SYNTHESISING FACIAL MOVEMENT: REAL TIME VISUAL SPEECH

R.E.E.ROBINSON

Speech Hearing and Language Research Centre
School of English Linguistics and Media
Macquarie University

ABSTRACT - This article describes the method by which a phonetic string is processed into visemes, with adaptability to other languages. The visemes are paired into disemes and lip sequences are extracted from a lip database. The resultant lip sequences are joined by using transition tables and placed in a play buffer. The number of frames in the play buffer is adjusted to match the duration of the required sequence. The system timer is used to play back the sequence in real time on a Sun Ultra 1/170 workstation.

INTRODUCTION

Acoustic speech synthesis has been in use for many years and the search for realistic sounding synthesis continues. Visual speech synthesisers are relatively new, but they have the advantage of the existing knowledge base of how the acoustic synthesis has developed and to some extent, the lessons learned. The acoustic synthesiser has developed to the point where Text To Speech (TTS) front ends are being tuned to add realism to the synthesis.

The same TTS front end can be used to control a Visual Speech Synthesiser (VSS). The intermediate level where the input text has been converted to phonetic information applies equally well to visual synthesis. Whereas the acoustic synthesiser interprets the phonetic information as sound waves, the visual synthesiser interprets this information as lip shapes and movements. This is usually done by mapping the phonetic information to a set of lip shapes which are arranged in groups called visemes. Several studies have examined lip shapes for various languages and have produced viseme lists with phoneme or phonetic equivalents.

The search for realism, both in natural appearance, and natural movement applies to visual synthesis as it does to acoustic synthesis. Without realism the synthesis is dull and monotonous. There are many techniques for giving the acoustic synthesis a more natural sound, from the micro level with pitch perturbation, the intermediate level with prosody, and the macro level with rhythm. The VSS suffers the same problems. Some visual synthesisers completely synthesise the head, which allow great flexibility in movement and expression, but are difficult to make natural looking. This VSS uses actual images of a real person, which is the closest thing to a natural look. The problem then is how to make it move in a natural way when attempting to produce speech. Another technique used by Duffy is to combine both methods. This approach places high demands on the computer. Expressions and emotions which add interest to the face and carry extra meaning have been explored by Henton (1994).

The challenge to produce naturalness and interest while moving the lips has only become possible with advances in computers, both in processing speed and graphics displays. The search for increased resolution incurs a penalty in execution time and data size, which necessitates some trade off. At the beginning of this project, the computers in use could display the realistic looking face but not in real time, the synthesis appearing as slow motion. The current computers run the VSS in real time.

PHONEME ENTRY

The input to the VSS can take several forms. It can be a file from a TTS containing phonetic and timing information in the Macquarie University TTS Format. The file is analysed and phonemic information is extracted to form a string, which is displayed in a text field (as a diagnostic). The timing information is also extracted. The VSS will also allow direct phonemic input from the keyboard and expects characters and symbols from the Machine Readable Phonetic Alphabet (MRPA). It uses default timing when using direct keyboard entry.

The phonemic string is applied character by character to a lookup table (LUT) and the corresponding value is placed in a viseme string. Refer to the Phoneme to Viseme Transformations. Any characters not in the LUT are flagged as errors, and silence is substituted. Since the viseme string was constructed from the LUT, there should be no invalid symbols, but a check is made anyway, and any found are also replaced by silence symbols. The viseme string is examined further to ensure it has a silence symbol at the beginning and end. These are added if not present. The viseme string is then checked for repeated visemes, and any duplicates are removed. At this point the viseme string is displayed (as a diagnostic), and processing continues.

OTHER LANGUAGES

The use of a LUT to translate phonetic or phonemic input to visemes, is readily adaptable to other languages. There are four conditions that need to be considered.

The first condition is that a LUT can be easily reloaded, reselected, or changed. This is merely a matter of loading a default LUT and providing a method of loading others. These can reside on disc and can be selected from a LANGUAGE button and read into the VSS. The second condition is that a LUT has been constructed for another language. This will depend on the existence of a study that categorises lip positions for that language, into visemes or some equivalent group. There are many studies of this nature. The third condition is that those visemes have a corresponding viseme in the existing database. If there is no matching viseme, then a similar one will need to be chosen, or a new database loaded. The fourth condition is that there needs to be some method of producing a suitable input to select the viseme and control the VSS. This could be input from the keyboard into the phoneme field, or preferably a TTS for that language, which produces compatible input.

