Proceedings of the Institute of Acoustics

GENERATION OF MOUTHSHAPES FOR A SYNTHETIC TALKING HEAD.

A.D. Simons & S.J. Cox

British Telecom Research Laboratories, Martlesham Heath, Suffolk, IP5 7RE, England.

1. INTRODUCTION

Videophones have been available from BT for many years but they suffer from the disadvantage of requiring a large bandwidth for transmission. Recently, a facial image coding system has been developed at British Telecom Research Laboratories (BTRL) which is capable of producing lifelike movements of the head and face at data rates low enough to be transmitted over ordinary telephone lines. Such a system could be used to provide (amongst other applications) a pseudo-videophone, if realistic mouth movements matching the transmitted speech could be supplied. This paper describes a technique for generating mouth movements for a synthetic face based on analysis of speech data. The technique has been incorporated into a real-time demonstrator.

2. IMAGE SYNTHESIS SYSTEM

Model based coding methods [7] code an image or sequence of images using knowledge of the scene to model the objects in the scene. The motion of the objects is determined, and this motion reproduced in the models. This creates a synthetic image sequence which matches the original sequence.

One of the most popular areas of research in this area is in the generation of synthetic facial image sequences [4][6]. A real-time synthetic face generation system is now in operation at BTRL [8]. This system produces facial expression changes using a method based on Ekman and Friesens' system of action units [1]. An action unit defines a single facial movement such as a blink, eyebrow raise or lip purse. Each action unit can be invoked with a parameter which defines the proportion of that action unit to be applied to the model.

The ability to control such high-level actions as mouthshape, jaw-movement, eyebrow position etc. raises the intriguing possibility of animating a facial image to make it apparently 'speak' an incoming speech signal. This technique would find applications as a pseudo videophone, in cartoon animations, recorded announcements etc. Clearly many levels of sophistication are possible in what is detected in the speech signal and conveyed in the image. This paper concerns itself with the basic requirement of generating realistic mouthshapes. In the demonstrator, these are supplemented by rotations and translations of the head, blinks and eyebrow movements. These movements are random (within the constraints of realism) and contribute greatly to the naturalness of the demonstrator.

GENERATION OF MOUTHSHAPES FOR A SYNTHETIC TALKING HEAD

3. DESCRIPTION OF THE TECHNIQUE

One possibility for generating mouthshapes is to recognise (using some speech units) the incoming speech and then transmit facial image codes which correspond to these units. It is likely that such recognition would need only a few broad acoustic or phonetic classes to obtain realistic facial images. In practice we do not need to make the decoding of the speech explicit because we can use training data in which every video frame has been labelled with an image code (coding the mouthshape) and the corresponding speech frame has been labelled with a code derived from a speech vector-quantiser. These correspondences can be used for automatic training of our model; the process of generating the facial image codes for some speech input is then one of deciding what the most likely sequence of image labels is, given a sequence of speech labels. Such a model needs to exploit information about both the co-occurrences of image and speech symbols and also about likely sequences of image symbols (the latter is required to generate a smooth sequence of output symbols). An attractive method of combining both types of information is a fully connected Markov Model (MM) (Figure 3.1).

Here each state is uniquely identified with an image code and also has associated with it a set of probabilities of emitting each of the speech codes

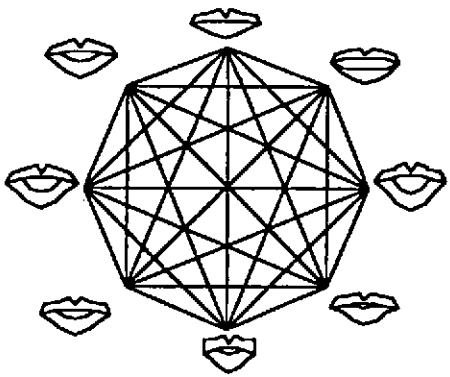

Figure 3.1 A fully connected eight state Markov Model (states represent mouthshapes).

Note that because the states are explicitly identified at training time the model is not "hidden" at training time and conventional techniques for optimisation are not required. However, the model is "hidden" when it is required to decode the speech signal, in the sense that the most likely sequence of states (mouthshapes) which generated the speech data is estimated. A further advantage of using an MM is that the standard Viterbi decoding algorithm can be used to estimated this sequence.

An utterance produced by the model can be visualised as a sequence of states with explicitly associated mouthshapes emitting a sequence of speech segments (figure 3.2).

