SYNTHESISING FACIAL MOVEMENT: DATA CAPTURE

R.E.E.Robinson
Speech Hearing and Language Research Centre
School of English and Linguistics
Macquarie University

ABSTRACT - A method of data capture and analysis of lip movement is described. Cine film was digitised by computer and a frame by frame comparison was done to generate a codebook of lip shapes for the synthesis of facial movement. By various methods of data reduction, real-time image playback was achieved.

INTRODUCTION

The Speech Hearing and Language Research Centre (SHLRC) at Macquarie University has a project under way which will synti..sise facial movements in concert with a speech synthesiser. This is a project to produce a realistic talking image coupled with a voice using the unlimited vocabulary Text To Speech (TTS) system also under development at SHLRC. The TTS system takes as input any typed text and has a set of inbuilt rules that provide a sequence of parameters, to control a speech synthesiser. The same sequence will be used to control a facial synthesiser when this project is complete.

The current approach uses a natural image that has been captured on film at 100 frames per second, digitised with a frame grabber on a Sun 4/330 computer under SunOs V4.1.1 and Version 2 of Openwindows. The frame grabber is a Data Translation DT1451 which provides 512x512 resolution with a 256 level grey scale, and uses the DT IRIS subroutine library with control programs written in the C programming language. The monitor is a Mitsubishi C-3910 RGB monitor and the video camera is a colour Hitachi FP7.

To overcome some of the computing limitations imposed by current technology, the disc storage was reduced, and synthesis speed increased by Area Of Interest (AOI) processing and frame by frame comparison, subtraction and storage. The captured images were analysed to look for the rapid movement associated with some fricatives. This method produced a large amount of redundant data for other slower articulations, especially steady state vowels.

The captured data has been partially analysed, and will form the beginning of a data base that will contain lip positions that correspond to speech. A code book will be devised to correlate the lip positions to the controlling parameters coming from the TTS controlling program. The code book will be manageable as estimates of different lip shapes vary from 15 (Storey and Roberts, 1988) to 21 (Benoit et al 1992).

FILM

The original cine film was shot on black and white film (AGFA GEVAERT GEVAPAN 36) with a Bolex 16mm cine camera that was modified to expose a new frame every 10ms (100 frames per second). A purpose built digital counter using high output Light Emitting Diodes (LEDs) was constructed and synchronised to the camera such that each frame had a unique number visible in the corner. The voice was recorded on audio cassette. The subject was prompted by cards which were placed on a floor stand and then revealed as each new word was to be spoken. A mirror was placed in the field of view so that the cue cards could be seen (in reverse) to show the viewer which word was being spoken. A head restraint allowed the subject to lean their forehead against it in a comfortable and stable position. A side mirror was angled at 45 degrees to show a profile view of the subject. A graduated rule was fixed next to the subjects mouth to allow scaling if required, and arranged to be visible in both views. The camera was arranged to show the front and side of the subjects face and moved as close as possible to fill the frame. The film was shot as a component part of a physiology study yet to be published in its entirety. At the same time the film was being shot, the subjects had EMG electrodes attached to several facial muscles which allowed the speech, muscle, and frame count to be recorded on computer magnetic tape. The speech consisted of 7 consonants (P B M S Z ZH SH) and 3 vowels

(I A OR) arranged to cover all permutations, and three subjects were recorded speaking these nonsense words.