The default LUT is based on the Plant (1980) and Plant & Macrae (1977) studies of Australian English. This study has identified 11 visemes. A British English LUT could be based on the studies by Storey & Roberts (1988) and Summerfield (1985) with 15 visemes. An American English LUT could be based on the studies by Montgomery & Jackson (1983) and Owens & Blazek (1985) with 11 visemes. A French LUT could be based on the study by Benoit et al (1992) with 19 visemes. A Dutch LUT could be based on the study by van Son et al (1993) with 10 visemes. A Japanese LUT could be based on the study by Fukuda &Hiki (1982) with 13 visemes.

VISEME ENTRY

The diagnostic field will normally show the result of the analysis of the phonemic input or TTS file. For testing purposes, the visemes can be entered from the keyboard and the Playv button pressed. The viseme string is checked in a similar manner as the processed phonemic string, for start and stop silence symbols, invalid symbols and duplicates. The corrected string is displayed and processing continues.

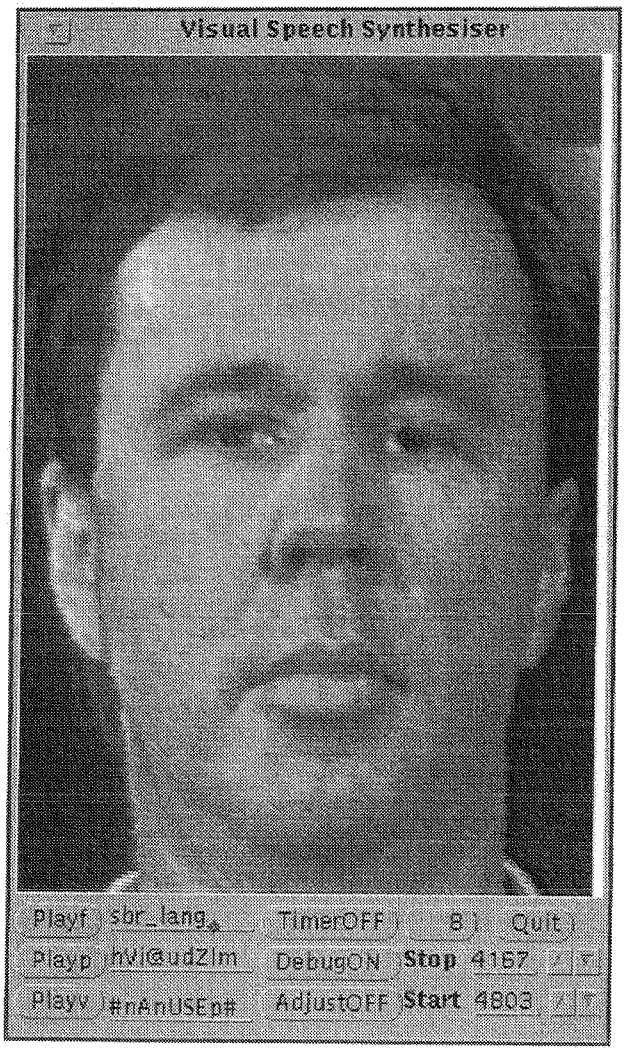
The database was constructed using the visemes determined by Plant (1980) and Plant & Macrae (1977). Refer to the Phoneme to Viseme Transformations. In the Robinson (1994) paper, the visemes were given arbitrary names. The 4 vowel visemes were called A, B, C, D, and the 7 consonant visemes were called 1, 2, 3, 4, 5, 6, 7. The names have been changed to a MRPA symbol that represents the group. They are now called /E/, /A/, /U/, /O/, and /p/, /v/, /T/, /w/, /S/, /n/, /d/, respectively. The new names are used in the tables.

ANALYSIS

The visemes are now analysed in a sequential manner, with particular attention focused on the viseme and their adjacent partners. This is because the VSS assembles a series of images, which show the lip change from one viseme to another, rather than showing target visemes in an isolated manner. This method shows natural transitions between visemes and will show any coarticulation that may occur. This is analogous to auditory speech synthesis using two phonemes or a diphone. Since viseme pairs are used in this visual speech synthesiser, these two viseme units can be called a diseme.

PHONEME TO VISEME TRANSFORMATIONS

/i/ /ɪ/ /ɛ/ /ɪə/	=	/E/
/a/ /æ/ /ʌ/ /aɪ/ /ɛə/ /eɪ/ /aʊ/	=	/A/
/ɒ/ /oʊ/	=	/U/
/ɜ/ /ə/ /ʉ/ /ʊ/ /ɔ/ /ɔɪ/ /ʊə/	=	/O/
/p/ /b/ /m/	=	/p/
/f/ /v/	=	/v/
/θ/ /ð/	=	/T/
/w/ /r/	=	/w/
/ʤ/ /ʃ/ /ʧ/	=	/S/
/l/ /n/ /j/ /g/ /k/ /h/	=	/n/
/s/ /d/ /t/	=	/d/
silence	=	/#/



