VIDEO SPEECH SYNTHESIS: A FLEXIBLE, INTEGRATED SYSTEM FOR ANALYSING AND SYNTHESISING SPEECH PRODUCTION IN THE VISUAL DOMAIN

N MICHAEL BROOKE
DEPARTMENT OF COMPUTER STUDIES, UNIVERSITY OF LANCASTER

Earlier papers (1,2) have suggested the potential of a reproducible, controllable and well-specified visual stimulus for analytical investigations of speech perception by normal hearing subjects in the bimodal, audio-visual domain. The output of such a 'video speech synthesiser' would be an animated real-time graphical display of the essential facial topography during speech utterance production. This would be synchronised with the output of an audio speech synthesiser to generate the audio-visual stimulus. The kinds of experiments which may be undertaken depend upon the sophistication of the video synthesiser. For example, discrimination experiments to explore the nature of categorical perception in certain bimodal phonetic spaces such as the /ma,ba,pa/ continuum (3) may be carried out using a simple terminal-analogue version of the video speech synthesiser. More complex experiments using purely visual and coherent or conflicting audio-visual stimuli (4) require a more detailed control of the visual articulatory movements and timings. A sufficiently well-specified video synthesiser might ultimately be used to help rehabilitate the hearing-impaired by providing a device in which small but significant variations could be controlled and exaggerated during a period of perceptual training.

The use of a computer offers the most flexible approach to the implementation of a video speech synthesiser. Implementation of a simple terminal-analogue model of facial articulation has been described previously (1), in which the positions of a subset of the visible articulators are defined and updated at small time intervals under the control of a digital computer. The terminal-analogue model is now the lowest level in a hierarchy at whose higher levels it is intended to include algorithms for generating articulatory trajectories by rule from phonetic transcriptions. Currently, quantitative data describing the facial movements during speech articulation are required. They will serve three purposes: first, to supply time-varying positional data for driving the terminal-analogue model and for generating stimuli for perceptual experiments designed to refine our understanding of the cues for optical speech perception; second, to derive parameters of a functional representation of facial kinematics (or dynamics), such as has been applied to formant movements (5), so that synthesis-by-rule may be attempted; and third, to evaluate the decisions taken in the facial modelling process.

Facial movement data

The measurement of facial articulatory movements during speech production poses two major problems, which are i) the definition of a measurement frame, and ii) the separation of positional variations due to the articulatory gestures from those due to the global head movements which are inherent in natural speech production. Some investigators have taken measurements directly from high-speed cine films of speakers' faces which have been assumed (6) or arranged (7) to be in an essentially fixed position. In an initial approach, frames from a reel-to-reel videotape recording of the front face of a speaker enunciating a series of VCV utterances have also been measured, using a video-

cursor device linked to a microprocessor which logged the screen co-ordinate values of a series of defined facial points. A limited correction of measurements to a 'standard' head position was attempted, assuming that i) the global head movements were small, and ii) the variation in distance of the facial points from the camera was negligible compared with the mean separation. Body movements other than facial movements are excluded at present, not necessarily because they are linguistically uninformative, but because a thorough analysis requires movements at this level to be dissociated from those at other levels.

An improved experimental system has been developed. The speaker is videotaped in both front and side views simultaneously, using a plane mirror set at 45° to the principal axis of the camera optics. A Sony rotary-shuttered camera is used. It achieves short exposure times, thus reducing blurring, particularly during the fast consonantal articulations. The recordings are replayed through a video disc storage device (Sony SVM1110) which permits single frames to be displayed for measurement with the videocursor. A data-logging programme running on a microprocessor system supervises the accumulation of files of screen co-ordinate data for a set of specified reference, scaling and articulatory points. Four reference points are chosen to mark fixed, non-articulatory locations on the face. The scaling points are fixed points in the picture area (but not on the face). From their measured screen co-ordinates a linear scale can be independently established for every frame, so overcoming potential errors arising from changes in the calibration of the equipment. The articulatory points are the facial positions whose time-varying displacements are to be measured. The files of raw screen co-ordinate data are transferred over a link to a minicomputer for analysis, which proceeds in two stages:

i) One of the four reference points is selected to mark the origin of an axis-frame emanating within the head itself. The raw screen co-ordinates from each frame are corrected for the effects of projections in the measurement system and computed as true (x,y and z) distances from the origin of the head axes, but measured in the axis-frame defined by the television screen orientation.

ii) A 'standard' head orientation is fixed by the co-ordinate positions of the four reference points in a single 'master' frame. In this frame, the head and screen axis systems are assumed to coincide. Standardisation of head orientation for every other frame is accomplished by multiplying the (x,y,z) vector of co-ordinates of each point by a rotational transformation matrix of order 3x3. The matrix is derived from the measured co-ordinates of the three remaining non-articulatory reference points and establishes a co-ordinate vector (x*,y*,z*) for each point which is measured in the head's axis-frame.

The output of the analysis programme is a file of articulatory trajectories representing the positions of each of a set of points on the face at specified moments during the course of an utterance. The effects of global head movements are eliminated without constraints having been placed upon either the position of the head or the physical emplacement of the videorecording set-up. Statistics regarding experimental accuracy and reproducibility can also be output if replicated measurements of any frame are provided.

Facial graphics display

An animated graphics display of a minimal, fixed-topography facial diagram has already been implemented on a PDP-11/60 host computer with a GT40 graphics satellite. The four lip margins and jaw could be moved independently. The nose and eyes were indicated as static outlines. The articulatory movements were

controlled in a terminal-analogue mode, using streams of positional data derived from the facial measurement experiments. Fixed measurement intervals of 1/25 sec. were used, from which a quasi-cinematographic display was generated (1,2). The package formed a useful first-approximation video synthesiser potentially suitable for simple visual discrimination experiments, but, like many other current animated displays (6,8), was not inherently adaptable.