FRAME CAPTURE

The film was transferred to video by a commercial company using a flying spot scanner. The initial method of projecting the film image onto a ground glass screen proved to be unsatisfactory. This method consisted of a Kodak PE1A Print Copier, with a video camera mounted perpendicular to the screen. The image was degraded in two significant ways. The centre of the screen had a higher light intensity than the periphery making the exposure uneven throughout the frame, and the image lost resolution due to the ground glass screen. The Print Copier had no positive frame clamping, which made vertical alignment imprecise. After much trial and error this method was abandoned. The flying spot method produced a video tape that could then be played with a Video Cassette Recorder (VCR) and the freeze frame function used to stop a frame for digitisation by the computer. Figure 1 shows a typical frame. It was found that VCRs differed remarkably in their ability to provide a stable frame that the frame grabber could capture. Capturing a frame to computer was nontrivial. This was despite the frame grabber manufacturer's claims that their circuitry was designed to grab frames from VCRs. The resultant captured image was sharp but grainy, with an individual frame number, as per the original. The vertical and horizontal alignment then needed adjustment as the flying spot process, although better than the Print copier, had not resulted in perfect alignment between subsequent frames. A frame at the start of the sequence was recorded and used as a master frame, and all subsequent frames compared to this. The two frames were subtracted and then the differences were highlighted by manipulation of the output Look Up Table (LUT). The second frame was then moved pixel by pixel in a horizontal and vertical direction until the differences between frames was reduced to lip movement, frame counter increment, and random noise. On those frames that aligned perfectly, there was a random spot pattern across the image due to the grain in the film. The head restraint and scaling rule were used as the datum. Figure 2 shows the comparison of 2 frames with the differences highlighted. This frame has the same spatial features as Figure 1. The cluster of dots in the top left hand corner is the difference in the frame count. The small horizontal line 75% down from the top of the image, and 75% across from the right hand side of the image, is due to lip movement. The frame was then saved to disc as an individual file with a unique name for later analysis. The file format was DT IRIS. Each frame used 256Kilobytes (256K) of storage when stored unencoded or 437K (approximately) when Run Length Encoding (RLE) was used. The value of RLE is especially good for graphic images, normally resulting in over 50% data reduction. In the case of a realistic image, with a grey scale, it has little value.

ANALYSIS

Since only a small proportion of the image was changing from frame to frame, Area Of Interest (AOI) processing was used, both to reduce disc storage and to achieve higher frame update rates. There were 3 areas where the image changed between frames. The frame counter, the profile view of the lips, and the front view of the lips. A subroutine was written to show a box on the screen xored with the image, that could be moved around to different areas of the screen. The region within the box was the AOI to be used. Two sizes of box were tried.

A large box of size 128x128 pixels was centered on the front view of the mouth and included the nose and jaw (in the vertical dimension) and the cheeks and scaling bar (in the horizontal dimension). This reduced the image storage to 16K per frame rather than 256K for the whole frame. Each sequential AOI was saved in a buffer equal to the original frame size and saved in the same manner as a normal frame. This allowed all the existing DT IRIS library routines to be used and provided a 16:1 data compression and subsequent speed up of data processing. Figure 3 shows a frame containing a sequence from the utterance PORB. The sequence is left to right, top to bottom. In the first 2 regions, the mouth is closing for the beginning of the stop /b/ and the next 12 show the occlusion. The final 2 regions show the mouth opening for the release. The sequence PORB required 128 frames which used 32Megabytes (M) of disc storage. With AOI processing and the 128x128 region, it required the master frame and 8 other frames each containing 16 regions, or 9 frames total. This used 2.25M of disc storage.

