SYNTHESISING FACIAL MOVEMENT: DATA BASE DESIGN

R.E.E.ROBINSON

Speech Hearing and Language Research Centre
School of English and Linguistics
Macquarie Univeristy

ABSTRACT - This article describes the design and construction of a database for lip movement during speech. A set of Australian consonants and vowels were arranged into nonsense words and recorded on cine film and video tape. The images of these words were then digitised and analysed to form diphone pairs. The visual diphones were arranged in a database with an entry lookup table for access and a set of transition tables of points of similarity. This allowed phonetic strings to be translated into lip movement sequences.

INTRODUCTION

The video capabilities of computers and the channel capacity of communications systems is improving to the point where high bandwidth video information is able to be added to traditional audio only channels. In order for Speech Synthesisers to enter this multimedia arena, they need a video realisation of the synthesised speech. A Text to Speech (TTS) system needs the added versatility of providing control of a lip movement generator in addition to the sound generator. Benoit (1992) calls this a Text to Audio Visual Speech (TTAVS) synthesiser and concludes that speech intelligibility will improve, especially in noisy environments. As the bandwidth of communications channels is not infinite, some data reduction is required, to not only make realtime facial synthesis possible, but to improve channel usage and efficiency.

One way to incorporate facial synthesis, is to completely synthesise a lip movement sequence and synchronise this with the acoustic signal. This is difficult as the compute load and display task is formidable. Synthesis in real time is usually only possible with stylised facial images. Another way to produce facial synthesis, is to have a limited set of facial images prestored and attempt to construct all possible lip positions from this set. In both of these cases, and in many other strategies, some trade off between versatility and a realistic image is necessary.

This article describes the design and construction of a database suitable for synthesising lip movement during speech. The database consists of a master image of a face, to which is added sub images of different lip positions. This is similar to the method used by Storey & Roberts (1988). Only the lip images are changed during speech. This drastically reduces the data required and consequent channel bandwidth. The database has a set of tables to allow the appropriate lip images to be found and assembled.

INPUT CONVERSION TABLE

The output of the Speech Hearing and Language Research Centre's TTS system provides a string of phonetic symbols that attempts to describe the input text. This string is used to drive the existing acoustic part of the speech synthesiser. It will also be used to access the database. An input lookup table is used to provide the location of the lip sequences and a translation of the phonetic symbols to the reduced set of visible diphthongs.

A SET OF LIP POSITIONS

The parts of the speech mechanism that are visible, are principally the lips and the teeth, with occasional appearances of the tongue tip. The tongue positions for the vowels are largely hidden, so that the set of lip positions for vowels is small. The consonants usually require more lip movement than the vowels and so consequently require a larger set.

Plant (1977) examined the visual perception of 20 Australian consonants, and 20 Australian vowels

STIMULUS

RESPONSE	p	b	m	f	v	w	r	ð	θ	l	n	j	g	k	s	d	t	dʒ	ʃ	tʃ
p	61	61	62																	
b	28	33	31																	
m	10	5	6																	
f				91	85															
v				5	11															
w						87	91													
r						2	2													
ð								52	42											
θ								43	49											
l										86	43	38	23	17						
n										2	6	4	6	1						
j										0	3	5	3	4						
g										2	5	5	7	9						
k										3	19	22	50	58						
s															58	39	25			
d											8	7			5	11	7			
t											11	10			17	25	16			
dʒ															7	7	22	45	42	33
ʃ																		26	27	29
tʃ																		26	29	35

Consonant Confusion Matrix

STIMULUS

RESPONSE	i	ɪ	ɛ	ɪə	a	aɪ	ɛə	ʌ	æ	eɪ	ɜ	ɔɪ	u	ju	ə	ʊ	ɔ	ɒ	oʊ	aʊ
i	75	55	30	35																
ɪ	5	40	35	15																
ɛ	0	0	20	10																
ɪə	0	5	5	10																
a					80	40	35	25	10	10										
aɪ					0	30	10	10	0	0										
ɛə					0	0	20	0	0	5										
ʌ					5	5	5	50	45	15										
æ					5	10	10	0	35	15										
eɪ					0	5	0	0	0	35										
ɜ											25	5	0	0	0	10	0			
ɔɪ											5	65	25	0	0	10	45			
u											45	5	55	85	15	45	30			
ju											10	5	0	10	10	0	0			
ə											0	0	0	0	25	5	0			
ʊ											0	10	5	0	15	25	0			
ɔ											10	0	10	0	5	5	20			
ɒ																		30	0	0
oʊ																		25	85	10
aʊ																		5	5	50

Vowel Confusion Matrix

after Plant (1977)

and diphthongs. He produced confusion matrixes for the consonants and for the vowels and showed some definite clustering. His study grouped the consonants into 7 visually different groups, and grouped the vowels and diphthongs into 4 visually different groups. While these are not 100% categorically distinct, they provide a starting point for the database. Testing of the database can further refine the grouping if required. See the vowel and consonant confusion matrices.

