Reading the Speech of Digital Lips: *Motives and Methods for Audio – Visual Speech Synthesis*

Darryl Storey and Martin Roberts

Department of Computer Studies, Loughborough University of Technology

Visible Language XXII, 1
Darryl Storey and Martin Roberts, pp. 112–127

The widespread practice of lipreading among the hearing impaired has, for a number of years, stimulated research into the feasibility of transmitting visible images of articulation to accompany acoustically conveyed speech, in those circumstances where visual reinforcement of the speech signal is typically lacking. Although there already exist several systems which, exploiting computer graphics, are capable of generating animated images of articulation while allowing for eventual audio/visual synchrony, each is open to criticism on the grounds of its perceptual inadequacy and/or cost. This paper offers a brief review of these initiatives to date and describes the recent development of a relatively simple, effective, and hence economical method of audio/visual speech synthesis.

Introduction

Aspects of visible speech

The skill of lipreading (speech reading) is practised by many partially hearing listeners in order to offset, at least to some degree, their own aural limitations. For the hearing-impaired exponent, lipreading provides a perceptual supplement by means of which the intelligibility of heard speech may be substantially enhanced. It is important to recognize however, that such improvement is not brought about as an consequence of the lipreader's experiencing some visual analog of amplification. Speech which is both seen and heard frequently appears more intelligible than any wholly acoustical counterpart, because the separate visual and acoustical representations are perceptually complementary.

Those of us with normal hearing can, of course, hold intelligible conversations without necessarily facing each other in order to do so. Unimpaired, the auditory system is perfectly able to detect and resolve all of the changes in acoustical frequency, and their relationships in time, which typify human speech. All too often however, hearing loss manifests itself as a reduced sensitivity to specific frequency regions of the acoustical spectrum; those very regions within which a variety of linguistic/phonetic distinctions are portrayed in sound.

Under such circumstances one's auditory perception of the temporal order of acoustical events may remain reasonably acute while the 'identity' of certain phonetic events may be obscured. The benefit of lipreading resides in the perceptual restoration of such lost phonetic identities. For example, a commentator's pronouncement to the effect: "I sought the President's opinion. . ." could easily be misconstrued by a listener with a hearing loss as "I thought President's opinion. . ." or worse still, "I fought the President's opinion. . ." due entirely to the perceived similarity between the productions of "s", "th", and "f". Visually, however, these elements of our phonetic repertoire are distinct. This situation pertains virtually across the board as far as place of articulation is concerned. For a hearing–impaired listener to distinguish "pot" from "tot", and "cot", seeing the word spoken is at least as useful as hearing it, if not more so.

Simply being able to view a talker's face in action however, is not necessarily the *sine qua non* of lipreading.

Simply being able to view a talker's face in action, however, is not necessarily the *sine qua non* of lipreading. Even visible enunciations may be 'more' or 'less' clear to interpret. Moustaches and beards can be detrimental to the lipreader's art and poor lighting conditions, coupled with an idiosyncratic articulatory style, can result in relatively little of a talker's speech being visually informative. Notably, for the purpose of lipreading, it is particularly helpful for the viewer to be able to discern the behavior of the talker's tongue!

In point of fact, perceptual reference to the visual aspect of speech communication is not a trait of the hearing-impaired exclusively. The capabilities of highly skilled lipreaders actually reflect something approaching the asymptote in exploitation of a tendency which normal hearing viewer/listeners share, i.e. the *de facto* integration of synchronous auditory and visual images of speech. Such a tendency is born out, not so much by the perceptual integration of ostensibly compatible components (there were many connected with the silent film industry who doubted that 'the talkies' would work) as by convincing experimental demonstrations of the ease with which seemingly incompatible elements could be fused, resulting in audio-visual 'illusions' of the type first reported by McGurk and Macdonald (1976). For many observers, the perceptual consequence of attending to an audio-visual presentation of the acoustically unambiguous syllable /ba/ synchronized with a video image of the syllable /ga/ is not /ba/ nor /ga/ but /da/.

Considerable research has gone into mapping out the common articulatory ground which visual and acoustical instances of speech clearly share (e.g., Summerfield, 1979; Campbell and Dodd, 1980). Running parallel with such examinations have been various endeavours to analyze the visual concomitants of speech, both for the purpose of understanding articulatory processes in speech production (e.g., Perkell, 1986; Abry and Broe, 1986), and in the hope of providing the fundamental data from which 'artificial' representations of human talkers might be reliably reconstructed (e.g. Montgomery, 1983; Montgomery and Jackson, 1983; Brooke and Summerfield, 1983).