As an example, consider the following phonetic string and the subsequent analysis. The top line is the output of the Macquarie University TTS in the old format. The next line is the International Phonetic Alphabet (IPA) equivalent. The Third line is the MRPA representation of that line. The fourth line is the viseme translation of the sequence. The fifth line is the same string after diseme analysis.

old TTS		/H/	/15/	/L/	/'5/	/DZ/	/"2/	/M/	silence
IPA		/h/	/ʌ/	/l/	/oʊ/	/dʒ/	/ɪ/	/m/	
MRPA		/h/	/V/	/l/	/@u/	/dZ/	/I/	/m/	
viseme	/#/	/n/	/A/	/n/	/U/	/S/	/E/	/p/	/#/
diseme	/#n/	/nA/	/An/	/nU/	/US/	/SE/	/Ep/	/p#/	

The analysis begins at the first symbol in the viseme string and this will be the silence symbol. This corresponds to closed resting lips. It is the /#/ symbol. The second symbol is examined and this viseme (in this case /n/) is joined to the preceding symbol to form a pair and becomes the first diseme which is /#n/. The third viseme (/A/) is examined and with the preceding viseme (/n/), becomes the second diseme (in this case /nA/). The entire viseme string is processed in this way to make diseme units, and this is shown as the fifth line in this example. The database consists of disemes that can be joined to provide visual speech. The place that the disemes are joined can be extracted form the transition tables.

COARTICULATION

Visemes do not exist in isolation, even though some studies may imply that this is so. Since coarticulation occurs, each viseme can be altered by its preceding or following neighbour., sometimes in a negligible way, and in other cases in a noticeable way. This is why Benoit et al (1992) has identified 19 visemes in his study of the French language. He has taken several vowels and consonants and arranged them in many ways. He has then categorised them into viseme groups and has found more than one viseme for a some vowels and consonants, because the effect of its neighbour has modified it to such an extent as to make it visually different. Some other studies have used a single carrier phrase or word and so have identified a smaller number of visemes which are more stereotyped.

The Plant (1980) and Plant and Macrae (1977) studies identify 4 vowel visemes and 7 consonant visemes for Australian English. Robinson (1992) arranged these 11 visemes in exhaustive combinations and recorded them on film. They were then digitised after Robinson (1992) and placed in an image database. This method captured any coarticulations that occurred. The playback method reproduces the coarticulations over two visemes. The /p/ viseme (which includes the /p/ /b/ and /m/ consonants) in Australian English exists in the database in several variants. It can be extracted as a diseme with any of four vowel visemes, seven consonant visemes, and silence (either preceding or following) . This should cover most coarticulation over two visemes and has redundancies.

ACCESSING THE TRANSITION TABLES

Continuing with the example, the /#n/ diseme has to be joined to the /nA/ diseme. The /n/ transition table will be used to provide the first join. The entry diseme is /#n/ and the target diseme is /nA/. The transition table cell at this address contains two numbers. See Robinson (1994). The numbers represent the closest match of two lip image sequences that contain the appropriate lip shapes. The first is the frame number of the end point of the previous diseme and the second number is the frame number of the start point of the second diseme. Now the /#n/ diseme can be joined to the /nA/ diseme. The analysis moves on to the next diseme. The /A/ transition table is now used to determine the transition point of the /nA/ and /An/ disemes. The analysis continues until the final /#/ is reached. The lip sequences are now known and can be assembled for display.

TIME CONSIDERATIONS

The lip sequences can be played back as they are now, or they can be further modified. If the source for the phonetic input was a TTS file, then timing information will be available and the length may need to be changed. Otherwise they are played as they are, which is the default timing, and this corresponds to the speed at which they were recorded. The lip sequences were recorded at a speed of 100 frames per second, which is one frame every 10 milliseconds (ms). The Macquarie University TTS uses 10ms

as its timing unit. The TTS file provides the durations in these units. The disemes need to be adjusted to match the phonetic input string. If the diseme is too long, then some frames are removed. If the diseme is too short, then some frames are duplicated. Care is taken to do the editing away from the ends as these are the matching points with other disemes. Only the middle 80% of a diseme is altered. The editing is evenly distributed through the diseme.

It was envisaged that the refresh rate of the graphics device and monitor would be a significant factor affecting realtime playback. Some graphics devices show a wipe as images are updated and others show a flicker. Double buffered graphics devices refresh the hidden screen, and show it as the next refresh. The play buffer was designed to be adjustable so that, a new lip image could be added at each refresh, including frame rates faster than 100hz. As it turned out, there has been no visible refresh artefact observed, on several different Sun platforms. Even the older slower Suns displayed the lip sequences faithfully, albeit at a slower speed. Since there has been no problem so far, this consideration has been postponed. It will be implemented in a similar manner as the timing is adjusted for TTS durations, if required.

