Proceedings of The Institute of Acoustics 2A13

DEVELOPMENT OF A VIDEO SPEECH SYNTHESISER

N MICHAEL BROOKE

UNIVERSITY OF LANCASTER, CURRENTLY AT MRC INSTITUTE OF HEARING RESEARCH

It can readily be demonstrated that visual cues are important in the everyday perception of speech, especially where there is noise or hearing impairment. The development of a video speech synthesiser, that is, a computer-generated dynamic visual display of a simplified but realistic facial schema which can be synchronised with the output of an audio speech synthesiser, would therefore provide a vital tool for analytical investigations of normal speech perception and also of speech-reading by the hearing-impaired thus offering a potentially valuable technique for rehabilitating them. For example, by varying the time of lip-opening in an optically specified /ba/ relative to the onset of a complex tone, it is possible to produce an audio-visual continuum of syllables ranging from /ma/ through /ba/ to /pa/ (1). It would therefore be possible to use a simple, terminal-analogue version of the video speech synthesiser in basic discrimination experiments to decide whether categorical perception can be a general property of bimodal phonetic events or is limited to a certain range of acoustical values. As a potential tool in rehabilitating the hearing-impaired, it may eventually be possible to isolate and then exaggerate, during an initial training period, the subtle visual differences between normally homophonous contrasts like /ma/, /ba/ and /pa/. This would require more sophisticated control over the timings and motions of the displayed articulatory gestures than that provided by a terminal-analogue video synthesiser and articulatory synthesis-by-rule from phonetic transcriptions of simple utterances may be considered as the means of computing the streams of movement data which operate the terminal-analogue model.

Whilst the internal topography of the vocal tract during speech articulation has been widely studied (2,3), there is only very sparse quantitative data concerning the visible, front-face topography (4); even the qualitative literature treats facial gestures as static, phonemic units (5). In the same way, the sophisticated systems now available for modelling articulatory mechanisms in the vocal tract domain (6) are not paralleled by visual facial models other than rather limited and inflexible real-time hardware (7), or complex and computationally expensive software types (8).

Current theories viewing speech articulation as the co-ordinated activity of a relatively slow vowel-producing system and a faster, anatomically quasi-independent consonant-producing system (2,9) have lead to the postulation of a powerful mathematical model for the simulation of vocal-tract articulatory movements (9). Such a model has already been applied to the simulation of formant movements (10). Its application as a transition algorithm in the synthesis-by-rule of facial articulatory movement is novel in that simulation of coarticulation may be modelled as a coproduction process rather than by the concatenation of a series of context-dependent target configurations. This depends, however, like the implementation of a terminal-analogue video synthesiser, upon the availability of reliable quantitative data on facial movement.

DEVELOPMENT OF A VIDEO SPEECH SYNTHESISER

The development of the synthesiser consequently involved two, related lines of work:

a) Production of a computer graphics software package for displaying a full-face outline with all essential topographical features and animation capability, built from simple modules enabling extra features and modes of animation to be quickly implemented through simple software modifications.
b) Development and execution of a systematic programme to record, measure and analyse continuous facial movements during speech utterances by human speakers.

Both lines were conditioned by the desire to produce a plausibly lifelike video synthesiser using the simplest possible computational model.

Graphics model for facial displays

Since a terminal-analogue model corresponds closely to a frame-by-frame dynamic graphics display and represents the lowest level of a synthesis-by-rule system, this model was chosen for initial development.

The assumed 'minimal useful' facial display consisted of a face outline with a moveable jaw and lip margins, plus static outline indications of the nose and eye positions. Inspection of still photographs of facial postures during vowel and consonant productions lead to the postulation, as a first approximation, of circular arcs for the representation of each lip margin (5). The centre of each lip margin was allowed to move vertically and the corner both vertically and horizontally. This models effectively the gross actions of the lip musculature (11). The face outline was specified as separate cheek and chin arcs. The cheek arcs originated at a fixed point near the temple and terminated at the angle of the jaw. The chin arc originated and terminated at the angle of the jaw and passed through the horizontal extremum at the chin. Jaw articulation was accomplished by conjoint variations in the vertical positions of the chin extremum and jaw angles. The stationary facial schema was computed so that the outlines overlaid a set of point co-ordinates obtained from photographs of human faces in a resting position which was supplied as input data. The display was animated by computing a sequence of frames corresponding to 1/25 sec. intervals in which jaw and lip co-ordinates were varied in accordance with the data obtained from facial measurements (see below). The frame-by-frame data computations and subsequent real-time display were implemented as two separate phases of a Fortran program running on a PDP 11/60 computer and calling special-purpose graphics routines to drive a satellite GT40 hardware display processor (12).

Facial measurements

A pilot experiment has been carried out to record on a Sony reel-to-reel videotape recorder the front-facial image of a human speaker, both in a resting position (that is, with the jaw and lips lightly closed) and during the enunciation of a series of VCV utterances. The speaker's lip margins were marked for better visibility and his head was held steady in a frame. The videotapes thus obtained were replayed through a video interface device (13) which superimposed moveable video 'crosswires' on the television monitor screen and permitted the co-ordinate positions of their intersection with respect to the screen axes to be transmitted at command to a linked Cromemco Z2 microprocessor. For each frame of the video replay measured, the inner eye corners (a and b) and a spot mark on the left earlobe (c) were recorded as fixed, non-articulatory

points for standardisation purposes. The remaining facial points to be measured were determined by the need to supply the driving data for the graphics programme described above. They were the central horizontal extremum and corner co-ordinates of each of the four lip margins and the chin horizontal extremum.