Computer faces: The problems that computers face

A variety of methods for computer-controlled synthesis of visible articulatory gestures have been explored to date (e.g. Boston, 1973; Erber and De Filippo, 1978; Montgomery, 1978; Brooke and Summerfield, 1983), although not as contributors to audio-visual productions necessarily. Arguably the most sophisticated of such synthetic faces, both computationally and graphically, is that developed by Parke (1975, 1982). This three computationally dimensional model *has* been incorporated into a system for synchronous audio-visual synthesis (Pearce, Wyvill, Wyvill and Hill, 1986).

None of these methods are entirely without limitation however, either as regards their visual sufficiency or their cost-effectiveness. The vector (outline) graphic images generated by the system of Montgomery (1978) for example, deal somewhat less than adequately with the problem of representing the behavior of the talker's tongue. Brooke's (1982) method, and the infinitely more complex model of Parke (1975), each avoid this issue, although ultimately they cannot escape it, by omitting the tongue altogether. It could of course be argued that this actually constitutes a more, rather than less authentic interpretation, since, during face-to-face communication, the talker's tongue is largely obscured within the shadow of the oral cavity. However, since the objective is to supplement the listener's acoustical analysis with a complementary visual one, strategies for the development of visual prostheses aimed at enhancing the perception of speech, and which exclude from the outset certain known and powerful visual cues, must be open to question.

The readiness with which the hearing-impaired appeal to lipreading, and its demonstrable efficacy as a perceptual strategy (Summerfield, 1987), even among 'normal' listeners (Reisberg, McLean and Goldfield, 1987), argue strongly for the visual reinforcement of speech in those settings where it is not usually provided, e.g., during public address annoucements, telephone con-versations, and radio broadcasts. Nevertheless, success at such a task would require a robust methodology for inferring

high-level articulatory 'primitives' from the acoustical outcome of articulatory gestures, i.e., the capability for working 'backwards' toward some explicit, unambiguous, and continuously changing specification of an articulatory source from what are, in effect, residual acoustical data. These primitives would then have to be reinterpreted as graphical, rather than acoustical coordinates. However, there is, as yet, a complete absence of any appropriate repertoire of such articulatory parameters.

The difficulties these objectives present have been emphasized elsewhere by others (e.g. Sondhi and Resnick, 1983; Levinson and Schmidt, 1983). Of the many methodological and engineering problems, two in particular need to be overcome. The first is the one-to-many relationship between speech events and the variety of possible configurations of the sound source from which these events might plausibly originate. The second is an analogous, and inverse, many-to-one mapping. Listeners commonly assign an equivalent perceptual label to quite discrepant acoustical structures (the different members of any linguistic community are perfectly intelligible, each to the other, despite their acoustically distinct spoken versions of particular language tokens. Although the phonological rules of language offer some hope for one's being able to constrain the number of options which might need to be considered by automated 'decision' processes, any undertaking of the kind considered here would remain formidable.

Of the many methodological and engineering problems, two in particular need to be overcome. The first is the one-to-many relationship between speech events and the variety of possible speech configurations of the sound source from which those events might plausibly originate. The second is an analogous, and inverse, many-to-one mapping.

Given that graphical enhancement of acoustically realized speech would be both useful and desirable, the most straightforward approach to these issues would be to deal, in the first instance, with synchronous audio-visual synthesis. With the output of both graphical and acoustical media under one's direct control, it is pos-sible to address, purposefully, at least one significant problem; that of achieving flexible, yet synchronous audio-visual output.

First Principals

A cornerstone of much speech science research is the well documented discrepancy between the phonetic nature of our perceptual interpretation of speech, and the acoustical signal's being wholly lacking in any

correspondingly discrete elements (e.g. Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967). When speech is articulated it emerges as a dynamic, continuously changing pattern of sound. When speech is perceived it registers as a succession of phonologically defined components of the listener's language. Thus a more-or-less continuous signal is, apparently, *analyzed* into a sequence of more-or-less descrete percepts. An analogous, though converse 'discrete-continuous' dichotomy pertains with respect to our perception of apparent motion. It has been known since the earliest days of animation with the Victorian Zoetrope, that discrete images, when presented in sufficiently rapid succession, can induce the perception of continuous movement. In this instance discrete events are synthesized into a perceptual representation of seemingly continuous activity [see Ramachandran and Anstis (1986) for a more detailed discussion].