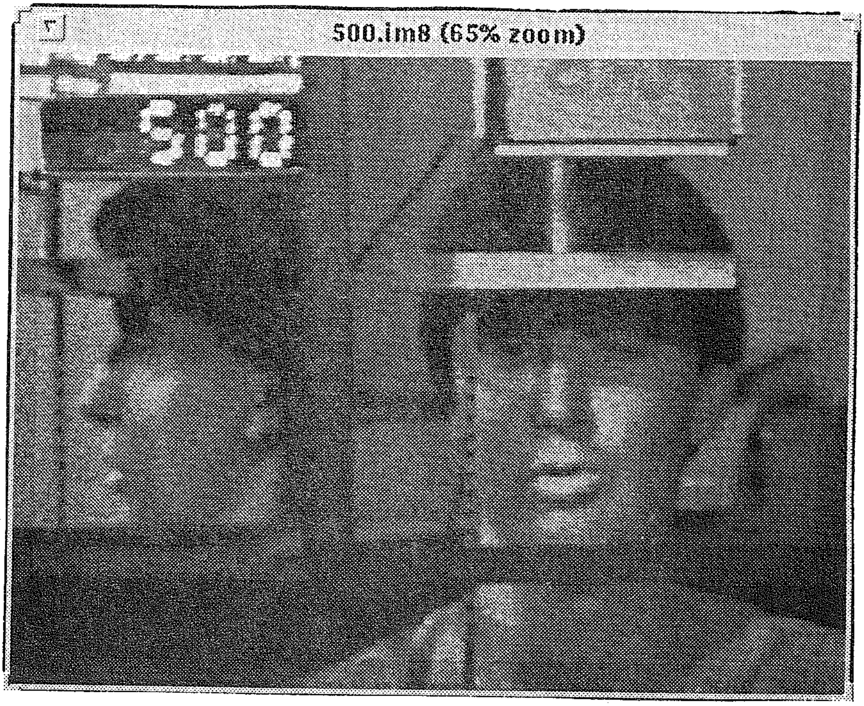

Figure 1. Frame 500 from Cine Film

A small box of size 64x64 pixels was centered of the front view of the lips. This reduced the storage to 4K per frame rather than the 256K for the whole frame. Each sequential AOI was saved in a similar manner to the larger AOI and achieved a data reduction of 64:1 over the original frame and a corresponding speed up of data processing. Figure 4 shows a frame containing the same but extended sequence from the utterance PORB. This sequence is also left to right, top to bottom. The first 16 regions show the occlusion of the /p/ followed by the vowel. The second last row shows the /b/ occlusion. The last 5 regions of the last row show the /b/ release. This reduced the data storage to the master frame and 2 other frames each containing 64 regions, 3 frames total, or 0.75M of disc storage.

PLAYBACK

The stored images can be played back as entire frames, as a master frame with the large AOI region continuously updated, or as a master frame with the small AOI continuously updated. Double buffering was used to allow the hidden frame to be updated, before being viewed. The update image appeared seemingly instantaneously. Without this the viewed image could be seen to be updated and this caused a distracting wave like discontinuity to roll down the screen. When the entire frames are played back, the sequence is not real time, and the differences in frame grain are perceptible. Some frames were Low Pass Filtered (LPF) to reduce the grainy effect, causing a softening the image.

It requires 76 seconds to replay a sequences of 64 frames, which corresponds to 640ms (100 frames per second original filming rate). When the large AOI regions are played back, discontinuities can be seen where the mouth AOI joins the master image. This can be reduced by Low Pass filtering the image in this area. It requires 25 seconds to replay a sequence of 64 frames. When the small AOI regions are played back, similar discontinuities are visible. It requires 2 seconds to replay a sequence of 64 frames. By reducing the number of frames replayed to one quarter (i.e. 25 frames per second,

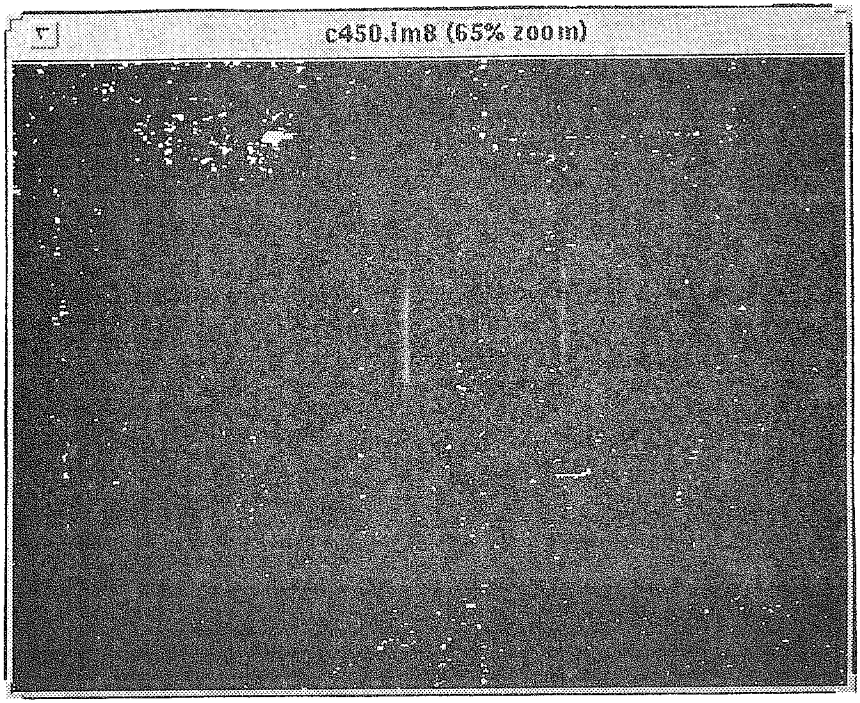

Figure 2. Comparison of Two Sequential Frames

this is similar to video speed) real time playback is achieved. The image appears realistic and no obvious flickering is evident.