A Fortran program running in the microprocessor was used to compute the results. A 'master' frame of the resting face provided the master standardisation co-ordinates of a, b and c against which the equivalent data from all succeeding frames were reduced to a common morphology (14,15). All recorded co-ordinates for every frame were adjusted so that the line from a to b was rotated into the screen x-axis with its mid-point as origin, thus fixing an x-axis. All recorded co-ordinates for every frame except the master frame were also sheared in the x-axis to fix a standardised y-axis based on the position of point c, then linearly scaled in the newly-fixed x- and y-axes. With the co-ordinates of a, b and c standardised for all frames, variations in the co-ordinates of the facial articulators could be attributed to real movement, with projection errors of less than 1%. Statistics concerning experimental accuracy were derived from replicated observations of the resting face and true movements of the lip and jaw during utterance enunciation were computed as percentage deviations from the resting face positions of i) the y-ordinates of the chin and central lip horizontal extrema, ii) the mean y-ordinates of the corresponding lip margin corners, and iii) the linear separation of the corresponding lip margin corners. These position data defined each frame of the animated computer graphics display and were derived in the way described so as to overcome the bilateral asymmetries of natural facial features.

Results and discussion

Facial measurements were obtained for repeated utterances of the syllable /aba/. They confirm earlier findings regarding the relatively slow but continuous movement of the mandible compared with the lips, which rapidly reach a nearly stationary configuration following the plosive release (4). Unfortunately the pilot studies were not totally successful because the camera was unshuttered and the time constants of the tube produced heavy picture smearing, especially when recording the rapid consonantal articulations. In addition, the videotape transport system was incompatible with the reproduction of clear single-frame pictures during 'stopped motion' replay. This introduced errors into the accurate positioning of the video crosswires. Both problems may be overcome by using more sophisticated shuttered cameras with a storage disc video replay machine. However, the position data were adequate to drive the computer graphics display programme and a more thorough-going programme of measurements seems to be justified.

The computer graphics package has been used to produce an animated display in real-time of VCV utterances. The currently displayed facial schema does not include articulators such as the teeth or tongue-tip. These deficiencies are clearly important if VCV utterances including, for example, high vowels or labiodental fricatives are to be produced, but do not seriously affect utterances in which the consonant is, for example, a bilabial stop, as in /aba/.

Other, possibly significant, perceptual cues such as facial skin-folds are likewise omitted. Furthermore, the use of simple circular arcs to represent the lip

margins becomes questionable in certain articulations, particularly those in which there is a high degree of lip-protrusion; a situation which may be improved by substituting more complex outline shapes such as ellipses, in place of the circular arcs. This extension, like the inclusion of further facial features, is facilitated by the modular, feature-oriented design of the existing graphics package. Beyond the solution of hidden-line problems, the addition of further features poses no conceptual difficulty. However, the attainment of reasonable display quality within an acceptable computational time-scale for an utterance of about 3/4 sec. duration (or about 19 frames) is exploiting almost the total power of the current graphics system software. Further graphics display developments will therefore impose rigorous demands upon the use of existing or extended software.

The author wishes to thank Prof M P Haggard, Director of the MRC Institute of Hearing Research under whose auspices the work was carried out, Dr A Q Summerfield for his advice and guidance and the trustees of the Piggott-Wernher Trust for their invaluable assistance.

References

1. 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/.
2. J.S. PERKELL 1969 Physiology of speech production: results and implications of a quantitative cineradiographic study (MIT Press).
3. S. E. G. OHMAN 1966 J. Acoust. Soc. Amer. 39, 151-168. Coarticulation in VCV utterances: spectrographic measurements.
4. O. FUJIMURA 1961 J. Speech and Hearing Res. 4, 233-247. Bilabial stop and nasal consonants: a motion picture study and its anatomical implications.
5. D. JONES 1976 An outline of English phonetics (Cambridge University Press).
6. P. MERMELSTEIN 1973 J. Acoust. Soc. Amer. 53, 1070-1082. Articulatory model for the study of speech production.
7. D. W. BOSTON 1973 Brit. J. Audiology 7, 95-101. Synthetic facial communication.
8. F. I. PARKE 1975 Comput. and Graphics 1, 3-4. A model for human faces that allows speech-synchronised animation.
9. C. A. FOWLER, P. RUBIN, R. E. REMEZ and M. T. TURVEY 1978 Language production (ed. B. Butterworth, Academic Press). Implications for speech production of a general theory of action.
10. H. FUJISAKI 1978 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.
11. W. J. HARDCASTLE 1976 Physiology of speech production (Academic Press).
12. DIGITAL EQUIPMENT CORPORATION 1973 Picture Book reference manual.
13. N. M. BROOKE and J. R. TRINDER Paper in preparation.
14. G. P. RABEY 1978 Brit. J. Surgery 15, 97-109. Current principles of morph-analysis and their implications on oral surgical practice.
15. G. P. RABEY 1971 Proc. Roy. Soc. Med. 64, 103-111. Craniofacial morphanalysis.