A flexible graphics package to simulate facial speech articulations should permit variability in i) the kind of facial diagram to be drawn, ii) the utterance to be simulated, and iii) the graphics hardware to be used. A new, modular system is being implemented which distributes the handling of these logically distinct sets of characteristics between four separate computational phases, using stored data files as the common interfacing medium:
Phase 1: A file of tables (the facial diagram specification file) is generated which describes the nature of the facial diagram to be drawn. A user-defined file is supplied. Its first section gives the ordered sequence of line and arc segments to be displayed, together with indices referencing tables of key co-ordinates whose contents subsequently define the actual screen positions of the segments. The second section specifies the kind of articulatory movement which each of the key co-ordinate points is permitted. Movement may be i) independent, that is, determined by the articulatory trajectory data for an utterance, or ii) dependent upon the linear displacement of one of the independently moveable points.
Phase 2: A user-defined file containing the 'resting' co-ordinate positions of the key facial points is combined with the data computed in phase 1. The result is a definition of the initial facial shape to be constructed. Having dealt with the facial characteristics, the utterance characteristics are incorporated into the model via a file of articulatory trajectories. These are used to relocate the key facial points and compute the geometrical parameters of each of the moveable pictorial elements for each of the frames in a quasi-cinematographic animated display sequence. The results are output to file.
Phase 3: The file computed in phase 2 is used to generate and output the normalized screen co-ordinates of a set of linear vectors for each frame of the display.
Phase 4: A user-defined specification of the graphics device characteristics, such as screen size and maximum vector length, is supplied. Each set of vectors formed in phase 3 is then recomputed in true screen co-ordinates and presented for its defined time period, thus generating the animated display.

Results and discussion
Experimental measurements of lip and jaw movements during the production of a selection of VCV utterances are now being accumulated. Simple visual discrimination experiments are also being prepared using existing data for a set of syllables such as /aba/, /abi/ and /abu/ to drive the original terminal-analogue graphics display.

Phase 1 of the new graphics package has been implemented and phase 2 is under development. Pictorial elements are currently limited to line and circular arc segments. However, the design of the package readily permits the possibility of using elliptical arc segments at a later stage (6). In addition, the removal of hidden lines such as would result, for example, from the occlusion of the teeth by the lips is not yet possible but should also be easy to incorporate in the future.
The division of the comuptations into phases, the form and the use of intermediate data files all confer important advantages on the new package:

i) Changes may be made to the facial diagram, utterance and graphics device characteristics independently, whilst retaining existing data files unchanged for future use. Reprocessing is reduced to the recomputation only of those phases of the task whose input files are altered. In particular, graphics devices are assumed to draw only linear vectors and device dependence can thus be restricted to the final phase of the package. Full error checking and diagnosis are provided for all user-defined file inputs, eliminating the risk of continuing computations with bad data.

ii) The articulatory trajectory files now contain the timing data for each frame of the animated display. It is therefore possible to vary the frame display rate with articulatory rates of change, consequently reducing the computational overheads.

iii) User-defined files are specified by format only. Their source is undefined. Consequently they may be set up by any independent computational process. For example, articulatory trajectory files are presently derived from the analysis of the facial measurement experiments,using a link between the data-logging and data analysis systems to integrate the facial measurement and display synthesis procedures. The trajectory data might also, however, be obtained from an independent articulatory synthesis-by-rule computation (2,5) or by concatenating user-defined trajectories. In these cases, both spatial and temporal characteristics of the trajectories would be very easy to modify and a wide range of test stimuli could rapidly be generated.

Acknowledgements

The author wishes to thank Prof. M. P. Haggard, Director of the MRC Institute of Hearing Research, Nottingham, for providing the facilities to carry out this work; Dr. A. Q. Summerfield for his advice and assistance; and the MRC for the award of a research grant to support the project.

References

1. N. M. BROOKE 1979 Proc. Autumn Conf. of Institute of Acoustics, 41-44. Development of a video speech synthesiser.
2. N. M. BROOKE 1980 Proc. International Seminar on Labiality (Lannion, France). Towards a video speech synthesiser: a model for computer graphics displays and its relation to experimental measurements.
3. N. P. ERBER and C. L. DE FILIPPO 1978 J. Acoust. Soc. Amer., 64, 1015-1019. Voice/mouth synthesisers and tactile/visual perception of /pa,ba,ma/.
4. H. McGURK and J. MACDONALD 1976 Nature, 264, 746-748. Hearing lips and seeing voices.
5. H. FUJISAKI 1978 Proc. Joint Meeting Acoust. Soc. Amer. and Acoust. Soc. Japan (Hawaii). From discrete functional units to continuous speech characteristics- a functional formulation of articulatory and phonatory dynamics.
6. N. P. ERBER, R. L. SACHS and C. L. DE FILIPPO 1980 in 'Advances in Prosthetic Devices for the Deaf: A Technical Workshop' ed. D. L. McPherson (Nat. Technical Inst. for the Deaf). Optical synthesis of articulatory images for lipreading evaluation and instruction.
7. M. GENTIL, L. J. BOE and R. DESCOUT 1980 Proc. International Seminar on Labiality (Lannion, France). EMG study of the lips (OOINF, MENT, DLI and LLSA) in the production of some French syllables.
8. A. A. MONTGOMERY 1980 Proc. Meeting Acoust. Soc. Amer. (Los Angeles, USA). Development of a model for generating synthetic animated lip shapes.