A more-or-less continuous signal is, apparently, *analyzed* into a sequence of more-or-less discrete percepts.

Taking these two observations together, suggests a way in which a graphical simulation of articulation might be arranged to coincide with an acoustical complement. A set of individual frames, representing discrete phonetic states, and presented in rapid succession, could be perceptually representative of articulatory motion. The complexities inherent in alternative approaches, and the force of 'apparent motion' phenomena, encouraged our exploration of this particular class of audio-visual relationship.

An Audio-Visual Synthesis System

The components

Real-time audio-visual synthesis has been accomplished using a BBC microcomputer (plus associated disk drive and monitor) in conjunction with a digital speech synthesizer provided by Loughborough Sound Images (LSI) limited. Acoustical productions are defined by a number of parameters stored on, and retrieved from disk. These parameters are interpreted by the synthe-sizer's digital circuitry, at a uniform sample rate, as settings for a series of electronic filters, which, in turn, simultaneously pass modified signals to the loudspeak-er. Each sample is separated from its predecessor by an interval of 10 milliseconds. The filtered signals, when realized in analog form, constitute speech sounds.

The graphical constituents of the system owe their origin to a Micro-Robotics 'Snap Camera'. This small device connects directly to the microcomputer. It focuses light, not onto a film plane, but onto an array (256 x 128) of light-sensitive cells. The inital, stable voltage output of each cell defines its state to the computer as 'on'. Light striking the cells of the camera causes them to discharge until they reach a voltage 'ceiling' at which point they turn themselves 'off '. These 'on/off ' states are interpreted by the computer as 'black' and 'white' respectively. The resultant distribution of black and white locations within the 256 x 128 point (pixel) array can then be stored, manipulated, and displayed to the VDU.

These are essential devices in themselves, but the key to operational success is a computer program which allows a recorded image to be edited interactively. The values stored within the global array can be changed, i.e., reversed, by manipulating them within 8 x 8 pixel blocks, a block at a time. By editing the data in this way, any recorded image can be completely transformed, as figures 1- 3 illustrate. The result can be made to appear as 'nice' or as 'nasty' as may be necessary.

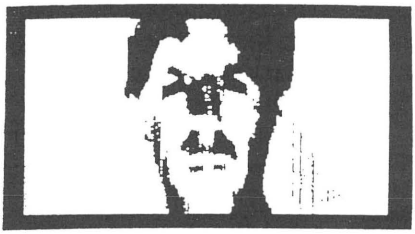

Figure 1
Initial face image recorded via the 'Snap Camera'.

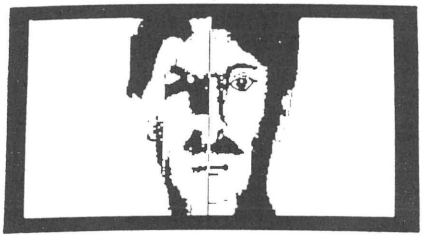

Figure 2
Transitional image achieved through the use of a 'pixel editor '.

Figure 3
Final image for animation.

Illumination and image enhancement

Under normal daytime conditions, the light we receive originates entirely from the one source, i.e., the sun. Although diffused, diffracted and scattered to some extent by the earth's atmosphere, natural sunlight tends to illuminate objects differently according to their position relative to the light's origin. The perceptual experience of 'highlight' and 'shadow' in our visual interpretation of objects in the world is therefore commonplace. In view of the inadequacy of natural daylight as a source of illumination for digital images captured via the Snap Camera, the face eventually recorded as a basis for graphical articulation, was illuminated in the first instance using two artifical light sources (100 watt standard lamps). This resulted in an altogether 'unnatural' image in terms of inherent light and shade. The artificial effect was counteracted through subsequent editing of the image, which was reconstructed to give the more naturalistic impression of its having been illuminated from one angle rather than two. Only those 'shadows' were retained which were consistent with this interpretation.

It was considered fundamental that the image area representing the mouth should be enhanced so as to give prominence to the lips, teeth, and tongue. With only black and white pixels available, this proved rather difficult at first. However, careful study of the monochrome photographs reproduced in newspapers, confirmed that reasonable effects may be obtained by highlighting black lips with white edges. The teeth, for example, could be represented as a few white pixels with a black outline, while the tongue could be made to appear white upon a black background, representing the oral cavity. Individual static positions for the mouth were determined by scrutinizing that of a subject asked to prepare, as it were, the articulation of various phonemes. These were modelled in an exaggerated fashion initially in order to arrive at an adequate approximation of the phonemes for storage by the computer. 'Fine tuning' of the various representations was done later with 'highlights' and 'shadows' touched in via the pixel editor where necessary.