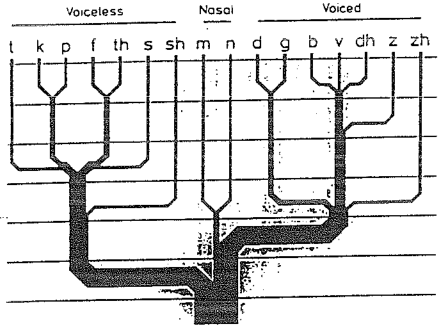
The 4 vowel groups are labelled A,B,C,D for the database, and the 7 consonant groups are called 1,2,3,4,5,6,7 with silence being represented as a # symbol. This gives 11 groups. These groupings are largely supported by Summerfield (1987). See the auditory and visual confusion trees.

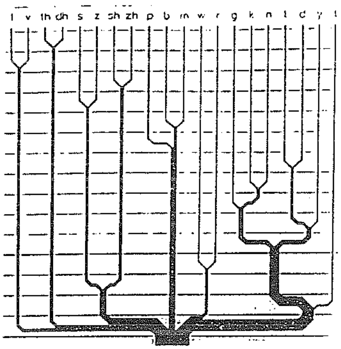
The A vowel group consists of high and mid-high front vowels, small aperture, spread lips and contains /i/,/ɪ/,/ε /,and the diphthong /ɪə/. The B group consists of mid-low front and low central vowels, large aperture, neutral lips and contains /a/,/æ/,/ʌ/, and the diphthongs /aɪ/,/εə/,/eɪ/,/æʊ/. The C vowel group consists of mid-low back vowels, large aperture, rounded lips and contains / / and the diphthong /oʊ/. The D vowel group consists of mid-high, back-high central and central vowels, small aperture, rounded lips and contains /ɜ/,/ə/,/u/,/ʊ/,/ɔ/ and the diphthongs /ɔɪ/,/ʊə/.

The consonant group 1 consists of bilabials and contains /p/,/b/,/m/. The consonant group 2 consists of labio dentals and contains /f/,/v/. The consonant group 3 consists of interdentals and contains /θ/,/ð/. The consonant group 4 consists of labio velar glides and contains /w/,/r/. The consonant group 5 consists of palatals and contains /dz/,/ʃ/,/tʃ/. The consonant group 6 consists of alveolar nonfricatives and plosives and velar plosives and contains /l/,/n/,/j/,/g/,/k/. The consonant group 7 consists of alveolar fricatives and plosives and contains /s/,/d/,/t/.

The database is constructed with the transitions from group to group, that is group pairs (similar to diphones), rather than as 11 fixed targets. This gives 143 transition pairs when arranged in exhaustive permutations (including silence).

These group pairs were arranged into clusters of cvc vcv vvv and ccc and produced 93 nonsense words, with some duplication. They were filmed and digitised after Robinson (1992) to generate the image sequences. Some of the words were difficult to say because they contained pairs that did not exist in Australian English. These were included for completeness but are of dubious value. Each word had a duration in excess of 1 second which at a filming speed of 100 frames per second and a digitising resolution of 512x512 8 bit pixels, produced over 2.3 gigabytes of data. A pixel is a picture-element and represents the horizontal and vertical sampling resolution. The number of bits allocated to each pixel determines the z resolution. In this case it is 8 bits, which provides each pixel with one of 256 possible grey colours. The lip area was selected in a 128x128 pixel box and was extracted as a sequence of subimages. This represents 1/16th of the screen area and





Auditory Confusion Tree

Visual Confusion Tree

after Summerfield (1987)

consequently a similar reduction in data, to just over 143 megabytes. The elimination of the duplications will reduce this further. A final figure has not been arrived at yet.

The film was examined for a typical front view of a neutral face and this was saved as a master frame. The lip area was saved as a string of subimages from the nonsense words, containing the diphones. These were entered into the database as strings of video subimages. The subimages were examined further to find the places of similarity (between all sequences), near the centres of the targets. This information was arranged in a set of tables which effectively contained the points that can be used to merge the visible diphone sequences for lip movement synthesis. These are the transition tables.

