PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Lipreading and Audio-Visual Speech Perception

Quentin Summerfield

Phil. Trans. R. Soc. Lond. B 1992 **335**, 71-78
doi: 10.1098/rstb.1992.0009

Lipreading and audio-visual speech perception

QUENTIN SUMMERFIELD

MRC Institute of Hearing Research, University Park, Nottingham NG7 2RD, U.K.

SUMMARY

This paper reviews progress in understanding the psychology of lipreading and audio-visual speech perception. It considers four questions. What distinguishes better from poorer lipreaders? What are the effects of introducing a delay between the acoustical and optical speech signals? What have attempts to produce computer animations of talking faces contributed to our understanding of the visual cues that distinguish consonants and vowels? Finally, how should the process of audio-visual integration in speech perception be described; that is, how are the sights and sounds of talking faces represented at their confluence?

1. INTRODUCTION

Lipreading† is useful to all sighted people, including those with normal hearing and those with impaired hearing. It confers benefits because the visible articulators, primarily the lips, teeth, and tongue, are among those whose configurations determine, and therefore are correlated with, the resonances of the vocal tract. The changing frequencies of the resonances help convey important phonetic aspects of speech including the identities of vowels and diphthongs and the place of articulation of consonants (e.g. 'b' versus 'd'). The evidence of the resonances is the least robust aspect of the acoustic speech signal. Thus, lipreading is beneficial because it can compensate rather specifically for the deficiencies of audition. As an example of the circumstances in which it is consequently useful, in quiet, the redundancy between the evidence of the resonances provided by vision and by audition enables subjects with normal hearing to perceive speech more accurately, even when the acoustics are undistorted (Reisberg *et al.* 1987). In noise (and also in reverberation and in cases of sensorineural hearing impairments), the auditory representation of the resonances in the mid to high frequencies is often distorted (see, for example, Summerfield (1987*a*)). Here, lipreading can play a complementary role. It enables subjects to tolerate an extra 4–6 dB of noise while maintaining performance at the level achieved when only listening (Middleweerd & Plomp 1987; MacLeod & Summerfield 1990). Four to six decibels may seem unimpressive, but for sentence material each decibel of signal-to-noise ratio (SNR) gained is equivalent to a 10–15% improvement in intelligibility. Listeners with sensorineural hearing impairments require the SNR to be improved by about 1 dB for every 10 dB of hearing loss if they are to perceive speech as accurately as listeners with normal hearing (Duquesnoy 1983). Thus, lipreading can compensate quite well for some of the consequences of moderate hearing losses.

In cases of profound or total hearing loss, lipreading on its own allows speech to be understood fluently by very few people, but it provides a basis for understanding by the majority when supplemented by appropriately tailored acoustical or electrical signals. These are signals that provide evidence of the articulatory activity that lipreading cannot detect, most obviously the activity of vocal folds in the larynx. The evidence may be conveyed directly by a sinewave whose frequency is modulated according to the rate of vibration of the vocal folds (Fourcin *et al.* 1979; Rosen *et al.* 1981), or indirectly by signals whose amplitude is correlated with the presence or absence of vocal-fold activity (Breeuwer & Plomp 1984; Grant *et al.* 1991).

Given the limited access to articulation which vision has, it may seem surprising that anybody can understand connected speech solely by lipreading. Even in carefully articulated syllables, the 22 consonants of English cannot be identified as such, but are (at best) categorized in eight or nine distinguishable groups (Walden *et al.* 1977). So how is the task possible? Some clues are provided by accounts of the auditory perception of words in fluent speech. These accounts stress the problems posed by the virtual absence of acoustic markers of boundaries between words and by the existence of phonological recoding across word boundaries (Klatt 1979). One solution is for the perceptual process to work sequentially, with the identification of each word establishing that a new word is beginning (Cole & Jakimik 1980). Perception of word boundaries would occur to the extent that phonetic, lexical, syntactic, and semantic factors rapidly eliminate alternative word candidates (Marslen-Wilson &

† Lipreading is the perception of speech purely visually by observing the talker's articulatory gestures. Audio-visual speech perception is the perception of speech by combining lipreading with audition. Speechreading embraces a larger set of activities. It is the understanding of speech by observing the talker's articulation and facial and manual gestures, and may also include audition.

Phil. Trans. R. Soc. Lond. B (1992) **335**, 71–78
Printed in Great Britain

Welsh 1978). Logically, the same factors should apply in lipreading. Words that are easy to lipread.should be familiar, should start with visibly distinctive consonants that define their onsets clearly, and should look like few other words. The word 'boy' conforms to these three criteria. It is familiar; it starts with a visibly distinctive bilabial closure; the diphthong 'oI' looks like few other vowels or diphthongs; the syllables that look similar to it, 'poy' and 'moy', are not words. Thus, 'boy' should be easy to lipread provided the syntactic or semantic context, or the onset of the following word, eliminate the possibility that is the first syllable of 'buoyant', 'boycott', 'boisterous', etc. MacLeod (MacLeod 1987; MacLeod & Summerfield 1987) partially confirmed these principles by analysing the ease with which words in a corpus of test sentences could be lipread. 'The boy's running away', 'The boy got into trouble', and 'The boy forgot his book' are lipread correctly by many people. On the other hand, sentences composed of words that do not conform to the three criteria, such as 'The sun melted the snow' and 'The three girls are listening', defeat nearly everyone. Thus, the visual ambiguity of individual consonants and vowels can be overcome if they form words that look like few other words.

