Proceedings of The Institute of Acoustics

AUDIO VISUAL PERCEPTION OF EVENTS AND SPEECH

D.W. BOSTON

ROYAL NATIONAL INSTITUTE FOR THE DEAF, LONDON

The issue is not so much as to whether there are a simple set of invariants relating speech to phonemes but whether hearing like vision has evolved to perceive events and whether such capacities or faculties play a part in speech perception.

Is Hearing Direct or Indirect?

In 1946 the phonetician Joos (1) wrote "The deaf lip reader perceives the speech signals directly and the rest of us indirectly through the ears". Joos stated explicitly the indirect view of hearing which is implicit in nearly all specialist publications on speech and hearing, in which hearing is mainly thought of as frequency analysis, rather than as the perception of events.

Like Joos, few would deny that the lip reader sees the visible part of the speakers articulation, yet many speech scientists deny that the listener hears the speakers articulation. This denial that articulation is perceived directly in hearing as in vision is implied not only in the motor theory, to which Joos subscribed and in which speech perception is thought to be based on the listeners knowledge as a speaker; but also in the now more widely held acoustical theories, in which the perception of articulation in the process of speech perception is denied, so that the term 'articulatory theory' has even come to imply the (indirect) motor theory.

Similarly the literature in psychophysics is still mainly preoccupied with continuous signals, or the location of events, but very rarely their identification, but what is the evolutionary selective value of location without identification? This indirect view of hearing has recently been challenged by Neisser (2) who stressed that speech perception is an audio visual process concerned with events in "someone's mouth" but he criticised the motor theory on the grounds that we do not need to dance in order to see the motion of the dancers. Neisser is the first to rediscover lip reading of which there has been practically no mention in the many articles on speech perception written in the 30 years,since Joos above, apart from literature on deafness. However the idea that hearing was concerned with events had been stressed in books on general psychology, thus Woodworth (3) asked "Do we hear sounds or do we hear objects". J.J. Gibson (4) suggested that sounds "mirror" the distinguishing features of the source. Earlier the audio visual approach had been implied by Sir Richard Paget (5) a physicist, who wrote "We lip read by ear" which has similarities to MacDonald and McGurk's (6) title "Hearing lips and seeing voices".

It should also be stressed that the question of whether articulatory information is available to the perceiver does not necessarily imply that articulatory patterns are more invariant than acoustical patterns.

Katherine Harris (7) writes that no transformation of the speech signal has been found, articulatory or otherwise that provides simple invariant units.

Perception of Events and evolution

In 1952 W.H. Huggins (8) whose contribution seemed to be in danger of being forgotten, wrote that in any communication system the optimum receiver needs to be matched to the transmitter and implied that animal hearing must have evolved to interpret meaningful events. In particular an important question with regard to such events was what (the excitation) was happening to what (the structure) and hearing must have evolved the capacity to separately identify these which would have an important role in speech perception. This subject is relevant to speech training aids for the deaf which have till recently been mainly concerned with the excitation, since it has been thought that the structure i.e. articulation can be seen or otherwise described.

Another point stressed by W.H. Huggins is that naturally occuring sounds are damped and hearing must have evolved to optimally identify damped sounds.

More recently the evolution of perceptual systems has been more widely studied but the relevance of excitation structure and damping to hearing seems to be unique to W.H. Huggins, but the perception of real events is often more complex as will be illustrated below.

Object and Event Perception

In our normal environment we see stable objects and structures despite the effect of our movement and of changes in lighting altering the patterns of light that fall upon our eyes. Artists have to learn to see these patterns (3b) and conversely Luria (9) describes how brain damage can cause a failure to convert these patterns into objects.

Most non speech research in hearing has been on continuous signals but a few contrasting examples mainly from work with musical notes will show how what we preceive is determined by the nature of the event.

1. Two complex harmonically related notes one low and one high, will fuse to form one note if they begin together but if one, the low begins as much as 10 millisecs before the other, two notes will be perceived. The notes will seem to begin simultaneously even if the starting delay of one note is increased to 30 millisecs. (10).

2. A single saw tooth note will be perceived either like a bowed or a plucked string according to the rate of onset.

2b. In the above example the rate of onset also influences the perceived timbre of the following note (11).

2c. It is well known that the character of the sound of several musical instruments including the piano is determined by the first 50 millisecs.

3. There are many examples in the perception of plosives in speech of event

identification, but the perception of silence as a clue to a following plosive is no doubt the most striking; the silence denotes a closure and therefore gives a clue to the articulation of the following plosive.

In the perception of speech plosives it has been demonstrated that the speech event we hear may be determined by visual clues. (6,12). It is interesting to consider when events are perceived as heard,and when seen. Speech events and musical events are generally perceived as heard even by the partially deaf, but spatial information is generally perceived as seen even when heard by the blind as in 'facial vision'.

Part of the reason why we do not think of hearing as being concerned with objects and events is because of the domination of hearing by vision. This does not occur with the blind who use hearing like we use vision. Thus a blind girl asked Dodd (12b)who has an Australian accent if she had a stone in her mouth.