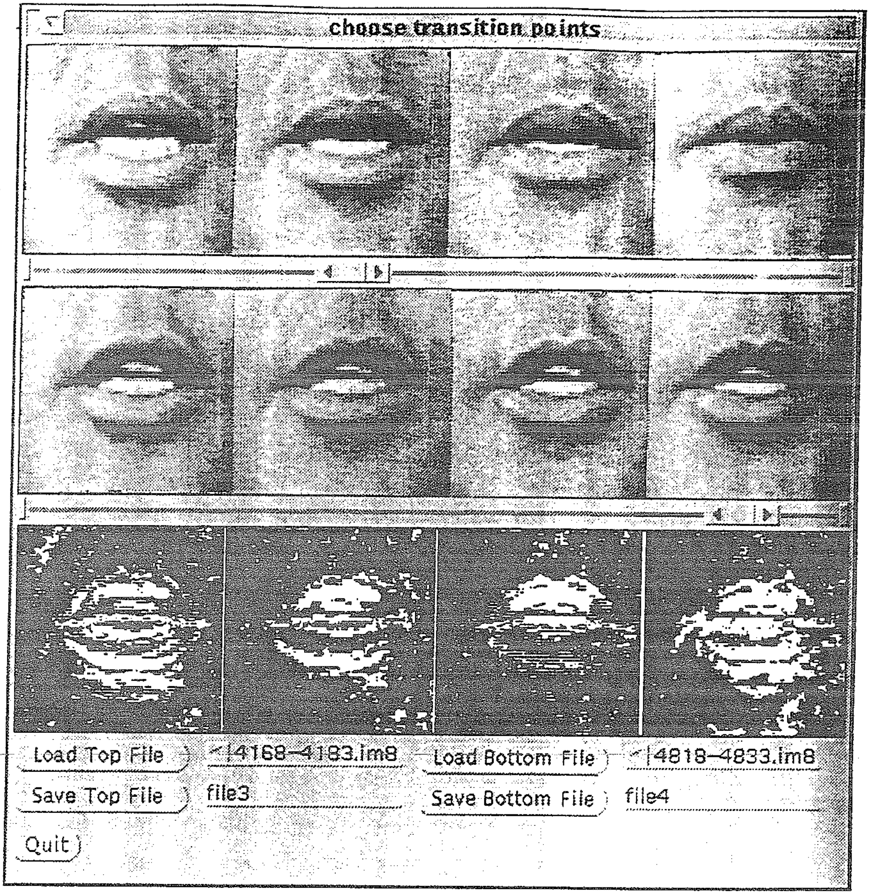
COMPARING IMAGE SEQUENCES

A program was written to run on a Sun Workstation under Openwindows to compare sequences and determine the points of similarity. The program displayed 2 image sequences, each in it's own window, aligned one above the other. The sequences are shown as a time series of lip positions, from left to right increasing in time. The 2 sequences are much longer than their windows, and only 4 frames out of each of them are shown at once. They can be moved left and right by placing the mouse cursor on the scroll bar and pulling it left or right. A third window, below these two, contained a comparison of the images, frame by frame, updated each time the other images are scrolled.

Since each frame consisted of a matrix of size 128x128 and each point in this matrix contained a pixel, the values could be compared. The pixel values could be any value between black and white (all the various grey colours) which corresponded to continuous values between 0 and 255.

All digitisation processes contain resolution (sampling) errors of up to 1 pixel in both x and y directions, and the differences in the cine film media and video tape media, would provide similar variations in the z direction from frame to frame. The registration of the cine film, as it moved through the camera at high speed was imperfect and manifested itself as elliptical wobble when played back. This was corrected by hand by examining each frame, and moving it pixel by pixel, until the best registration was achieved. All these errors were usually cumulative, and added visual noise to the compared image. However they were largely subliminal during playback, the registration being the most obvious.

Since the images were displayed as a grey scale, comparing the images gave many differences of 1 or 2 between corresponding pixels, and displaying them as their difference value would be hard to see. So an arbitrary threshold of 20 was settled upon and any values above this were drawn as white, any below were drawn as black. This gave a good idea of the comparison. It showed of



Load Top File 4168–4183.im8

Load Bottom File 4818–4833.im8

Save Top File file3

Save Bottom File file4

Comparing Frame Sequences to Determine Transition Points

images would show only this artefact of the recording and digitisation process. Since frame by frame adjustment had also attempted to reduce global head movements (despite the head rest) slight aspect changes showed as lines along any edges or rapid grey changes, such as at the lip boundaries or nose and nostril curves. The comparison window showed the differences, and highlighted them. The difficulty lay in finding a close match between frames of different sequences. Each diphone had to be compared to every other diphone sequence to find a perfect match, a close match, a near match, an approximate match, or any similarity at all!

The Figure shows the comparison between the "pee" diphone from "peep" and the "een" diphone from "neen". This would allow the two to be assembled sequentially to form a "...peen..." or "...neep..." sequence. The differences are shown below the 2 sequences. The top sequence shows the end of the vowel and the beginning of the closure for the final /p/ in "peep". The other sequence shows the vowel near it's centre of the word "neen". The view is arranged at the closest match.

