Handbook of Research on Face Processing
A.W. Young and H.D. Ellis (Editors)

READING GESTURES BY LIGHT AND SOUND

Michael Studdert-Kennedy

University of Connecticut, Yale University and
Haskins Laboratories, New Haven, Connecticut

Let me begin where Ruth Campbell ends, with the double dissociation between site of lesion (left/right hemisphere) and mode of facial processing (lipreading/recognition of facial identity and expression). This important finding has a parallel in the first systematic studies of aphasia in American Sign Language (ASL) (Poizner, Klima & Bellugi, 1987). At issue in this work was whether ASL, given its elaborate visuo-spatial structure, would lateralise to the right hemisphere, or, given its formational ('phonological') and syntactic status as a language, independent of spoken language, to the left hemisphere. The answer came from patients, all native signers of ASL, with different lesions and correspondingly different patterns of impairment. Patients with right hemisphere lesions displayed normal perception and production of ASL, but were severely impaired on standard tests of visuo-spatial function. Patients with left hemisphere lesions performed normally on the standard spatial tests, but were severely impaired in the perception and production of ASL. Thus, sign language, in a fashion at present well beyond our grasp, seems to rest on some of the same neurophysiological underpinnings as spoken language. This discovery, together with much other recent work on ASL and other sign languages (e.g. Klima & Bellugi, 1979; Stokoe & Volterra, 1985), places a severe, rarely acknowledged constraint on theories of language universals: any adequate theory will have to be formulated in terms that capture the common properties of both signed and spoken language.

By the same token, the dissociation of lipreading from facial recognition according to site of lesion, and certain other recent discoveries in lipreading, place sharp constraints on theories of speech perception: they force us to formulate accounts of speech perception, and of hemispheric specialisation for speech perception, in terms that capture the common properties of both lipread and heard speech. As Summerfield (1987) has remarked: '... any comprehensive account of how speech is perceived should encompass audio-visual speech perception. The ability to see as well as to hear has to be integral to the design, not merely a retro-fitted afterthought' (p.47). Of course, the challenge from lipreading is very much less than from sign languages, because the optic signal that we read from lips arises from exactly the same physical source as the acoustic signal that we hear. That is why lipreading bears more directly on speech perception than does the reading of print.

The most important recent work on lipreading stems, in my view, from the discovery of auditory-visual interference effects in both perception (McGurk & MacDonald, 1976) and short term memory (Spoehr & Corin, 1978; Campbell & Dodd, 1980). Unfortunately, Campbell chooses not to discuss the latter work (to which she has been a notable contributor in many papers), perhaps because its details are complicated, and often difficult to interpret. But the broad message of both bodies of work is clear.

Interference effects occur at a relatively low level of the perceptual process, and they arise because the auditory and visual forms of speech are structurally identical: they share a common metric (Summerfield, 1979).

Despite the common metric, optic structure is in no way intrinsic to, or necessary for, the linguistic function of speech. Perhaps this should be obvious enough from the simple fact that blind children learn to speak normally. Yet Campbell quotes Miller and Nicely's (1955) observation 'Lipreading ... provides just the information that ... noise or deafness removes', and comments: 'Is this a lucky biological accident? If not, it may be that spoken language has developed [i.e. evolved] in an essentially bimodal manner'.

If selection pressures had indeed shaped spoken language into forms accessible to both eye and ear, we would expect languages to avoid phonetic contrasts that cannot be seen, and to prefer contrasts that can. Yet, as Campbell herself remarks of English: '... the number and type of speech sounds that can be seen are few'. And Maddieson's (1984) survey of the phonemic inventories of a systematic sample of 317 of the world's languages does not seem to support the predicted biases. In fact, the amount of information to be gathered from the lips probably varies from language to language, and is therefore not a biological constant. For example, while liprounding distinguishes back vowels from front vowels in English, it does not do so in languages, such as French and Swedish, that have both front and back rounded vowels. Even within a language, dialects may vary in what they offer to the eye. Thus, the dental fricative of the English definite article (of potential syntactic value to lipreaders, as an index of many noun phrases), though interdental and clearly visible in some dialects, is more difficult to see in others, where it is executed with an apical articulation behind the upper front teeth, scarcely different in placement from alveolar /d/ (Ladefoged & Maddieson, 1986).

In short, speech has evolved to be heard, not seen - just as sign languages have evolved to be seen, not heard. To the arguments sketched above readers may add their own intuitions by whispering the word, east, for example. Here, acoustically sharp contrasts among a high front vowel, a palatal fricative and an alveolar stop are accomplished by lingual traverse across a few millimeters of vocal tract space, invisible to the lipreader, though detectable on a high resolution X-ray. In all languages, consonantal contrasts carry a heavier functional load than vowel contrasts and exploit similarly small shifts in the degree and placement of an intraoral constriction to yield distinctive acoustic contrasts. The adaptation of ASL signs to vision is illustrated by Siple (1978). She shows that signs executed within the foveal focus of a sign language viewer exploit finer contrasts of hand shape and movement than signs executed outside the foveal focus. Thus, the phonetic forms of both language modes are shaped by the modalities to which they are addressed. Since, then, the optic properties of speech are purely coincidental to its linguistic function, we must look for an understanding of lipreading outside the linguistic system, in more general aspects of perceptual function.