Against this general background, four specific issues have interested me and my colleagues. They are framed as questions in the abstract of this paper. The following sections summarize what is known about them and pose the questions still outstanding.

2. WHAT DISTINGUISHES BETTER FROM POORER LIPREADERS?

Individual differences in the ability to lipread are large. Scores on sentence tests of lipreading often range from less than 10% correct to over 70% correct among a group screened to have normal vision and to be homogeneous in respect of hearing status and age (e.g. profoundly hearing-impaired children, 11–93% (Heider & Heider 1940); normal-hearing young adults, 1–50% (MacLeod & Summerfield 1987); moderately hearing-impaired adults, 15–85% (Dodd *et al.* 1989)). In the case of young adults with normal hearing, this pattern contrasts with the relative lack of variability found in tests of auditory speech reception (see, for example, van Rooij *et al.* (1989)). Thus, individual differences in lipreading skills reflect something other than normal variation in speech-perceptual abilities.

The extent of these individual differences is intriguing from the point of view of cognitive science and has been of practical concern to teachers of the deaf. It prompted a large number of studies which attempted to identify sensory and cognitive correlates of the ability to lipread (see Jeffers & Barley (1971) for a review). Some caution has to be exercised in interpreting their results because many studies do not report the reliability of their scores. However, informative conclusions can be drawn by noting patterns of correlations over several studies. If this is done (e.g. Summerfield 1991), correlations with intelligence and verbal reasoning, the two abilities that might most obviously be expected to relate to the ability to lipread, are found generally to be low and not significant, although positive. McGrath (1985), reviewing the literature, concluded that good lipreading depends on a minimal level of intelligence and verbal ability but, provided this is attained, further ability is not important.

Demonstrations that lipreading is difficult to teach and to learn further support the idea that it is independent of other cognitive abilities. The difficulty has been shown in a variety of ways. Consider first that post-lingually hearing-impaired adults, who might in principle be expected to have had an interest in improving their lipreading skills, often prove to perform no better than normal-hearing adults (see, for example, Plant & Macrae (1981); Middleweerd & Plomp (1987); Lyxell & Ronnberg (1989)). Second, groups who have received training in lipreading may perform no better than untrained groups (see, for example, Conrad (1977); Dodd *et al.* (1989)). Third, training itself may produce no improvement (Binnie 1977). Alternatively, where improvement occurs, though significant and worthwhile, it is often small (an additional 10% correct) in relation to the spread of scores among subjects (Dodd *et al.* 1989, and references therein). Findings such as these initially called into question the tradition of oral education for profoundly inpaired children (see Farwell 1976) and latterly prompted calls for a reconsideration of the aims of lipreading classes for post-lingually deafened adults (Brooks 1989).

MacLeod & Summerfield (1990) showed the cognitive independence of lipreading in a different way. They used two tasks to measure the ability of a group of 20 young adults with normal hearing to lipread the content words in sentences. First, the subjects transcribed test sentences by observing a video-recording of the face of the talker with no sound. This test provided a measure purely of lipreading ability. It had a test–retest reliability, expressed as a correlation coefficient, of 0.92. Second, the subjects attempted to identify the words in a different set of sentences presented acoustically against a background of noise. An adaptive psychophysical procedure (Plomp & Mimpen 1979) was used to determine the minimal SNR at which subjects could perform the task. The test was run in an audio-alone mode and an audio-visual mode. The difference between the two thresholds averaged 6.1 dB and provides a measure of the average benefit from vision to speech perception in noise. It had a test-retest reliability of 0.80. The correlation between lipreading ability (from the first test) and visual benefit (from the second test) was 0.89. This is as high as could have been expected, given the reliabilities of the individual measures. It may seem trivial to report that the ability to lipread correlates highly with the ability to benefit from lipreading when listening is difficult. However, the force of the result depends on the fact that the measure of visual benefit was obtained by subtracting two measures made on the same subjects. In this way, individual differences among the subjects in all other cognitive and intellectual abilities were partialled out, yet the correlation remained. Thus, little

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

of the individual variation in the ability to lipread can reside in variation in these cognitive and intellectual abilities.

What has been found to correlate significantly with the ability to lipread words in connected speech? First, hardly surprisingly, are scores on other tasks of lipreading, such as the ability to lipread consonants and vowels in nonsense syllables or isolated words (Summerfield 1991). Second, more surprisingly, were reports of high negative correlations between the ability of listeners with normal hearing to lipread words in sentences and the latency of components of the visual evoked potentials recorded from the scalp about 130 ms after a bright flash of light. Although the size of the correlation declined and the mathematical sophistication required to reveal it increased over a set of replications of the original finding, the correlations remained impressively high (-0.90 to -0.91 (Shepherd *et al.* 1977); -0.61 to -0.89 (Shepherd 1982); -0.57 (Samar & Sims 1983)). The results suggest one answer to the question of why lipreading does not correlate consistently with other measures. Skill in lipreading may depend largely on the speed of aspects of low-level visual-neural processing. If the ability resides in aspects of subjects' hard-wired physiology, no wonder it is so hard to teach and learn.

More recently, the picture has been complicated by failures to replicate the original finding (Samar & Sims 1984; Ronnberg *et al.* 1989). Although these failures undermine the conclusion that the roots of good lipreading are set in fast visual-neural processing, they reinforce the idea that lipreading is a particular skill unrelated to other cognitive and sensory abilities. Summerfield (1991) observed that it fulfilled many of the requirements of impenetrability purported to be characteristics of cognitive modules (Fodor 1983). Montgomery & Demorest (1988) put it more generally: lipreading is an independent trait, probably hardened and untrainable in adulthood.