In the top sequence, the top teeth are being covered as the mouth closes, to pucker for the final /p/ burst, with more bottom teeth being shown than the bottom sequence. In the lower sequence, the burst, with more bottom teeth being shown than the bottom sequence. In the lower sequence, the top teeth are visible most of the time, with not as much bottom teeth area visible. The lips are more dynamic in the top sequence, not only opening and closing, but with some protrusion. There is very little movement in the lower sequence. Cupids bow is more visible in the top sequence, but more arched in the lower sequence. The difference sequence shows much more white area for the left and right frames, and more top lip and teeth differences, for frame 3, but probably the best match is frame 2, the least white area for top lip, bottom lip, and teeth changes. I would therefore select a transition point at frame two. So moving from left to right for "...peen..." would consist of all the top sequence (off screen) including the left frame shown, and then the bottom sequence from the second frame shown. Similarly for the "...neep..." sequence.

TRANSITION TABLES

If the image sequences are to be assembled in some order, then a set of splicing rules needs to exist. The points at which the sequences can be joined vary within the sequence. The diphones are not a fixed length. They are a maximum length with several transition points within that sequence. Changes from one diphone to another occur sometimes near the beginning for one particular diphone, and sometimes near the end for another. The set of diphones were compared to each other to determine points of similarity. The points at which their similarity was the greatest was noted and placed in a lookup table. When it is time to assemble the sequences, this lookup table of transition points provides the frame numbers.

There are 4 tables for the vowels, and 7 tables for the consonants. There is no table for # (silence) but there are entries in all the tables for silence/vowel and silence/consonant transitions. The table for the A vowel consists of an 11x11 matrix with 2 members in each cell. The table for the 1 consonant consists of an 11x11 matrix with 2 members in each cell. All tables were constructed like the table shown and the cells were filled with transition points. The vertical axis is all the combinations of consonants and vowels for the A vowel (for discussion purposes). This is the entry diphone. The horizontal axis is the target diphone consisting of all combinations of that vowel and the consonants. There may be some duplication of data, but this done to make the tables general and easy to read. Otherwise the reading program would have to read the table backwards for some entries. Each table is arranged for an entry diphone to elicit a transition pair for an target diphone. Considering the same example as above, the transition points will be held in A transition table. The entry diphone (for "...neep..") is 6A and the target diphone is A1 to assemble "...6A1..." and the pair derived from the cell 6A,A1 will provide the transition point. The same data is in cell 1A,A6 since the transition point is the same (usually, but not always).

	A1	A2	A3	A4	A5	A6	A7	AB	AC	AD	A#
1A						p,q					
2A											
3A											
4A											
5A											
6A	n,m										
7A											
BA											
CA											
DA											
#A											

Continuing with the example of the A Transition Table we can find that we must use the first "n" frames of the 6A diphone and then use the remaining frames of the A1 diphone starting "m" frames from its beginning. Similarly, for the "...1A6..." sequence we can get from Transition Table A that we need to use the first "p" frames of the 1A diphone and then use the rest of the A6 diphone from "q" frames from the its beginning.

CONCLUSION

The database is not finished and will probably change as the methods of data extraction are tested in real time and refined. The system is workable but large, too large for micro computer applications as yet. It still requires the power of large graphic workstations to realise any practical results. This is, however, a good base to work from.

ACKNOWLEDGMENTS

My thanks go to Robert Mannell, Jonathon Harrington, Chris Callaghan, John Telec, Andrew McVeigh, Steve Cassidy, Jim Powe, the Macquarie University Workshop, the Macquarie Audio Visual Technical Services, and the ABC for their help and support during this project.

REFERENCES

Benoit, C., Lallouche, T., Mohamadi, T. & Abry, C. (1992) *A Set of French Visemes for Visual Speech Synthesis,* Talking Machines: Theories, Models, and Designs, Baily, G., Benoit. C., & Sawallis, T.R. (Editors), Elsevier Science Publishers B.V. Pages 485-504

Plant, G. & Macrae, J. (1977) *Visual Perception of Australian Consonants, Vowels and Diphthongs,* Australian Teacher of the Deaf. Vol 18, Pages 46-50

Plant, G.L. (1980) *Visual Identification of Australian Vowels and Diphthongs,* Australian Journal of Audiology 1980 2:2 Pages 83-91

Summerfield, Q. (1987) *Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception,* Hearing by Eye: The Psychology of Lipreading, Dodd, B. & Campbell, R., Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Storey, D. & Roberts, M. (1988) *Reading the Speech of Digital Lips: Motives and Methods for Audio-Visual Speech Synthesis,* Visible Language Vol XX11 Number 1, Pages 113-127

Robinson, R.E.E. (1992) *Synthesising Facial Movement: Data Capture,* Proceedings of the Fourth Australian International Conference on SPEECH SCIENCE AND TECHNOLOGY Brisbane, December 1992 Pages 207-212