, N. M. "Integrating audio and visual information to provide highly robust speech recognition," *Proceedings of ICASSP 98*: 821-824, 1996.