Three questions remain for further research, therefore. What aspects of heredity and environment lead to good lipreading? Is there a critical period for learning? Could it be exploited more thoroughly than in Nature's basic provision as an investment for old age and an insurance against hearing loss?

3. AUDIO-VISUAL ASYNCHRONY

Asynchrony between the image of a talker and the sound of her voice is distressing to infants (Dodd 1979) and irritating to adults (CCIR 1990). Three questions arise. What is the minimal detectable asynchrony? What is the minimal asynchrony that disrupts the audio-visual intelligibility of speech? Are they the same? Answers are important in designing signal-processing algorithms for hearing aids to be used by listeners with severe to profound hearing impairments. For this group, these devices function as 'aids to lipreading' part or all of the time. However, the processing that they perform to enhance the acoustical speech signal delays the acoustical signal. How much delay can be introduced before the benefits to intelligibility from the processed signal are undermined by the detrimental consequences of the asynchrony?

In 1980, estimates of the minimal detectable asynchrony varied widely. In the limit, where tests were performed by practised observers with brief abrupt non-speech signals such as a click and a flash of light, temporal order could be judged reliably when the asynchrony of onsets was as small as 20 ms (Hirsh & Sherrick 1961). On the other hand, when untrained observers were asked to adjust the delay between the sound track and image of a videorecording of a talker until they just perceived asynchrony, they set the sound to lead the image by as much as 150 ms or to lag by as much as 250 ms (Dixon & Spitz 1980).

Accordingly, McGrath & Summerfield (1983) sought to define the minimal detectable asynchrony using speech-like stimuli and a robust psychophysical technique. Subjects judged whether a pair of synthetic 'lips', generated as Lissajoux figures on an oscilloscope, opened before or after the start of a low-pitched buzz which was presented through headphones. On average, the acoustical signal had to lead lip opening by 80 ms or lag by 140 ms for subjects reliably to distinguish it from the case of simultaneous onset.

These results predict that the audio-visual intelligibility of connected speech should not be perturbed until the audio signal is delayed by 140 ms, or so. To test this prediction, we simulated one existing form of signal-processing aid to lipreading (Rosen *et al.* 1981) by detecting each moment when the talker's vocal folds closed, and generating an acoustical pulse at that moment. The result is a train of pulses which indicates the moments at which the talker is producing voice and, if so, what the fundamental-frequency contour (pitch contour) of her voice is. Subjects with normal hearing attempted to transcribe test sentences accompanied by this signal which was delayed by different amounts.

Hearing the acoustical pulse train with no delay increased identification of the content words in the sentences by about 15% compared with lipreading alone. Delaying the audio signal by 160 ms eliminated the benefit. Performance never fell below the level achieved by lipreading alone. Thus, subjects were able to ignore the delayed signal, rather than integrating it detrimentally. In the group mean data there was little effect of delays up to 80 ms. We concluded, rather tentatively given the particular form of stimuli used and the restricted range of age and the normal hearing of our subjects, that signal-processing should delay the signal as little as possible, but that delays of up to 40 ms (half the delay giving no overall decrement) might be acceptable†. Pandey *et al.* (1986) subsequently confirmed and extended this conclusion. They presented unprocessed speech in a background of speech babble and also found that delays of up to 80 ms had no significant effect on intelligibility. Thus, there should be ample time for many signal-processing algorithms to do their work. Problems would arise only with algorithms that needed to look ahead over

† The European Broadcasting Union has made a similar recommendation for the limits to asynchrony in television transmissions. Sound should lag vision by no more than 60 ms and should lead by no more than 40 ms.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

an appreciable stretch of speech to track parameters in noise.

Two further questions arise, therefore. Why is the minimal detectable asynchrony so long, and why is it asymmetrical? The asymmetry probably reflects the different response latencies of the receptor organs (20–100 ms in the retina; 1–4 ms in the cochlea; see King & Palmer (1985) for a discussion). The large absolute size of the thresholds may reflect broad distributions of response latencies in bimodally sensitive cortical neurones. This broad tuning, in turn, may be a necessary feature of a system that has evolved to use sight and hearing collaboratively to detect and interpret events occurring over a range of distances. The sight and sound of events occurring at distances up to 50 m are perceived to be synchronous if sound must lag sight by 140 ms to be detectably asynchronous. If the registration of audio-visual synchrony were more precise, audio-visual events would not be perceived to be in synchrony over this wide range of distances.

Unfortunately, there is a flaw in this tidy picture. Occasional findings suggest that some subjects can detect, and be disrupted by, shorter asynchronies. For example, when concerts are broadcast simultaneously on television and VHF radio channels, the sound may be delayed by up to 30 ms with respect to the picture. Apparently, broadcasting companies receive complaints from professional musicians who can detect the asynchrony. A second example was provided by our own data. Although intelligibility did not decline in the group mean data until sound lagged vision by 160 ms, the best five subjects who gained most from the acoustical signal with no delay (and therefore had the most to loose when delay was introduced) showed a linear trend of decreasing performance as delay increased up to 80 ms.

Thus, a key question remains to be resolved. Do those who gain most from the synergy of sound and vision, and who therefore have most to lose when delay is introduced, display reduced intelligibility with delays shorter than 80 ms, even if they cannot detect the asynchrony?