From an ethological perspective the function of perception is to control action. We negotiate the physical world by adapting our actions to the structure of objects and events that we perceive. The sensory modality specifying that structure is a matter of indifference: we jump from the path of a moving car whether we see it or hear it. Some events are accessible by only one modality, but many, perhaps most, are accessible by more than one, and we would be puzzled if the expected correlates were missing: a glass shattering in silence, the thud of an invisible book falling from a desk.

Of course, we would not be puzzled by the thud of an invisible book if the book fell in a completely dark room, because we know that the optic properties of a falling book are not intrinsic to the event: they are only available if there is an extrinsic source of light to be reflected from the book. Other objects or events have no intrinsic acoustic properties, but can be heard if we supply an extrinsic source of energy. Such events may be unfamiliar to many of us who are sighted, but are commonplace for the blind who tap their sticks on the pavement and listen for the sound reflected from obstacles in their path. They are also commonplace for bats or dolphins who feed by 'illuminating' their prey with radiated sound, and for, say, flute players who may practice in silence before making the resonances of the flute audible by exciting its column of air with their breath.

The application of these simple principles to speech is obvious. If we articulate a sentence in a lighted room without setting the vocal cords into vibration, viewers may pick up some of what we say by watching our lips,. but they will lose most of it, because they cannot see our articulations. (Even if our faces and teeth were transparent, viewers would find the task difficult because, as already noted, articulatory movements are delicate and have evolved to be heard, not seen.) Of course, as soon as we blow air through the glottis, as in whispering, or set the cords into vibration, as in voiced speech, we excite the changing resonances of the column of air in the vocal tract, so that our articulations become audible and our speech intelligible.

Notice, incidentally, that we can apply the same principles to sign language. Toward the end of the 18th century, an Austro-Hungarian aristocrat, named Wolfgang von Kempelen, constructed a primitive speech synthesiser (Flanagan, 1972). The device consisted, in part, of an energy source (a pair of bellows) which set a reed into vibration, and a flexible leather resonating chamber. Von Kempelen formed different vowels by modulating the form of the leather chamber with his hand. Thus, an attentive viewer might have seen the vowels as well as heard them. By the same token, we might, at least in principle, construct an acoustic device sensitive to the hand shapes and movements of a signer. We could then hear the signs as well as see them.

In short, there is nothing intrinsically acoustic about speech, or intrinsically optic about sign language. Certainly each has evolved to exploit its characteristic medium, and therefore has properties peculiar to that medium. By 'a lucky biological accident' speech can in some measure be seen as well as heard.

What is it, in fact, that we hear? Do we hear the movements of the articulators, or do we hear the shapes and volumes of the resonating cavities that these movements configure? And which of these does the lip reader perceive? Campbell reports Summerfield's (1987) 'usefully provocative idea' that lipreaders may 'compute' the vocal tract filter function (that is, the shapes and volumes of the vocal tract cavities). But she does not mention his equally provocative, and no less useful, alternative proposal that lipreaders may perceive the kinematic surface structure (and hence the dynamic deep structure) of modality-free articulatory movements. The omission is surprising because it is, after all, precisely the movements that lipreaders try to see and, if deaf, to imitate.

We can resolve (perhaps dissolve) the choice between cavities and movements by recalling the blind child learning to speak. How does the child do this? The only information (that is, the only structure) available to the child is a time-varying acoustic signal isomorphic with the changing

configurations of the cords and the tract. A neural description of the signal in these terms will evidently suffice for recognition of speech, as when a victim of congenital cerebral palsy, unable to speak, nonetheless learns to understand speech and even to read and write (Fourcin, 1975). But for the blind child who learns to speak, mere recognition is not enough. The signal must also specify the movements to be made for its reproduction. Some blind children display patterns of error in speaking that are seldom observed in sighted children, such as confusions between /m/ and /n/, and the substitution of /0/ for /f/. These errors, evidently reflecting the lack of visual information, are transitory and have typically disappeared by the third year of life (Mills, 1987). It is difficult to avoid the conclusion that the blind child, like every normal child, comes into possession, through early postnatal development, of a specialised link between sound and articulation. One piece of the link is a neural model of the speech apparatus, and of the way its several parts may be coordinated to modulate its configuration over time.

What, then, of the deaf child learning to speak? Since, as argued above, the acoustic and optic signals are simply radiations of sound energy or reflections of light energy, from the same articulatory events, deaf children do not need '... detailed acoustic experience on which to base their analysis of seen speech'. Their task in learning to speak is to superimpose on their vocalisations (which will have developed almost normally during the first six months of life) the articulatory patterns given to them by eye. They often do not succeed in this - in either speaking or understanding speech - as well as normal children, because the optic structure is fragmentary. They are striving to interpret a facial sign language, as it were, of which the gestures are partially obscured by an arbitrary grid.