It was considered fundamental that the image area representing the mouth should be enhanced so as to give prominence to the lips, teeth, and tongue.

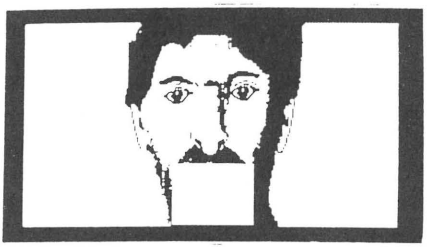

Figure 4
Enhanced facial image with animation window 'cleared'.

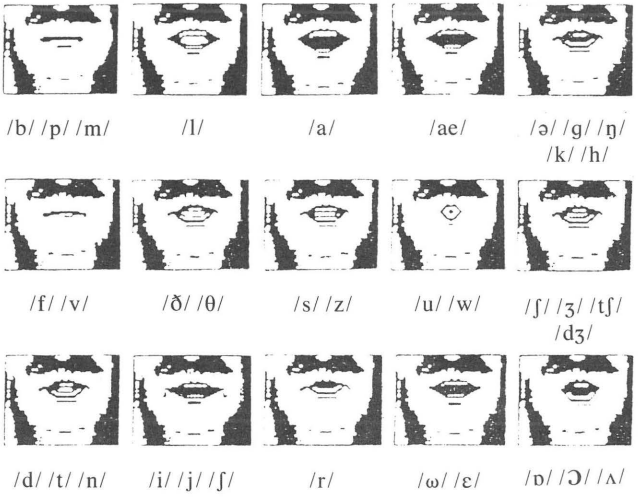

Figure 5
Individual frames representing 15 articulatory positions.

Animation and audio-visual synchrony

Animation of the digital face centers upon changes made to the values stored, and hence the images produced, in a small (72 x 64 pixel) 'window', the relative screen location of which is shown in figure 4. Fifteen static articulatory configurations (of lips, teeth, and tongue) have been predefined, as in figure 5. They are each stored at run time in separate areas of memory. From these locations the 'frames' can be recalled and transferred to the area supporting the graphical window, their insertion simply overwriting the prior resident data and concomitant screen image in less than 10 milliseconds. (The repertoire of fifteen discrete articulatory positions attempts to exploit the intuitive observation that groups of speech sounds, at least as phonetically defined, are visually similar if not identical. English equivalents of the phonetic captions employed are given in Table 1.) The perceptual effect of a change from the graphical realization

of an articulatory position appropriate to the bilabial stop consonant /b/, say, to that representing the vowel /a/, is unquestionably appearance of the syllabic gesture /ba/, motion being inferred between the two positional extremes. (This sequence could of course be construed as /ma/ or /pa/.)

The speed with which changes can be wrought upon the display of this smaller window to the VDU assumes particular importance for the eventual synchrony of such changes with complementary instances of acoustical synthesis. No modification of the instruction set passed to the speech synthesizer can become effective in less than 10 milliseconds, whereas shifting of the screen images can. Thus, even the most rapid of consonant-vowel articulations, as portrayed acoustically by the LSI device at least, can be adequately encapsulated within a synchronous graphical event.

Any instance of speech prepared in a format admissable as input by the LSI synthesizer can be coupled with a synchronous graphical interpretation. Manual scrutiny of the parameter values required to drive the synthesizer reveals the locations of disjuncture within this table of values, corresponding to concomitant acoustical, and hence articulatory change. Since the speech synthesizer receives its instructions via the microcomputer it is necessary only to interleave these with the requisite code for generating the appropriate graphical sequence, each frame in the sequence occupying the window only for so long as is necessary to accompany the speech synthesis to its next point of acoustical departure. The preliminary visual analysis implicated in this set of procedures has in fact been automated. A program has been written which distills the discontinuities from within the synthesis parameters automatically. Determining which of the 15 frames has to be invoked at any point however, remains a manual, or rather auditory exercise for the present.

Text–to–(audio-visual) speech conversion

Anyone familiar with the operation of software-controlled speech synthesizers will appreciate how tedious is the preparation of perceptually adequate utterances.