COMPARING CINE TO VIDEO

A test run was shot on video to compare to the cine film. The sequence was easier to setup and record than the cine film, but the frame rate was slower. A colour video camera (Hitachi FP7) and VCR (NEC 895EA) was used to record a subject in a similar manner to the cine film recordings. A circuit was built to extract the vertical synchronising signal from the camera composite video and this signal was used to increment the frame counter. Since it was normal video, with interlaced scanning (2 fields = 1 frame), the extracted field signal was divided by 2 to give the correct frame rate of 25 frames per second (40ms per frame). This signal correctly incremented the counter and ensured there was a unique number on each frame. When the video was examined, it was found that the image quality was as good as or better than the cine film, there being no film grain influence. In the parts of speech that have rapid lip movements, there was a reduction in the number of frames available for analysis. On cine film, the word PORB showed 12 frames of the lips closed for the /b/ occlusion. The video film showed 2 frames. The video also showed 2 frames of the lips closed for PEEB, PARB, ARPAR, and 3 frames for ARPAR and ARMAR. This corresponds to the relative frame rates. Faster video cameras would overcome this. However, this appears to have a relatively minor impact on this application.

FURTHER DEVELOPMENT

The next step is to capture, analyse, and compress the rest of the cine film. The data will be categorised in the code book for logical retrieval. The rules for resynthesis are a many to one correlation, which the literature has already covered. However, I suspect there are many traps and

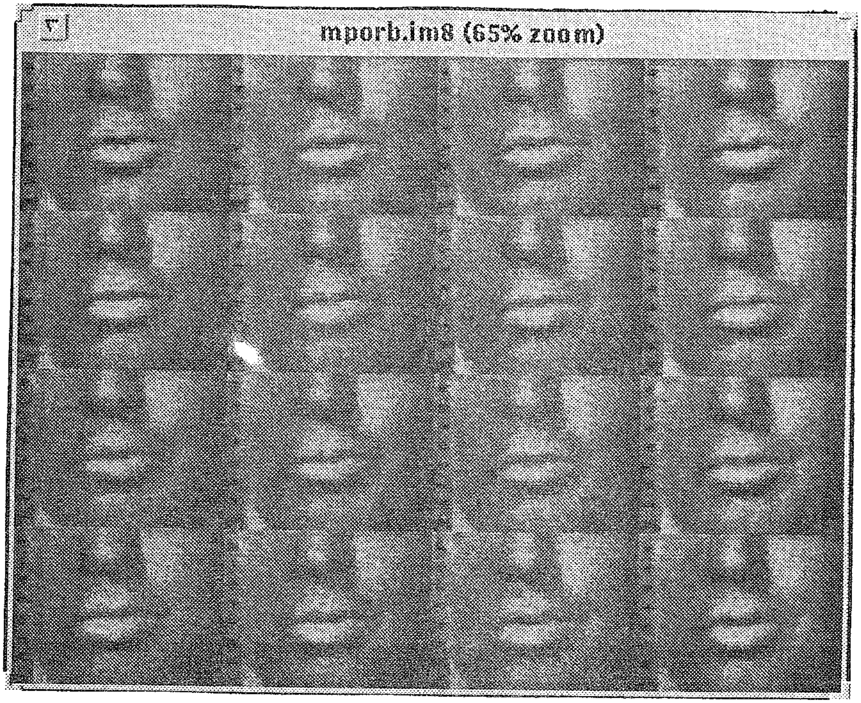

Figure 3. Mouth Positions for the /b/ in the Word PORB

inconsistencies. The approach will be to store the sequences, not as lineal paths from rest to utterance then to rest, but to build a matrix of lip shapes. The matrix will have coordinates that will be analogous to vowel space. This will enable the rules to be developed as pointers to the matrix, rather than a 1 to 1 correspondence of sequences. This will also overcome to some extent the problem of a fixed resynthesis for each identicle utterance. The pointers may be biased left or right, up or down, to achieve more or less lip spreading, greater or smaller lip rounding. The matrix will have coordinates that offer lip spreading as the X axis, and lip rounding as the Y axis. By moving top to bottom, in different columns will produce different lip area openings. By moving diagonally, tight lip rounding will change to an open mouth oval shape. Other matrices will be required for other lip dynamics. The object will be to offer great flexibility, and not be restricted to behaving like a tape recorder and merely replaying a prerecorded sequence.