I should emphasise that I am not proposing a 'motor theory' of speech perception. The assumption that the brain of an animal contains neural structures isomorphic (perhaps after some transformation) with its body, and with its possible modes of action, is an axiom without which behaviour would be unintelligible to the neurophysiologist. The further assumption that animals able to imitate the vocalisations of their conspecifics - an ability confined to a few species of songbirds and marine mammals, and to man - have a specialised neural link between the processes of perception and production is also hardly controversial. Thus, I am simply making, for speech, the common sense assumption we must make for facial expression, manual gesture, gait, and every other mode of action we can imitate, namely, that the signal induces a perceptual structure isomorphic with the action that produced it. The structure is then a modality-free pattern of pieces, and their relations, corresponding to the coordinated pattern of pieces - movements, gestures - by which the action was executed (cf. Studdert-Kennedy, 1987).

On this account, then, lateralisation of lipreading to the language hemisphere is merely a different aspect of the well-established lateralisation of auditory speech perception to that hemisphere. Speech motor control is vested in the left hemisphere, perhaps because speech evolved by duplicating and exploiting neural structures that had originally evolved for right-handed manipulation and bimanual coordination (MacNeilage, Studdert-Kennedy and Lindblom, 1984). The neural organisation required for speech perception may then have been drawn to the same locus in the course of the evolution of a capacity for vocal imitation.

Finally, I cannot refrain from remarking that, if '... the theories so far offered to account for lipreading are underelaborated', they are scarcely more so than our theories of speech perception. In fact, we have

no well-developed theories of speech perception, susceptible of systematic testing and elaboration. One reason for this is that theorists have not been able to agree on a set of perceptual primitives. An approach to lipreading research of the kind sketched above may help to supply this lack by encouraging attention to the coordinated patterns of gesture that the speech signal conveys (cf. Browman & Goldstein, 1986). As long as we have no adequate theory of speech production we are unlikely to have an adequate theory of speech perception.

Acknowledgement. Preparation of this paper was supported in part by the National Institutes of Health. Grant HD-01994 to Haskins Laboratories.

REFERENCES

Browman, C. and Goldstein, L. (1986) Towards an articulatory phonology. Phonology Yearbook, 3, 219-252

Campbell, R. and Dodd, B. (1980) Hearing by eye. Quarterly Journal of Experimental Psychology, 32, 85-99

Flanagan, J. L. (1972) The synthesis of speech. Scientific American, February 1972, p.52

Fourcin, A. J. (1975) Language development in the absence of expressive speech. In E. H. Lenneberg and E. Lenneberg, (Eds). Foundations of Language and Development, Vol. 2. New York: Academic Press, 263-277

Klima, E. S. and Bellugi, U. (1979) The Signs of Language. Cambridge, Mass.: Harvard University Press

Ladefoged, P. and Maddieson, I. (1986) Some of the sounds of the world's languages: Preliminary version. University of California at Los Angeles: Working Papers in Phonetics, 64 (entire issue)

MacNeilage, P., Studdert-Kennedy, M. and Lindblom, B. (1984). Functional precursors to language and its lateralization. American Journal of Psychology, Vol. 246 (Regulatory Integrative and Comparative Physiology, Vol. 15) R912-14

McGurk, H. and MacDonald, J. (1986) Hearing lips and seeing voices. Nature, 264, 746-748

Maddieson, I. (1984) Patterns of Sounds. Cambridge: Cambridge University Press

Miller, G. A. and Nicely, P. (1955) An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America, 27, 338-352

Mills, A. E. (1987) The development of phonology in the blind child. In B. Dodd and R. Campbell (Eds). Hearing by Eye. London: Lawrence Erlbaum Associates, 145-161

Poizner, H., Klima, E. S., and Bellugi, U. (1987) What the Hands Reveal About the Brain. Cambridge, Mass.: MIT Press

Siple, P. (1978) Visual constraints for sign language communication. Sign Language Studies, 19, 15-24

Spoehr, K. T. and Corin, W. S. (1978) The stimulus suffix as a memory code phenomenon. Memory and Cognition, 6, 583-589

Stokoe, W. and Volterra, V. (1985) SLR '83. Rome, Italy: Istituto di Psicologia, CNR

Studdert-Kennedy, M. (1987) The phoneme as a perceptuomotor structure. In A. Allport, D. MacKay, W. Prinz and E. Scheerer (Eds). Language Perception and Production. London: Academic Press, 67-84

Summerfield, A. Q. (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds). Hearing by Eye. London: Lawrence Erlbaum Associates, 3-51

Summerfield, A. Q. (1979). use of visual information for phonetic perception. Phonetica, 36, 314-331