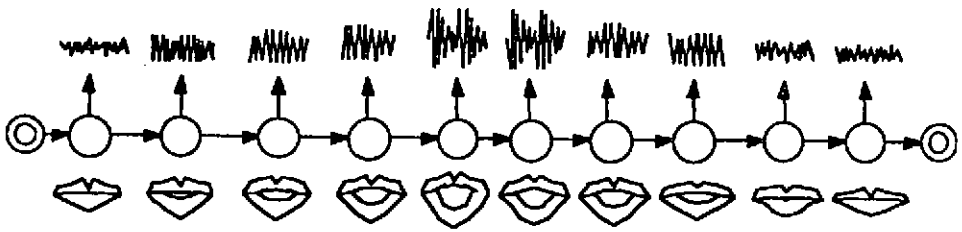

Figure 3.2 A diagrammatic representation of a model producing speech.

4. TRAINING THE MODEL

Synchronous speech and image data was collected from three speakers. In the experiments described here the data from only one speaker was used, because as yet only one data set has been fully annotated with the mouth height and width parameters.

The speech data was band limited between 50 Hz and 3.5 KHz. A pre-emphasis filter was applied and the speech was blocked into 40ms frames (320 samples). The frame overlap was 20ms. The Hamming window size was chosen to obtain parameters at a convenient multiple of the video frame rate, in this case double. A number of parameterisations of the speech data were tried; LPC coefficients, LPC reflection coefficients, LPC cepstral coefficients and Mel Frequency Cepstral Coefficients (MFCCs). A comparison of these different parameterisations is given in section 6.3.

During the recording the teeth of the subject were painted using a black spirit-based dye so that the inner lip margins could be reliably distinguished from the teeth. The subject was seated so that the mouth region of the face filled the frame when the more extreme vowels were spoken. Two 1KW quartz halogen floodlights were placed either side of the camera, and the lighting level was adjusted so that the mouth area would appear black and the rest of the image white when a thresholding algorithm was applied. The image data was then processed using a semi-automatic procedure to measure the height and width of the mouth opening. The data was captured at a rate of 25 frames/s so a set of parameters was obtained every 40ms.

Fifty phonetically rich sentences were recorded, each sentence being approximately four seconds long. This gave 10,000 vectors of speech coefficients and 5,000 image data vectors. Both sets of data were vector quantised using the same codebook generation and vector quantising algorithms. 16 image-codes were found to be adequate to represent the range of mouth movements, and 64 speech codes were used. This gives a MM with 16 states, each representing a particular vector quantised mouthshape. Each mouthshape has a probability associated with producing each speech vector. The model transition probabilities were estimated directly from the image data, and the speech code output probabilities for each state from the joint occurrences of image and speech symbols.

5. USING THE MODEL TO GENERATE A MOUTHSHAPE SEQUENCE

Once the model has been trained it is possible to calculate the most likely sequence of mouthshapes given a particular utterance. This is done using the Viterbi algorithm [5]. Given a sequence of vector quantised speech segments and a MM as described, the Viterbi algorithm calculates the most probable path through the model, which is equivalent to the most probable mouthshape sequence.

The mouthshape sequence calculated using the Viterbi algorithm is translated into action units with associated proportions using a simple lookup table.

6. EXPERIMENTS AND RESULTS

6.1. Error Measures

While it is possible to measure the percentage of time that the mouthshapes (image codes) predicted by the system match the vector quantised mouthshape in the original sequence exactly, this does not give a particularly good idea of how good the end result will appear. For example, if 50% of the mouthshapes chosen are correct and the other 50% are very different from the required mouthshape the synthesised sequence will be subjectively less acceptable than a system where 25% of the mouthshapes are correct and the other 75% are close (in some sense) to the required mouthshape.

Where percent correct is used as an error measure the percentage of time that the chosen mouthshape was in the top N candidates is also quoted. The second best mouthshape is defined to be the mouthshape with the smallest Euclidean distance between its parameters and those of the required one, the third best is the next smallest and so on. The other error measure which is used is mean squared difference (msd). The squared differences between the height and width of the required mouth and that of the chosen mouth are calculated, and the mean squared difference is calculated for the whole utterance. The msd which would result from an entirely random selection of mouthshapes (the expected msd) is also shown for reference. It is worth noting that an msd of half the expected msd is that which would be achieved by a system producing an error rate of 50% with incorrect mouths being chosen randomly.

6.2. Experiments

Two main sets of results are presented. The first set (figure 6.1) compare different speech parameterisations for this problem For these, the model was trained on the first 40 sentences in the database, and tested on the final 10. The second set (figure 6.2) show the effect of varying the amount of training data. Here, the training set size was initially 10 sentences and was increased to 20,30 and 40 while the test set was correspondingly decreased from 40 to 10. These experiments were carried out on all of the speech parameterisations. A summary of results are shown in the next section.

6.3. Results and Discussion

A comparison of four different speech data parameterisations is shown in figure 6.1. It can be seen that the LPC Cepstral coefficients give the highest percent correct and correspondingly the lowest msd. MFCCs give the worst results. This may be because LPC coefficients reflect the vocal tract shape and therefore the mouthshape better.