The aspect of moving lips in a master frame, provides a static face. While this may be of use to restrict data variables for some experiments, it will not offer any realism and may lead to the viewer losing interest or concentration. While not wishing to add to the matrices the dimensions of expression or emotion, some normal facial operations like blinking or looking down or away will be relatively simple. A matrix of eye positions will enable a variety of expressions to be conveyed using an identicle mechanism to the mouth dynamics.

APPLICATIONS

Facial synthesisers will have as many uses as speech synthesisers. As small computers have simple speech synthesisers, and research establishments have sophisticated synthesisers, so will be required a range of facial synthesisers to match the applications requirements. The telecommunications industry is still exploring methods of audio data reduction and channel bandwidth economies and a synthesiser

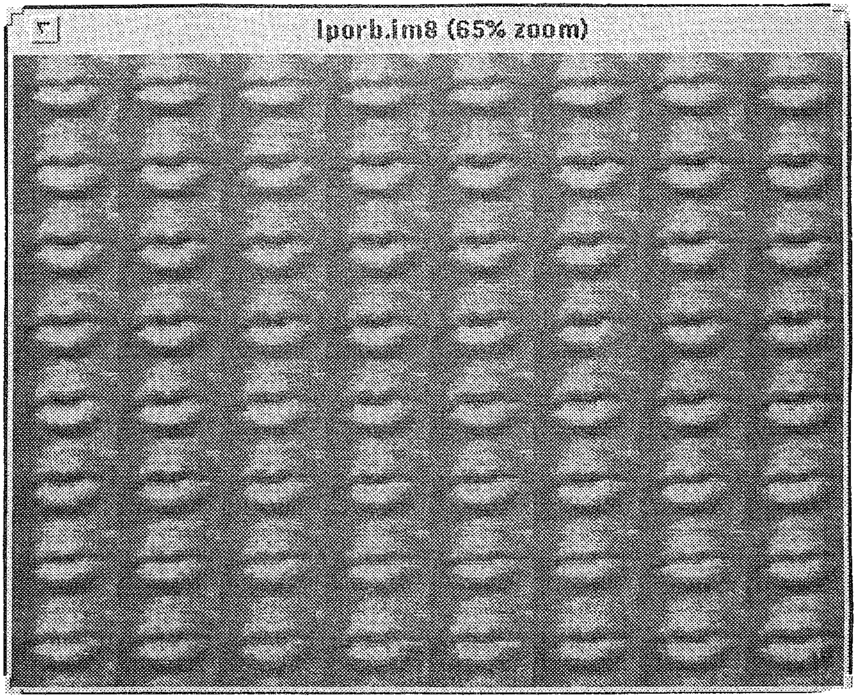

Figure 4. Lip Positions for the Word PORB

is a useful tool to generate repeatable and controllable tokens. Images will be required to be reduced to save bandwidth, and facial synthesisers will be a useful tool. Speech recognition will benefit from lip information to distinguish some ambiguities in acoustic cues, particularly where there are competing signals causing degradation.

CONCLUSION

The data required for generating a code book for a facial synthesiser at first appears simplistic, but as frames are gathered, the task begins to look insurmountable. With data reduction, slower frame rates, and data processing of the salient features, real time playback and data storage becomes manageable.

ACKNOWLEDGMENTS

My thanks go to Robert Mannell, Jonathan Harrington, Chris Callaghan, John Telec, Mark Stevens, Harry Purvis, Andrew McVeigh, Steve Cassidy, Jim Powe, and the Macquarie University Workshop for help and support during this project.

REFERENCES

Storey, D. & Roberts M. (1988) *Reading the Speech of Digital Lips: Motives and Methods for Audio - Visual Speech Synthesis,* Visible Language Vol XXII Number 1, 113-127

Benoit, C. Lallouache, T., Mohamadi, T., & Abry, C.(1992) *A Set of French Visemes for Visual Speech Synthesis,* Talking Machines: Theories, Models, and Designs, Eds. Bailly, G. Benoit, C. & Sawallis, T.R. (Elsevier Science Publishers B.V.)