Proceedings of The Institute of Acoustics

USE OF OPTICAL INFORMATION IN SPEECH PERCEPTION

QUENTIN SUMMERFIELD

MRC INSTITUTE OF HEARING RESEARCH

Most sighted people possess some sensitivity to linguistic information carried by the facial concomitants of the articulation of speech. A description of this basic sensitivity and of the maximum available information would usefully constrain procedures for training the post-lingually deaf to supplement residual hearing with speech (lip)-reading. The ergonomic benefits of delimiting the information available in each modality would be enhanced by an understanding of the perceptual processes involved in the co-registration of the significant optical and acoustical information. The two experiments reported here represent a first step in exploring the nature of this co-registration.

The traditional term 'lip-reading' implies that hearing-impaired people read the lips. This prompts the question of what information is provided solely by the lips. Experiment 1 sought to answer this and to establish what degree of advantage accrues with visual information of various degrees of verisimilitude. The experiment used a realistic task in which speech must be understood against the background of an irrelevant speech stream. The advantage for full-face presentation over audio-alone would establish a standard with which performance in the other conditions of the experiment could be compared, and would also extend to a more natural situation earlier findings showing that facial information aids perception of low-pass filtered or noise-masked speech. [See Erber (1975).] The experiment was also intended to provide some leverage in dissociating two roles that optical information might play in speech perception under demanding listening conditions. Vision could parallel audition, both providing detailed phonetic information; alternatively, perceivers might obtain only a rough indication of syllabification through vision by observing the coupled movements of the lips and jaw, using these to direct attention selectively to phonetic detail in the accompanying acoustics, or as a rough source of prosodic information.

Experiment 1. Ten adults with normal hearing and vision transcribed test sentences (Fry, 1961) presented 12 dB below a background prose passage in one audio condition (Condition A) and in four conditions supplemented by monochrome videorecordings. In each condition, 25 test sentences containing a total of 100 scored words were presented. In Condition B:(Full Video) subjects viewed the face of the talker speaking the test sentences. In Condition C:(Lips) the talker had been made-up with luminous lip-stick and videorecorded under ultraviolet illumination. Only his lips could be seen. In Condition D:(Dots) lip movement was specified by four point sources located at the centres of the lips and the corners of the mouth. Finally, in Condition E:(Circle) the amplitude envelope of the test sentences modulated the diameter of a Lissajous circle displayed on the TV monitor.

The percentages of words correctly transcribed in each condition averaged over subjects and ranked in order of improving performance were: E:(Circle), 20.8%; A:(No Video), 22.7%; D:(Dots), 30.7% C:(Lips), 54.0% B:(Full Video), 65.3%. Every subject performed more accurately in Conditions B and C compared to Condition A, demonstrating that normal, untrained listeners can utilise optical concomitants of speech articulation in speech perception, that useful information can be obtained from the lips alone, and that a highly reduced display is not, per se, a bar to speech-reading. The benefits of these displays contrast with the minimal (8%) and non-significant improvement that resulted from the dots display in Condition D. Here only a partial specification of articulating lips was provided; the talker never seemed to close his mouth, suggesting that important information for speech-reading may derive from the changing area of the oral orifice. In contrast to these anatomically veridical displays, the circle in Condition E did not improve performance at all. With practice subjects would possibly benefit from this otherwise unfamiliar indication of syllabification. However, the difficulties of relating its patterns to the ongoing acoustics may stem not only from its unfamiliarity, but also from its lack of obvious articulatory underpinning. While patterns of lip movement offer a limited, but direct, indication of articulation, no similarly direct relation applies to an amplitude modulated circle. [See also Risberg and Lubker (1978).] In summary, Experiment 1 confirmed that untrained observers can benefit from viewing the face of the talker whose speech they must understand, and showed that a reduced, but significant, improvement also occurs when only the talker's lips are displayed.

Experiment 2. Experiment 2 further explored the sensitivity of untrained observers to optical specifications of information for phonetic perception. The procedure used in Experiment 1 of pairing natural speech with contrived displays was reversed; natural videorecordings were synchronised with synthetic speech syllables. The paradigm is a modification of that used by McGurk and MacDonald (1976).

Three 11-member continua of VCV syllables modelled on the speech of an English adult male were created with an OVEIIIb speech synthesiser. They varied in a triangular arrangement from [aba] to [ada], [ada] to [aga], and [aga] to [aba] using changes in the trajectories of the second and third formants. Randomisations were prepared in which the members of each continuum were synchronised equally often to videorecordings of the same adult male talker uttering one or other of their end-point syllables. Six adults with normal hearing and vision identified 20 instances of each syllable in the audio-visual condition just described and 10 instances in an audio-alone condition. The experiment sought to determine the extent to which optical information would bias the interpretation of acoustical information in favour of one or other continuum end-point, and the extent to which the information in the two modalities would combine to specify phonetic events not defined in either individual modality.

Clear differences between the two video conditions and between each of these and the no-video condition occurred for each continuum. The necessarily open response set led some subjects to identify the inherently more ambiguous stimuli in the centres of each continuum as the clusters [abda], [adga], and [abga], even in the no-video condition. Thus, interpretation of the data in terms of movements of phoneme boundaries is not straightforward. Rather, the results are summarised below as proportions of identifications in the major response categories averaged over the 11 members of each continuum.