4. SYNTHESIS OF TALKING FACES

Much has been learnt about listeners' sensitivities to the acoustic cues that distinguish speech sounds by analysing natural productions and synthesising artificial utterances in which the spectro-temporal values of individual cues are manipulated precisely (see Klatt (1987) for a review). In the early 1980s, attempts were made to apply the same strategy to identify visual cues used in lipreading and to explore the perceptual processes that interpret the cues (e.g. Erber *et al.* 1980; Montgomery & Soo Hoo 1982; Brooke & Summerfield 1983; Walden *et al.* 1983; see Brooke (1991) for a review).

These attempts were informative, but ultimately were hampered by the limitations of computer-controlled vector graphics, as our own experience illustrates. Brooke (Brooke & Summerfield 1983; Brooke 1991) programmed an animated vector graphic of a talking face in which the configurations and movements of the lips, teeth, and jaw were generated from 13 points measured 25 times per second on the face of a natural talker. McGrath (1985; Summerfield *et al.* 1989) used the synthesizer to determine what role is played by the visibility of the teeth when subjects lipread vowels. A role for the teeth had been suggested by Montgomery & Jackson (1983). They had measured parameters describing the configuration of the mouth from videotapes of talkers producing American–English vowels and diphthongs. The parameters included measures of the height and width of lip opening and the area of the oral aperture. Multidimensional scaling was used to determine to what extent the physical measures could predict the perceptual confusions among the vowels made by lipreaders. Montgomery & Jackson found that only about 50% of the variability in the perceptual data could be accounted for in this way. They concluded that parameters not in the measured set, such as the visibility of the teeth, might play a role in determining perceptual distinctiveness of the vowels.

Accordingly, McGrath synthesized exemplars of the 11 monophthongal British–English vowels in 'b-vowel-b' syllables both with, and without, the teeth included in the synthesis. Without the teeth, subjects identified the vowels correctly on 51% of presentations. With the teeth, accuracy rose to 57% correct. Thus, the overall benefit of seeing the teeth was small, amounting to only 6%. However, the teeth played two important roles that led to larger improvements with particular vowels. First, they helped to distinguish close front vowels in syllables as 'beeb' and 'bib' from more open vowels in 'berb' and 'beb' which are articulated with the jaw lower and the teeth less visible. Second, the presence of the teeth helped subjects to distinguish rounded vowels in 'boob' and 'boub' (the vowel as in 'could') (where the teeth are not naturally visible) from unrounded vowels in 'beeb', 'berb' and 'barb' (where the teeth are usually more evident). Thus, overall, the experiment showed that subjects are sensitive to the visibility of teeth. They use it logically to distinguish vowels with otherwise similar lip shapes.

To calibrate the perceptual quality of the synthesizer, McGrath also ran conditions using a human talker. In two conditions, luminous lipstick and ultra-violet illumination were used to restrict the display to the same features as were animated in the synthesis. In these conditions, similar results were obtained. Performance was as good when only the talker's lips were visible (50% correct) or his lips and teeth were visible (56%) as in the analogous conditions using the synthetic face. Thus, the synthetic face conveyed the identities of vowels as accurately as might reasonably have been expected, given the data used to generate it and the restricted range of features that it included.

The limitations of the synthetic face were shown when the talker's face was presented naturally. Now, subjects identified 78% of the vowels correctly, significantly above the level achieved with the synthetic face, suggesting the importance of other features, such as the tongue and the wrinkling and protrusion of the lips, that are hard to convey with vector graphics and were excluded from the synthesis.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Other experiments also revealed the limitations of the synthesis. McGrath sought also to establish whether the synthetic face was sufficiently natural to be perceived 'in the speech mode'. In order words, did observers perceive syllables spoken by the synthetic face automatically, as they would with a natural face? Alternatively, did they intuit the intended vowel by a process of conscious analysis of the lip configuration? The difference is like that between an acoustic speech sound naturally 'naming itself' and a listener consciously hearing out the individual formants and working out what the intended speech sound must have been by an unnatural process of memory and matching.

To distinguish these alternatives, McGrath asked whether the synthetic face would cohere with a natural acoustical syllable and thereby give rise to audio-visual fusions of the type first described by McGurk & MacDonald (1976). Accordingly, he synthesized visual exemplars of 'ba' and 'ga' and made videorecordings of natural exemplars of the same two syllables. He synchronized each of the four visual tokens with natural acoustical exemplars of 'ba', 'ga', 'pa', and 'ka' and presented them to naive subjects. The instructions biased subjects against audio-visual integration. They were told to watch the face but to report only what they heard. The subjects were divided into two groups. One group ('natural–synthetic') saw the stimuli containing the natural visual tokens before the stimuli containing the synthetic visual tokens. The order was reversed for the other group. Both groups perceived fusions when the stimuli containing the natural visual tokens were presented. For example, they identified the combination of acoustical 'ba' with visual 'ga' as 'da', and acoustical 'pa' with visual 'ga' as 'ta'. Overall, such responses were made on about 75% of trials where incompatible pairings of visual and acoustical tokens were presented. The natural–synthetic group showed a similar pattern of responses when the stimuli containing the synthetic visual tokens were presented, although the overall proportion of fusion responses (44%) was somewhat lower. The synthetic–natural group, in comparison, did not produce fusion responses with the synthetic stimuli. Rather, they reported the acoustical syllables veridically. Control experiments established that prior experience of lipreading the synthetic face did not lead to fusions with the synthetic face. Rather, subjects needed prior experience of fusions generated by the natural face. McGrath hypothesized that exposure to stimuli incorporating the natural face helped to establish the slightly atypical phonetic categories specified by fusion percepts.