EDGE BLENDING

The lip database is described in Robinson (1994) and consists of a master image, several tables and many lip image sequences.

The master image is a font view of the head and shoulders of a male speaker, digitised after Robinson (1992) from a real image. The image is portrait oriented with a 320 pixel (picture element) wide and 460 pixel high resolution. Each pixel can have one of 256 grey levels. The image uses about one quarter of the screen under the Openlook windowing system. This is displayed using XView on the Solaris 2.5 operating system.

The lip area of the image is 160 pixels wide and 120 pixels high. Several sequences of lip images have been stored in the database, with indexed entry. The lip images are superimposed (or pasted) onto the master image. The join between the lip image and master image is visible unless some processing is done. All the edges of the lip images have been blended with the master image similar to the Duffy method. For a 10 pixel wide border around the edge of the lip image, each pixel is blended with the master image. A pixel from both, are added together and divided by the distance from the edge, which gives a variable weighting. The closest pixels to the edge are weighted towards the master image and those closest to the lips have a weighting favouring the lip image. There is a linear gradient across the border area. The lip images then blend with the master image and the joins are not visible.

PLAYBACK

When the lip sequences have been determined, they are ready to be played back. The images are copied from the database one by one, in order, to the lip area of the master image. A timer is started to control the playback. The timer is provided by the system clock and the period is set to 10ms. Every period, it interrupts and a new lip image is transferred to the screen. Normally the timer will be stopped when the sequence end has been reached, and the visual synthesis occurs just once. There is a timer button that allows free running of the timer and thus a continual repeat of the sequence. The VSS is now ready for more input.

The joins between the disemes can be adjusted after assembly, for testing reasons. The STOP and START fields are updated as the disemes are loaded giving some indication of the loading from the transition tables. When the ADJUST button is engaged, the transition points can be changed and the play buffer reloaded with new lip images. This allows the transition points to be varied to check that they are at the best or closest match. If a better match is found, the transition points can be noted and the transition tables edited to change the points.

The original workstation used for this project was a Sun 4/330 which played back the speech in slow motion. The subsequent workstation to this was a Sun Sparc 2 which played back at a slightly faster rate. The Sun Sparc 5 workstation was noticeably faster, but still less than real time. The Sun Ultra 1/170 succeeds in realtime playback.

CONCLUSION

Real time Visual Speech synthesis has been achieved only with a very fast workstation. The data required and screen update speed made the task quite difficult and only the current technology has achieved success. The database is quite large, as it contains many disemes, to allow for some coarticulation across visemes. Provision has been made for other languages in the form of readily changeable phoneme to viseme conversion tables. Work is planned to continue to add more features and refine the existing structure.

ACKNOWLEDGMENTS

My thanks go to Robert Mannell, Jonathan Harrington, Chris Callaghan, Steve Cassidy, David Mitchell, and all those at the Speech Hearing and Language Research Centre at Macquarie University for help and support during this project.

REFERENCES

Benoit C. Lallouache 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 (Editors) Elsevier Sciences Publishers 485-503

Duffy N.D. *Animation using Image Samples*

Fukuda Y. & Hiki S. (1982) *Characteristics of the Mouth Shape in the Production of Japanese - Stroboscopic Observation,* Journal of the Acoustical Society of Japan (E) 3,2 75-91

Henton C. (1994) *Techniques for Synthesising Visible, Emotional Speech,* Proceedings of the Fifth International Conference on Speech Science and Technology 581-586

Montgomery A.A. & Jackson P.L. (1983) *Physical Characteristics of the Lips Underlying Vowel Lipreading Performance,* Journal of the Acoustical Society of America Vol 73 No 6 2134-2144

Owens E. & Blazek B. (1985) *Visemes Observed by Hearing Impaired and Normal Hearing Adult Viewers,* Journal of Speech Hearing and Research Vol 28 381-393

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

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

Robinson R.E.E. (1992) *Synthesising Facial Movement: Data Capture,* Proceedings of the Fourth International Conference on Speech Science and Technology 207-212

Robinson R.E.E. (1994) *Synthesising Facial Movement: Data Base Design,* Proceedings of the Fifth International Conference on Speech Science and Technology 840-845

van Son N. Huiskamp T.M.I. Bosman A.J. & Smoorenburg G.F. (1993) *Viseme Classifications of Dutch Consonants and Vowels* Journal of the Acoustical Society of America 1341-1355

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

Summerfield Q. (1985) *Preliminaries to a Comprehensive Account of Audio Visual Speech Perception,* Hearing by Eye, Dodd B. & Campbell R. (Eds), Lawrence Erlbaum Associates, Hillsdale, New Jersey