Table 1
phonetic symbols for transcribing English consonants

p	pie	pea	
t	tie	tea	
k	kye	key	
b	by	bee	
d	dye	D	
g	guy		
m	my	me	ra*m*
n	nigh	knee	ra*n*
ŋ			ra*ng*
f	fie	fee	
v	vie	V	
θ	thigh		
ð	thy	thee	
s	sigh	sea	
z		Z	mi*z*zen
ʃ	shy	she	mi*ss*ion
ʒ			vi*s*ion
l	lie	lee	
w	why	we	
r	rye	re	
j		ye	
h	high	he	

Note also the following:

tʃ	chi(me)	chea(p)
dʒ	ji(ve)	G

phonetic symbols for transcribing English vowels

i	heed	he	bead	heat	keyed
ɪ	hid		bid	hit	kid
eɪ	hayed	hay	bayed	hate	Cade
ɛ	head		bed		
æ	had		bad	hat	cad
ɑ	hard		bard	heart	card
ɒ	hod		bod	hot	cod
ɔ	hawed	haw	bawd		cawed
o	hood				could
oo	hoed	hoe	bode		code
u	who'd	who	booed	hoot	cooed
ə	herd	her	bird	hurt	curd
ʌ	Hudd		bud	hut	cut
aɪ	hide	high	bide	height	
aʊ		how	bowed		cowed
ɔɪ		(a)hoy	Boyd		
ɪə		here	beard		
ɛə		hair	bared		cared
aə	hired	hire			

Hence our own demonstrations of audio-visual synchrony utilizing the LSI device have been confined, thus far, to two examples only, i.e. the phrases: "A bird in the hand is worth two in the bush" and "An apple a day keeps the doctor away". Nonetheless, these efforts were sufficient to establish that the principles are fairly robust. Interestingly, informal observation of the graphical animations in isolation reveals their fragmentary nature. It is the synchronous occurrence of an acoustically dynamic signal which lends coherence to what is then interpreted, perceptually, as a unitary event. This observation complements, in a somewhat contradictory way, those illusory audio-visual demonstrations mentioned earlier. The 'McGurk effect' is itself an example of perceptual coherence being maintained despite apparently *discrepant* information arriving at the senses. In such illusions, it is the visual component of the ensemble which tends to impose its overall perceptual structure. This particular situation is the converse of that noted here, although a formal examination of the degree to which our fragmentary animations may compensate for acoustically impoverished speech may yet bring the two rather divergent observations into line.

Informal observation of the graphical animations in isolation reveals their fragmentary nature. It is the synchronous occurence of an acoustically dynamic signal which lends coherence to what is then interpreted, perceptually, as a unitary event.

To provide for faster entry of speech data to the synthesizer than is customarily possible, LSI have implemented a software facility which admits the input of pseudo-phonetic strings. Such an input sequence is converted automatically into the most appropriate, acoustically context-dependent interpretation, in terms of filter source parameters for the machine, the output function of which remains as previously described. A recent project (Strawbridge, 1986), has resulted in the provision, additionally, of a module for 'translating' standard English orthography into the quasiphonetic form of representation adopted by LSI. Although this latter system currently resides within a mainframe computer, it should prove relatively straightforward to replicate its operation using the BBC micro. We therefore have all of the necessary components at least for realizing a system for the conversion of text to fairly high quality audio-visual speech. In the meantime, however, we have been exploring alternative, though admittedly less elegant vehicles for speech synthesis via the BBC microcomputer, in particular the "Speech!" utility

A recent project has resulted in the provision, additionally, of a module for 'translating' standard English orthography into the quasiphonetic form of representation adopted by LSI.

marketed by Superior Softward Ltd., which is both cheap and readily available.

Careful disassembly of the appropriate constituent program from the "Speech!" package revealed that it would admit the predication of our own graphics routines (also written in assembly language). The unadulterated 'speech' program enables, within reason, acoustical synthesis from orthographical input. Since our own graphical extension is designed to follow, explicitly, the sequence of phonetic events as determined by the host program, the fundamental text-to-speech facility is unimpaired by the modification. Indeed, phonetic misinterpretation of any orthographical string is carried over to the graphical aspect also, the two coexisting programs being synchronously compatible, literally to a fault.