Overall, therefore, the results showed clear limitations to the phonetic acceptability of the synthetic face. However, from a less critical perspective, the performance of the synthesizer was impressive. It showed that audio-visual integration is not undermined if the talking face is represented schematically and palpably unnaturally by a few dozen vectors. The result is analogous to the demonstration that the occurrence of audio-visual fusions is not precluded by a mismatch of gender between the visible and audible talker (Green *et al.* 1991). In general, the fact that the image of a face is perceived simultaneously to be inappropriate to the accompanying acoustics, yet to provide phonetic information, is compatible with the idea that the identity of talkers and the speech they produce are analysed in physically and functionally separate parts of the brain (Campbell 1989; Ellis 1989).

Compared with the naturalness that could be achieved in a real-time display of a synthetic face in the early 1980s using vector graphics, modern hardware can realize techniques of model-based image coding and texture-mapping in real time to create startlingly life-like images of human faces (Aizawa *et al.* 1989; Waters & Terzopoulos, this symposium). Real-time analysis of an accompanying speech signal can be used to up-date the values of a limited set of acoustic parameters which can drive either a Hidden Markov Model or a neural net to generate a sequence of visible articulatory configurations. The face then seems to be producing the acoustical speech signal (Morishima *et al.* 1990; Welsh *et al.* 1990). These devices have potential applications in human-computer interfaces and videophony. A question for further research is to establish how far the articulatory movements of the synthesized face contribute to its naturalness, and how far they contribute to its intelligibility. Can an artificial face whose articulatory movements are generated from an analysis of the acoustic speech signal, convey useful information to a good lipreader in the absence of sound? More generally, can its lip movements improve intelligibility over sound alone for a severely or profoundly impaired listener?

5. AUDIO-VISUAL INTEGRATION

From an early age, human beings are predisposed to relate what they see and what they hear. Within the first few days of life, infants orientate visually towards an audible click (Wertheimer 1961). By 2.5–4 months, given a choice of two adjacent films, they look selectively at the one whose sound-track can be heard (Spelke 1976). By 4 months, given the choice of adjacent displays of a talker articulating different vowels, they look selectively at the vowel that can be heard (Kuhl & Meltzoff 1982). By 6 months, they prefer to look at the image of a talker who has the same gender as the voice in the sound-track (Walker-Andrews *et al.* 1991). By the age of 6 years, they have learnt some aspects of the audio-visual structure of the phonemes of their language (Massaro 1987). The knowledge is well-established in adulthood. Confronted with an audio-recording of one syllable synchronized with a video-recording of a different syllable, they perceive the syllable or syllables whose natural audio-visual appearance is most like the appearance of the artificial combination (McGurk & MacDonald 1976; Summerfield 1979, 1987*b*; Repp *et al.* 1983; Summerfield & McGrath 1984; Massaro 1987).

These demonstrations of audio-visual integration occur so impressively in the laboratory because there are huge advantages from integrating visible and audible evidence of speech in everyday life. How though do observers know that the articulations which they see and the speech which they hear are usually manifestations of the same event? Co-occurrence in

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

space and time is useful evidence but not conclusive. It is buttressed by similar patterning occurring in the two streams over time. For example, the size of the lip opening is one of the factors determining the rate of air-flow through the vocal tract and this in turn determines the overall intensity of the speech stream. In addition, an increase in the size of the opening raises the frequencies of each of the first three formants (Stevens & House 1955). Thus, in natural speech, the visible modulation of oral area is correlated with the acoustical modulation of overall amplitude and formant frequencies. Correlated modulation tells observers that the visual and acoustic signals originated in the same articulatory event and should be interpreted together. From an information-processing perspective, the questions that arise are how are the two streams of information represented at their conflux and what form does their conflux take?

It has seemed to me (Summerfield 1979, 1983, 1987*b*, 1991) that audio-visual integration must occur before phonetic categorisation takes place. Two pieces of evidence are particularly persuasive. First, lipreading is most useful in those natural situations where noise, reverberation, and hearing impairment, alone or in combination, make it difficult to categorise the acoustical speech stream phonetically. Critically, speech can be perceived audio-visually when the acoustical signal is replaced by a signal that cannot be categorized phonetically. These are signals such as a sequence of acoustic pulses synchronized to the moments when the talker closes her vocal folds (Rosen *et al.* 1981), or a single sinusoid whose amplitude is modulated according to the intensity of the speech signal in a band of frequencies centred on the frequency of the sinusoid (Breeuwer & Plomp 1984; Grant *et al.* 1991). Second, Green & Miller (1975) have shown that the rate of speech specified by a syllable presented visually can influence the perception of an accompanying acoustical consonant as voiced ('bi') or voiceless ('pi'). Here, a decision about the phonetic feature of voicing is made on the basis of evidence from both modalities. Integration, must necessarily take place before categorization, therefore. This conclusion has received support from two mathematical analyses. Massaro (1987) modelled subjects' identification responses to incompatible audio-visual combinations of consonants. Such responses were predicted more accurately by a model which preserved continuous values of auditory and visual features up to the point of integration, than by a model which made categorical decisions about the phoneme presented in each modality. Braida (1991) derived models from signal-detection theory which predict the confusions made between naturally produced audio-visual consonants from a knowledge of the pattern of confusions when the stimuli are presented separately in auditory and visual modes. More accurate predictions resulted if pre-categorical integration was assumed rather than post-categorical integration.