Overall, two classes of effect occurred. First, stimuli ambiguous in the no-video conditions were assimilated into the response category of the phonetic event displayed visually. Every subject, with the [aba-ada] and [aga-aba] continua, and five of the six subjects, with the [ada-aga] continuum showed an increase in the proportion of responses corresponding to the phonetic event displayed visually. The second class of effect occurred only with the [aba-ada] and [aga-aba] continua. Here, the addition of visual information largely eliminated a response category. Independently of whether a bilabial was specified acoustically, a bilabial tended only to be perceived when lip closure was specified optically; and, in general, when lip closure was specified optically, a bilabial was perceived (often in a cluster), regardless of what consonant was specified acoustically. Principally, these effects reflect the visibility of the optical specifications of stop consonants uttered at different places of articulation: bilabial occlusion and release are fully displayed while more dorsal articulations achieve a less precise optical definition. As a result, an audio-visual display of a bilabial must entail optically specified bilabial closure and release. An audio-visual specification of an alveolar or velar, on the other hand, must entail optical specification of a non-labial articulation but it need not be specifically alveolar or velar, at least not for untrained observers. In summary, subjects behaved as if they appreciated the logical constraints that articulaton imposes on the audio-visual specification of phonetic events.

Acoustic Continuum	Video Display	%-Identifications averaged over six subjects [aba]	[ada]	[aga]	[abda]	[adga]	[abga]	Other
	[aba]	37.3	14.5	1.4	32.9	8.5	2.0	3.4
[aba-ada]	No Video	38.2	31.1	0.6	16.7	8.8	3.1	1.5
	[ada]	11.2	55.3	13.9	8.3	4.4	0.6	6.3
	[ada]	0.0	62.5	21.4	9.5	5.9	0.0	0.7
[ada-aga]	No Video	0.0	55.3	15.6	23.5	4.7	0.0	0.9
	[aga]	0.2	53.7	28.3	12.3	5.0	0.0	0.5
	[aga]	5.2	7.7	48.6	17.5	15.3	1.1	4.6
[aga-aba]	No Video	35.6	2.3	27.2	18.3	14.5	0.3	1.8
	[aba]	32.1	2.6	17.6	12.7	22.0	10.9	2.1

Discussion. Normally a talker's articulatory apparatus imposes structure on both light and sound, but the experience of watching and listening is of perceiving one speaker and one message. Explanations for bi-modal phonomena where percepts relate to the stimulus information in neither individual modality generally recognise that the information in the two modalities must be represented in a common metric for integration to occur. Phenomena of the type found in Experiment 2 then pose two questions. First, should the process of integration be viewed as a passive averaging or as an active process guided by an appreciation of articulatory constraints? Secondly, in the metric of integration, are phonetic events represented discretely as phonetic features or continuously in a form relating to articulatory dynamics?

Passive averaging can be ruled out. It predicts that audio-visual combinations of syllables such as [aba] and [ada] should either always, or never, yield percepts of clusters. The increase in clusters when the bilabial is specified optically, and their decrease when the bilabial is specified acoustically, demands a more subtle meshing of the information in the two modalities. The present data, in themselves, do not resolve the second question. Two independent pieces of evidence favour a continuous metric, however. First, current views of dichotic auditory integration in phonetic perception suggest that continuous rather than categorical (ie auditory rather than linguistic) descriptions of the information in each channel are combined (eg Repp, 1977). Secondly, distinctions of manner of production can be conveyed inter-modally in tactual-visual perception by variation in the relative timing of events in the two domains (Erber and DeFilipo, 1978). A demonstration that manner of articulation was jointly determined under more representative conditions by the auditory and the visual modality would serve further to enhance our understanding of how analogue information is combined in bi-modal phonetic perception. Since many of the phenemena of auditory speech perception can be rationalised by relating proximal acoustical stimuli to their origins in articulation (eg Liberman and Studdert-Kennedy, 1978), it is sometimes suggested that speech perception proceeds via the acoustical 'surface structure' to an appreciation of the 'deep structure' of the underlying articulatory dynamics -- a deep structure to which vision has only partial, but direct, access. In arguing that optics and acoustics are combined in a metric related to articulatory dynamics, we note, along with McGurk and MacDonald (1976), that the visual receptivity of observers to phonetic information demonstrates that phonetic perception is not solely the province of sound pressure variation and auditory analysis. The present results reinforce the distinction between the physical media which expound articulation and the dynamic patterns of articulation themselves. Potentially, it is these patterns and not the supporting media which are phonetically specific and phonemically relevant.

[A further account of these experiments appears in Summerfield (1979).]

References.

N.P. ERBER 1975 Journal of Speech and Hearing Research 40, 481-492. Audio-visual perception of speech.

N.P. ERBER and C.L. DEFILIPO 1978 Journal of the Acoustical Society of America 64, 1015-1019. Voice/mouth synthesis and tactual/visual perception of /pa,ma,ba/.

D.B. FRY 1961 The Lancet July 22, 197-199. Word and sentence lists for use in speech audiometry.

A.M. LIBERMAN and M. STUDDERT-KENNEDY 1978 In R Held, H.W. Leibowitz, and H.L. Teuber (Eds.), Handbook of Sensory Physiology, Vol VIII Perception, pp. 143-178. (New York: Springer-Verlag).

H. McGURK and J. MacDONALD 1976 Nature 264, 746-748. Hearing lips and seeing voices.

B.H. REPP 1977 Journal of Experimental Psychology: Human Perception and Performance 3, 37-50. Dichotic competition of speech sounds: the role of acoustic stimulus structure.

A. RISBERG and J. LUBKER 1978 Prosody and speechreading. Quarterly Progress Status Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm 1978:4, 1-16.

A.Q. SUMMERFIELD 1979 Phonetica (in press). Use of visual information in phonetic perception.