GENERATION OF MOUTHSHAPES FOR A SYNTHETIC TALKING HEAD

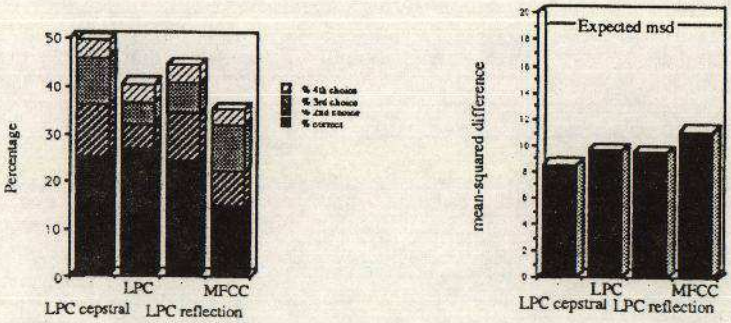

Figure 6.1 A comparison of different front-end parameterisations.

The effect of varying the amount of training data was investigated. The results were much as expected; increasing the training set size increased the error rate on the training set, and decreased the error rate on the test set. The percent correct seemed to level out at about 25% after 40 training sentences had been presented. This is not conclusive but there was no more data available for further tests.

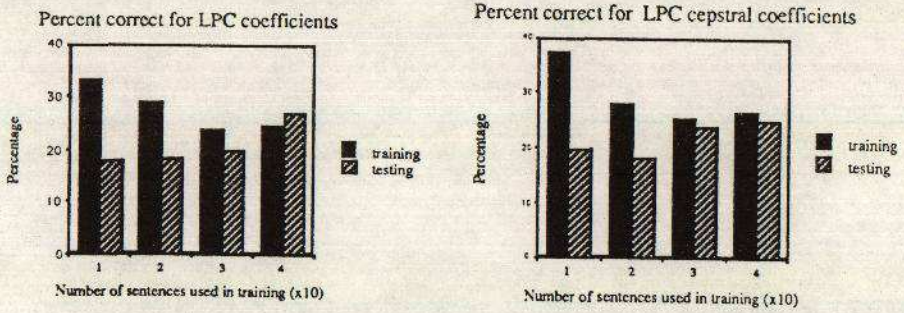

Figure 6.2 The effect of increasing the amount of training data on percentage of correct mouthshapes chosen.

GENERATION OF MOUTHSHAPES FOR A SYNTHETIC TALKING HEAD

7. REAL-TIME IMPLEMENTATION

The simplest implementation of the Viterbi algorithm waits until the speech is ended before tracing back, which is clearly not practical for any real-time application. Partial traceback (as described in [2]) waits until all the paths have converged before making a decision about the state sequence. Any delay introduced into our system must not vary, otherwise synchronisation of speech and image becomes very difficult. A constant traceback (maximum size 150ms to enable normal conversations to take place) was therefore performed after processing each frame.

Experiments were also carried out in which the traceback extended beyond the delay at which the last mouthshape was transmitted. At the delay point the mouthshape currently the most probable for that timeslot is transmitted to the hardware, but a different mouthshape might later become optimal for that timeslot if a different path subsequently had a higher probability. This method has the advantage of taking more information into account for any given mouthshape, but the inherent smoothness gained by using a single path could be lost. The two methods are shown diagrammatically in figure 7.1.

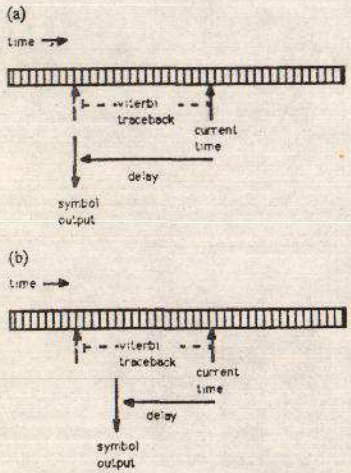

Figure 7.1 Two methods of calculating the mouthshape to be transmitted.

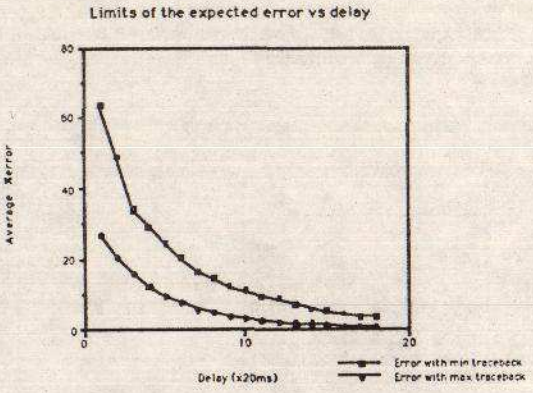

Figure 7.2 Maximum and minimum errors (compared with full traceback) for a given delay.