It is hard to be more specific about the form that the two streams possess at their conflux than to say that they are analogue and continuous. None the less, several possibilities can be outlined (Summerfield 1987*b*). For example, from the pragmatic orientation of engineering, there is no need for the two streams to be represented in a common metric and to 'flow together'. Knowledge of the audio-visual structure of phonemes and words can be represented in a computer program as sequences of linked auditory and visual templates. Pattern-recognition systems based on these principles hold promise for enhancing the performance of speech recognizers in noise (Petajan 1985; Nishida 1986). In contrast to this view, a phenomenologist would note that the products of audio-visual integration are auditory: Speech in noise sounds clearer when one can see the talker's face; the combination of an acoustical 'ba' with a visual 'ga' sounds like 'da' or 'dha'. Therefore, some aspect of the visual information must be converted into an auditory representation during audio-visual integration. Both views can be contrasted with the perspective of a direct realist (e.g. Fowler & Rosenblum 1991) where the production of speech is regarded as the production of appropriately coordinated articulatory gestures. The gestures modulate a sound stream to make themselves audible and modulate reflected light to make themselves visible. Subjects' knowledge of the gestures can be accessed by many different routes: audition and vision are but two. The objects of perception are the gestures, it matters little how they are sensed. It comes as no surprise, therefore, that naive subjects perceive 'auditory-tactile fusions' when an acoustical signal presented to their ears specifies one syllable but the face which their hand simultaneously feels utters a different syllable (Fowler & Dekle 1992). Thus, knowledge of the gestures is amodal. It cannot reside exclusively in audio-visual templates.

Each of these conceptions of the internal representation offers valuable ideas which will have to be reconciled in a comprehensive account of audio-visual speech perception.

6. CONCLUSIONS

Although much has been learnt about lipreading and audio-visual speech perception in the last 15 years, only incomplete answers can be given to the four questions with which this paper started. (i) We do not know what distinguishes better from poorer lipreaders. Several potential predictors of the ability to lipread, including intelligence and linguistic ability, can be ruled out. One measure, speed of low-level visual-neural processing, has been shown to correlate highly in some studies. This result reinforces the idea that lipreading is a particular skill divorced from other intellectual abilities. (ii) People are surprisingly tolerant of audio-visual asynchrony when perceiving speech. In group-mean data, sound must lag vision by more than 80 ms for audio-visual intelligibility to decline. The insensitivity should allow adequate time for signal-processing algorithms in future hearing aids to do their work. (iii) Perceptual experiments with relatively simple vector-graphics animations of talking faces have confirmed that subjects possess a detailed knowledge of the visible articulatory gestures that produce speech. Some computer-generated faces are now impressively natural. It is possible to animate their articulatory movements automatically by analys-

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

ing an accompanying acoustical speech signal. The next challenge is to regenerate the articulatory gestures sufficiently precisely for accurate lipreading to be possible. (iv) People find it natural and effortless to integrate the speech they see with the speech they hear. Integration occurs before speech is categorized phonetically, but it is not known how the auditory and visual streams are represented at their conflux.

Overall, therefore, research has confirmed that the ability to lipread a talking face is a useful and natural skill, but it remains a rather mysterious one.

I thank Mark Haggard and Vicki Bruce for comments on a draft of this paper, and Michael Whybray and Michael Brooke for advice about computer-based systems for transmitting, analysing, and animating talking faces.

REFERENCES

Aizawa, K., Harashima, H. & Saito, T. 1989 Model-based analysis image coding (MBASIC) system for a person's face. *Sign. Process.: Image Commun.* **1**, 139–152.

Binnie, C.A. 1977 Attitude changes following speech-reading training. *Scand. Audiol.* **6**, 13–19.

Braida, L. 1991 Crossmodal integration in the identification of consonant segments. *Q. Jl exp. Psychol.* **43A,** 647–677.

Breeuwer, M. & Plomp, R. 1984 Speechreading supplemented with frequency-selective amplitude-envelope information. *J. acoust. Soc. Am.* **76**, 686–691.

Brooke, N.M. 1991 Computer graphics animations of speech production. In *Advances in speech hearing and language processing*, vol. 2 (ed. W. A. Ainsworth). JAI Press. (In the press.)

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

Brooks, D.N. 1989 Guest Editorial: Lip-reading instruction and hearing-aid use. *Br. J. Audiol.* **23**, 275–278.

CCIR 1990 Tolerances for transmission time differences between the vision and sound components of a television signal. CMTT Document 1042-E presented to CCIR XVIIth Plenary Assembly, Dusseldorf, January 1990.

Campbell, R. 1989 Lipreading. In *Handbook of research on face processing* (ed. A. W. Young & H. D. Ellis). Amsterdam: North-Holland, pp. 187–205.

Cole, R.A. & Jakimik, J. 1980 A model of speech perception. In *Perception and production of fluent speech* (ed. R. A. Cole), pp. 133–163. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Conrad, R. 1977 Lip-reading by deaf and hearing children. *Br. J. educat. Psychol.* **47**, 60–65.

Dixon, N.F. & Spitz, L. 1980 The detection of audiovisual desynchrony. *Perception* **9**, 719–721.