Enhancements ought undoubtedly to follow, but the most significant, and demanding challenge; that of driving the graphics from an analysis of acoustical output remains to be confronted. At the very least we have an extremely cost effective medium with which to evaluate any theoretical steps taken in that particular direction in the future.

Acknowledgement

The guidance and encouragement of Dr. Q. Summerfield of the MRC Institute of Hearing Research, Nottingham, is gratefully acknowledged.

About the authors

Darryl Storey gained his B.Sc. in Mechanical Engineering from Lanchester Polytechnic, Coventry in 1982. He then spent a period in industry as a Development Engineer and a year in local government as a Computer Programmer. Since 1984 he has held the post of Experimental Officer in the Department of Computer Studies, Loughborough University of Technology.

Martin Roberts hold a Ph.D. in Experimental Psychology from Nottingham University. His research interests are in distinguishing sensory from cognitive influences upon speech perception and the application of audio-visual experimental techniques to that end. He is currently a Research Assistant in the Department of Computer Studies, Loughborough University of Technology.

References

Abry, C. and Broe, L. J. 1986. Laws for lips. *Speech Communication,* 5, 97–104.

Boston, D. W. 1973. Synthetic facial communication. *British Journal of Audiology,* 7, 95–101.

Brooke, N. M. 1982. Video speech synthesis for speech perception experiments. *Journal of the Acoustical Society of America,* 71, S77(A).

Brooke, N. M. and Summerfield, Q. 1983. Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics,* 11, 63–76.

Campbell, R. and Dodd, B. 1980. Hearing by eye. *Quarterly Journal of Experimental Psychology,* 32, 85–99.

Erber, N. P. and De Filippo, C. L. 1978. Voice/mouth synthesis and tactile/visual perception of pa, ba, ma. *Journal of Acoustical Society of America,* 64, 4, 1015–1019.

Levinson, S. E. and Schmidt, C. E. 1983. Adaptive computation of articulatory parameters from the speech signal. *Journal of the Acoustical Society of America,* 74, 4, 1145–1154.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychological Review,* 74, 6, 431–461.

McGurk, H. and Macdonald, J. 1976. Hearing lips and seeing voices: A new illusion. *Nature,* London, 746–748.

Montgomery, A. A. 1978. Generation and evaluation of synthetic facial images for lip-reading. *Paper presented at the annual meeting of the American Speech and Hearing Association,* November, 1978.

Montgomery, A. A. 1983. The search for invariant visible cues in lipreading. *Journal of the Acoustical Society of America,* 73, S15.

Montgomery, A. A. and Jackson, P. L. 1983. Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America,* 73, 6, 2134–2144.

Parke, F. I. 1975. A model for human faces that allows speech-synchronized animation. *Computers and Graphics,* 1, 3–4.

Parke, F. I. 1982. Parameterized models for facial animation. *IEEE Computer Graphics and Applications,* 2, 9, 61–68.

Pearce, A., Wyvill, B., Wyvill, G. and Hill, D. 1986. Speech and expression: a computer solution to face animation. In: *Proceedings of the graphics interface '86 conference*, Vancouver, Canada, May 26th–30th.

Perkell, J. S. 1986. Coarticulation strategies: preliminary implications of a detailed analysis of lower lip protrusion movements. *Speech Communication,* 5, 47–68.

Ramachandran, V. S. and Anstis, S. M. 1986. The perception of apparent motion. *Scientific American,* 254, 6, 80–87.

Reisberg, D., McLean, J., and Goldfield, A. 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In: B. Dodd and R. Campbell, (Eds.). *Hearing by Eye: Experimental Studies in the Psychology of Lipreading.* London, Lawrence Erlbaum Associates.

Sondhi, M. M. and Resnick, J. R. 1983. The inverse problem for the vocal tract: numerical methods, acoustical experiments, and speech synthesis.*Journal of the Acoustical Society of America,* 73, 3, 985–1002.

Strawbridge, K. P. 1986. The development of a grapheme-to-phoneme translation algorithm in Prolog for use with the LSI Phonetic Synthesizer. *Unpublished M.Sc. dissertation,* Loughborough University of Technology.

Summerfield, Q. 1979. Use of visual information for phonetic processing. *Phonetica,* 36, 314–331.

Summerfield, Q. 1987. Preliminaries to a comprehensive account of audio-visual speech perception. In: B. Dodd and R. Campbell, (Eds.), *Hearing by Eye: Experimental Studies in the Psychology of Lipreading.* London, Lawrence Erlbaum Associates.