Audio visual fusion and the ventriloquism effect

When an event has both audible and visual components fusion occurs. In real events as opposed to synthesized events the only error can be one of timing as when a film is out of synch. The effect of a mismatch in the visual aspects of speech has already been mentioned but the question of what is perceived when the wrong type of object or motion is synchronized with the sound would seem a useful area for research which will be referred to later.

When audible and visual aspects of an event are synchronized the sound may appear to come from the visual source even when it comes from a loudspeaker displaced from the visual source as on a television set. If the visual source is suddenly removed the apparent source of sound may still be displaced from the real source, due to an after effect. It has been reported (13) that when the TV image of a human speaker is shown a delay of 350 millisecs greatly reduces the probability of fusion when the sound is reproduced on a displaced loudspeaker. If a lamp modulated with speech is substituted for the face the probability of fusion is reduced, but the 350 millisec delay has now little effect so that fusion is more likely to occur with the delayed lamp than with the delayed face - but less likely to occur when there is no delay. It has been reported (14) that visual capture is independent of whether the language is understood.

Synthetic Audio Visual and the Relation of Vision and Sound

The evidence that the ventriloquism effect is more effective for a moving source such as lips and the evidence of audio visual fusion and lip reading, suggest that some information on motion is conveyed by the sound that can be related to the vision. The frequencies in the sound depend on the structure of the source, and the changing frequency to its motion, and therefore information is available for animal perception to relate sound to vision. The synthesis of moving visual stimuli which can have a controlled relationship to a sound modulation would be a useful tool for research into any such interactions. Ideally it would be desirable to control the time relationship or dealy between the sound and vision as well as the kind of motion. The synthesis of visual stimuli from the sound though useful is liable to be associated with some delay in the visual image. It would be preferable for the above research if the sound could be delayed

rather than the vision since this occurs naturally. The synthesis of synchronous visual information from sound has potential application for the deaf or speech disabled for communication and teaching.

Some system of speech analysis is necessary to control the visual image, Underwood using an analogue speech analyser (15) produced a side view of a mouth in 1970 on a computer display. Using a similar analyser without a computer I produced a front view of a mouth on an oscilloscope by controlling a lissajous figure (16). The gain of the X & Y inputs were controlled by the frequency of the formants F_2 and F_1 respectively which approximates the mouth shape for vowels. Erber (17) in the U.S.A. has described a modified version of my display, and claims that vowels may be read nearly as easily from his display as from a TV picture of a real mouth.

The synthetic mouth has been suggested as a communication aid for the deaf to use the telephone (16,17). In the U.K. the most promising applications so far appear to be for speech therapy for adults after strokes, and children with severe speech problems.

References

(1) JOOS M 1948. Acoustic Phonetics Supp. to Language Monograph 23.Page 62
(2) NEISSER U. Freeman and Co. San Francisco 1976. Cognition and Reality Chapter 8.
(3) WOODWORTH R.S. 1940.Psychology p.516. Methuen & Co. 3B page 481.
(4) GIBSON J.J. Allen and Unwin 1966. The senses considered as Perceptual Systems.
(5) PAGET SIR RICHARD 1930. Babel Kegan Paul. Page 27.
(6) McGURK H & MacDONALD J 1976. Nature Vol. 264.No.5588 746-748. Hearing Lips and Seeing Voices.
(7) HARRIS KATHERINE 1977. Haskins SR50 13-20. The study of articulatory organization, some negative progress.
(8) HUGGINS W.H. 1952. J.A.S.A. Nov.1952 582-589. A Phase principle of Complex Frequency Analysis.
(9) LURIA A.R. 1975. Penguin Books. p.37. The man with a Shattered World.
(10) RASH R.A. 1978. Acoustica Vol. 40 21-33. The perception of simultaneous notes as in polyphonic music.
(11) CUTTING J.F. & ROSNER B.S.1976.Quarterly J.Exp.Psych.28,361-378.Perceptual categories for music like sounds.Implications for theories of speech perception.
(12) DODD BARBARA 1977. Perception Vol.6 p31-40.The role of Vision in the Perception of Speech. 12b. Personal Communication.
(13) RADEW M & BERTELSON P. 1977. Perception & Psycho physics 1977. Vol. 2 47-53. Adaptation to Auditory Visual discordance and Ventriloquism.
(14) THURLOW W.R. & JACK C.E. 1973.Perceptual & Motor Skills 36, 1171-1184 Certain Determinants of the"Ventriloquism Effect".
(15) UNDERWOOD M. ADDIS BOSTON 1972. Physics Conf. series No. 13 117-125 The evaluation of certain parameters for automatic recognition of spoken words.
(16) BOSTON 1973. British Journal of Audiology. 7,95-101.Synthetic Facial Communication.
(17) Erber N. 1977 J.A.S.A. 62. Supp.No.1 S 76. Real time synthesis of optical lip patterns from vowel sounds.