Dodd, B. 1979 Lipreading in infants: attention to speech presented in- and out-of synchrony. *Cogn. Psychol.* **11**, 478–484.

Dodd, B., Plant, G. & Gregory, M. 1989 Teaching lipreading: The efficacy of lessons on video. *Br. J. Audiol.* **23**, 229–238.

Duquesnoy, A.J. 1983 The intelligibility of sentences in quiet and in noise in aged listeners. *J. acoust. Soc. Am.* **74**, 1136–1144.

Ellis, A.W. 1989 Neuro-cognitive processing of faces and voices. In *Handbook of research on face processing* (ed. A. W. Young & H. D. Ellis), pp. 207–216. Amsterdam: North-Holland.

Erber, N.P., Sachs, R.M. & DeFilippo, C.L. 1979 Optical synthesis of articulatory images for lipreading evaluation and instruction. In *Advances in prosthetic devices for the deaf: a technical workshop* (ed. D. L. MacPherson), pp. 228–231. Rochester, New York: NTID Press.

Farwell, R.M. 1976 Speechreading: a research review. *Am. Ann. Deaf* **121**, 19–30.

Fodor, J.A. 1983 *The modularity of mind*. Cambridge, Massachusetts: MIT Press.

Fourcin, A.J., Rosen, S.M., Moore, B.C.J., Douek, E.E., Clarke, G.P., Dodson, H. & Bannister, L.H. 1979 External electrical stimulation of the cochlea: clinical, psychophysical, speech-perceptual and histological findings. *Br. J. Audiol.* **13**, 85–107.

Fowler, C.A. & Rosenblum, L.D. 1991 The perception of phonetic gestures. In *Modulatory and the motor theory of speech perception.* (ed. I. G. Mattingly & M. Studdert-Kennedy), pp. 33–59. Hillsdale, New Jersey: Erlbaum Associates.

Fowler, C.A. & Dekle, D.J. 1991 Listening with eye and hand: Crossmodal contributions to speech perception. *J. exp. Psychol.: Hum. Percept. Perform.* **17,** 816–828.

Fromkin, V. 1964 Lip positions in American-English vowels. *Lang. Speech* **7**, 215–225.

Grant, K.W., Braida, L.D. & Renn, R.J. 1991 Single-band amplitude envelope cues as an aid to speechreading. *Q. Jl exp. Psychol.* **43A,** 621–645.

Green, K.P., Kuhl, P.K., Meltzoff, A.N. & Stevens, E.B. 1991 Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk Effect. *Percept. Psychophys.* (In the press.)

Green, K.P. & Miller, J.L. 1985 On the role of visual rate information in phonetic perception. *Percept. Psychophys.* **38**, 269–276.

Hirsh, I.J. & Sherrick, C.E. 1961 Perceived order in different sense modalities. *J. exp. Psychol.* **62**, 423–432.

Heider, F. & Heider, G. 1940 An experimental investigation of lipreading. *Psychol. Monogr.* **52**, 124–153.

Jeffers, J. & Barley, M. 1971 *Speechreading (lipreading).* Springfield, Illinois: Thomas.

King, A.J. & Palmer, A.R. 1985 Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Expl Brain Res.* **60**, 492–500.

Klatt, D.H. 1979 Speech perception: A model of acoustic-phonetic analysis and lexical access. *J. Phonet.* **7**, 279–312.

Klatt, D.H. 1987 Review of text-to-speech conversion for English. *J. acoust. Soc. Am.* **82**, 737–793.

Kuhl, P.K. & Meltzoff, A.N. 1982 The bimodal development of speech in infancy. *Science, Wash.* **218**, 1138–1141.

Lyxell, B. & Ronnberg, J. 1989 Information-processing skill and speech-reading. *Br J. Audiol.* **23**, 339–348.

MacLeod, A. & Summerfield, Q. 1987 Quantifying the contribution of vision to speech perception in noise. *Br. J. Audiol.* **21**, 131–141.

MacLeod, A. & Summerfield, Q. 1990 A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *Br. J. Audiol.* **24**, 29–43.

Marslen-Wilson, W.D. & Welsh, A. 1978 Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* **10**, 29–63.

Massaro, D. 1987 Speech perception by ear and eye. In *Hearing by eye: the psychology of lip-reading* (ed. B. Dodd & R. Campbell), pp. 53–83. London: Lawrence Erlbaum Associates.

McGrath, M. 1985 An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces. Ph.D. thesis, University of Nottingham.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

McGrath, M. & Summerfield, Q. 1983 Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J. acoust. Soc. Am.* **77**, 678–685.

McGurk, H. & MacDonald, J.W. 1976 Hearing lips and seeing voices. *Nature, Lond.* **264**, 126–130.

Middleweerd, M.J. & Plomp, R. 1987 The effect of speechreading on the speech-reception threshold of sentences in noise. *J. acoust. Soc. Am.* **82**, 2145–2147.

Montgomery, A.A. & Demorest, M.E. 1988 Issues and developments in the evaluation of speechreading. In *New reflections on speechreading* (ed. C. L. DeFilippo & D. G. Sims) (*The Volta Review* **90**), 193–214.

Montgomery, A.A. & Jackson, P.L. 1983 Physical characteristics of the lips underlying vowel lipreading performance. *J. acoust. Soc. Am.* **73**, 2134–2144.