Figure 7.2 shows the % difference between the sequence generated using full traceback from the end of the sentence and the sequence generated using a fixed delay. The top line shows the error using a delay as in Figure 7.1(a), the bottom line shows the error with a fixed delay, but calculating the traceback to the beginning of the sentence on every occasion. There is some advantage to be gained from using as much traceback as possible. Subjective viewings of sequences generated using this method show that no noticeable discontinuities are introduced. The delay used in the final system must be determined by the maximum acceptable to the users.

8. IMPLEMENTATION OF A REAL TIME DEMONSTRATOR

A hardware demonstrator has been produced which is capable of generating synthetic facial images in real-time. This uses a Texas Instruments TMS 320C30 DSP and an Application Specific Integrated Circuit (ASIC) developed at BTRL. This image processing hardware is connected to the speech processing unit via an RS232 serial link.

The speech processing unit is PC based, and uses a 56001 DSP card supplied by Loughborough Sound Images Ltd. The speech signal is fed into the 56001 via an A-law codec, the speech parameterised, vector quantised and then processed by the Viterbi algorithm. The packets containing the action units are transmitted to the image processing hardware. The PC adds extra action units to these packets to create realistic movements of the head such as rotations, translations eyebrow movements etc.

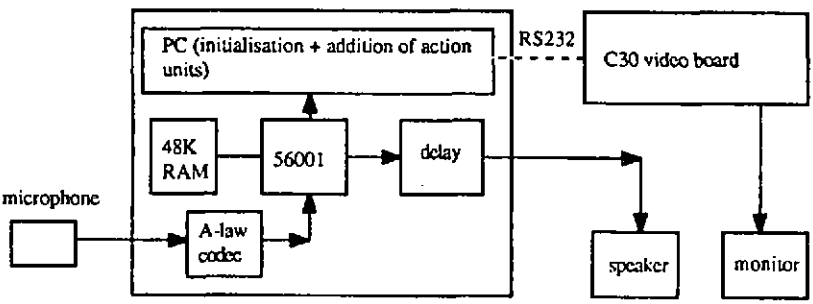

Figure 8.1 Block diagram of the real-time talking head demonstrator.

9. CONCLUSIONS

This method of generating mouthshapes for a pseudo-videophone application produces a realistic sequence of mouthshapes with a delay which is acceptable for normal conversations. While the percent correct mouthshapes chosen is relatively low, the mean-squared difference figures show that wrong mouthshapes are often close in height and width to that of the correct mouthshape. Informal demonstrations have shown that the system is generally considered acceptable, at least for speaker dependent use. For a pseudo videophone application, the next important step is to make the system speaker independent over telephone lines.

An important question is : "How close to real mouthshapes must the generated mouthshapes be to make them acceptable to users in a given application?". When the answer to this question is known (and it can be answered only by rigorous subjective testing) it may stimulate further research into topics such as finding representations of speech which model mouthshapes better or learning higher-order statistics in the model.

10. ACKNOWLEDGMENTS

Thanks to Mike Shaw for the DSP implementation of the algorithm. Also to all the members of RT5232 for many discussions about the mouthshape generator.

11. REFERENCES

[1] P.EKMAN, W.V.FRIESEN, 'Manual for the facial action coding system', Consulting Psychologists' Press, Palo Alto, CA, 1977.

[2] J.N.HOLMES, 'Speech synthesis and recognition' Van Nostrand Reinhold (UK) Co. Ltd.

[3] S.MORISHIMA, K.AIZAWA, H.HARASHIMA, 'An intelligent facial image coding driven by the speech and phoneme.' Proc IEEE ICASSP, pp 1795-1798, 1989.

[4] F.I.PARKE, 'Parameterised models for facial animation', IEEE Computer Graphics and Applications Vol 12 Nov 1982, pp61-68.

[5] L.R.RABINER, 'A tutorial on Hidden Markov Models and selected applications in speech recognition.', Proc IEEE, Vol 77,No2, pp257-286, 1989.

[6] K.WATERS, 'A muscle model for animating three-dimensional facial expression', Proc SIGGRAPH 1987.

[7] W.J.WELSH, S.SEARBY, J.B.WAITE, 'Model-based image coding', British Telecom Technical Journal, Vol.8 No.3, July 1990.

[8] W.J.WELSH, A.D.SIMONS, R.A.HUTCHINSON, S.SEARBY, ' Synthetic face generation for enhancing a user interface'. IMAGE COM '90 First International Conference on new image chains. Nov 1990.