Montgomery, A.A. & Soo Hoo, G. 1982 ANIMAT: a set of programs to generate, edit and display sequences of vector-based images. *Behav. Res. Meth. Instrument.* **14**, 39–40.

Morishima, S., Aizawa, K. & Harashima, H. 1990 A real-time facial action image synthesis system driven by speech and text. *SPIE Vis. Commun. Image Process.* **1360**, 1151–1158.

Nishida, S. 1986 Speech recognition enhancement by lip information. *Proceedings of CHI 86*, pp. 198–204.New York. Association for Computing Machinery.

Pandey, P.C., Kunov, H. & Abel, S.M. 1986 Disruptive effects of auditory signal delay on speech perception with lipreading. *J. Aud. Res.* **26**, 27–41.

Petajan, E.D. 1985 Automatic lipreading to enhance speech recognition. *IEEE CVPR'85*, 40–47.

Plant, G.L. & Macrae, J.H. 1981 The NAL lipreading test: development, standardization, and validation. *Aust. J. Audiol.* **3**, 49–57.

Plomp, R. & Mimpen, A.M. 1979 Improving the reliability of testing the speech-reception threshold for sentences. *Audiology* **18**, 43–52.

Reisberg, D., McLean, J. & Goldfield, A. 1987 Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. In *Hearing by eye: the psychology of lip-reading* (ed. B. Dodd & R. Campbell), pp. 97–113. London: Lawrence Erlbaum Associates.

Ronnberg, J., Arlinger, S., Lyxell, B. & Kinnefors, C. 1989 Visual evoked potentials: relation to adult speechreading and cognitive function. *J. Speech Hear. Res.* **32**, 725–735.

Rosen, S.M., Fourcin, A.J. & Moore, B.C.J. 1981 Voice pitch as an aid to lipreading. *Nature, Lond.* **291**, 150–152.

Samar, V.J. & Sims, D.G. 1983 Visual-evoked response correlates of speechreading performance in normal-hearing adults: a replication and factor-analytic extension. *J. Speech Hear. Res.* **26**, 2–9.

Samar, V.J. & Sims, D.G. 1984 Visual evoked-response components realtered to speechreading and spatial skills in hearing and hearing-impaired adults. *J. Speech Hear. Res.* **27**, 162–172.

Shepherd, D.C. 1982 Visual-neural correlate of speechreading ability in normal-hearing adults: Reliability. *J. Speech Hear. Res.* **25**, 521–527.

Shepherd, D.C., Delavergne, R.W., Frueh, F.X. & Clobridge, C. 1977 Visual-neural correlate of speechreading ability in normal-hearing adults. *J. Speech Hear. Res.* **20**, 752–765.

Spelke, E. 1976 Infants' intermodal perception of events. *Cogn. Psychol.* **8**, 553–560.

Stevens, K.N. & House, A.S. 1955 Development of a quantitative description of vowel articulation. *J. Acoust. Soc. Am.* **27**, 484–493.

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

Summerfield, Q. 1983 Audio-visual speech perception, lipreading, and artificial stimulation. In *Hearing science and hearing disorders* (ed. M. E. Lutman & M. P. Haggard), pp. 131–182. London: Academic Press.

Summerfield, Q. 1987*a* Speech perception in normal and impaired hearing. *Br. med. Bull.* **43**, 909–925.

Summerfield, Q. 1987*b* Some preliminaries to a comprehensive account of audio-visual speech perception. In *Hearing by eye: the psychology of lip-reading* (ed. B. Dodd & R. Campbell), pp. 3–51. London: Lawrence Erlbaum Associates.

Summerfield, Q. 1991 Visual perception of phonetic gestures. In *Modularity and the motor theory of speech perception* (ed. I. G. Mattingly & M. Studdert-Kennedy), pp. 117–138. Hillsdale, New Jersey: Erlbaum Associates.

Summerfield, Q., MacLeod, A.M., McGrath, M. & Brooke, N.M. 1989 Lips, teeth, and the benefits of lipreading. In *Handbook of research on face processing* (ed. A. W. Young & H. D. Ellis), pp. 223–233. Amsterdam: North-Holland.

Summerfield, Q. & McGrath, M. 1984 Detection and resolution of audio-visual incompatibility in the perception of vowels. *Q. J. exp. Psychol.* **36A**, 51–74.

Wertheimer, M. 1961 Psychomotor coordination of auditory and visual space at birth. *Science, Wash.* **134**, 1692.

Van Rooij, J.C.G.M., Plomp, R. & Orlebeke, J.F. 1989 Auditive and cognitive factors in speech perception in elderly listeners. I: Development of test battery. *J. Acoust. Soc. Am.* **86**, 1294–1309.

Walden, B.E., Montgomery, A.A. & Prosek, R.A. 1987 Perception of synthetic visual consonant-vowel articulations. *J. Speech Hear. Res.* **30**, 418–424.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K. & Jones, C.J. 1977 Effects of training on the visual recognition of consonants. *J. Speech Hear. Res.* **20**, 130–145.

Walker-Andrews, A.S., Bahrick, L.E., Raglioni, S.S. & Diaz, I. 1991 Infants bimodal perception of gender. *Ecol. Psychol.* **3**, 55–75.

Welsh, W.J., Simons, A.D., Hutchinson, R.A. & Searby, S. 1990 Synthetic face generation for enhancing a user interface. *Proceedings of Image-Com'90, International Conference on new Image Chains, Bordeaux*, France. pp. 177–182.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES