861

SOME THEORETICAL IMPLICATIONS OF CROSS-MODAL RESEARCH IN SPEECH PERCEPTION

MICHAEL STUDDERT-KENNEDY
Haskins Laboratories
270 Crown Street
New Haven, Connecticut 06511-6695
USA

1. Preliminary

Both face and voice carry information about an individual's identity and emotional state. But the visual and auditory channels conveying this information are largely independent. Apart from a speaker's sex and age, we cannot reliably pair the identities of face and voice; and the observation that a speaker's face and voice may express quite different emotions is commonplace. Not surprisingly, then, studies of voice and face recognition, or of vocal and facial affect, are typically carried out by different people in different laboratories. Lipreading, by contrast, is typically studied by people who also study speech perception. The reason for this is simply that the two signals, optic and acoustic, that carry the phonetic message, are not independent: they both arise from the same physical source, the speaker's articulations.

Studies of lipreading over the past fifteen years have taken on a new (though not yet widely recognized) theoretical importance in speech research. Certainly, speech has evolved to be heard, not seen: indeed, we can reliably apprehend relatively few phonetic structures by eye, because most of a speaker's articulatory maneuvers are concealed from view. But the fact that we can visually apprehend at least some phonetic structure, with fair reliability, demonstrates that speech is not purely auditory. More than this, the fact that we can integrate optic and acoustic information precategorically, so as to arrive at a categorical phonetic percept that we could not have achieved from either channel alone, demonstrates the formal correspondence, or isomorphism, of the two sources. Evidently, the perceptual primitives of speech are not the static entities—consonants, vowels, features—of standard linguistic description, but dynamic structures corresponding to a talker's gestures (cf. Browman & Goldstein, 1990).

What follows is a summary account of selected studies of cross-modal speech perception in adults and infants. Adult studies have been largely directed to exploring the nature of the cross-modal interaction. Infant studies have been directed both to establishing that infants are sensitive to correspondences between sound and gesture and to understanding the perceptual basis of their imitative responses.

B. de Boysson-Bardies et al. (eds.),
Developmental Neurocognition: Speech and Face Processing in the First Year of Life 461–466.

2. Studies Of Cross-Modal Speech Perception

Neuropsychological studies suggest that the capacities to perceive speech by ear and by eye can be dissociated (Campbell, Landis & Regard, 1986; Ellis, 1989). This fact demonstrates that the two processes are, in principle, redundant. Under certain conditions, however, when neither process is fully adequate, they may be complementary, the eye supplying what the ear lacks, and vice versa (Summerfield, 1987). It then becomes a matter of interest whether information from the two channels is combined additively after some hypothetical process of "phonetic feature" extraction, or is integrated into a continuous time-varying, precategorical structure. Several diverse experimental paradigms have yielded evidence that, under at least some circumstances, the latter is the case.

2.1. ADULTS

2.1.1. *Fundamental Frequency As An Aid to Lipreading.* A stringent test of the possibility that auditory and visual information can be integrated precategorically is provided by situations in which one or other signal, presented alone, cannot be understood at all. An example comes from the combination of the talker's face with a synchronized pulse train picked up from the talker's larynx. Alone, the pulse train carries no segmental information, only the talker's fundamental frequency, conveying intonation, stress and the timing of voice onset and offset. Yet it can appreciably facilitate lipreading.

In a test of the speed with which subjects can track passages of connected discourse (repeating a talker's words verbatim), Rosen, Fourcin & Moore (1981) found that the addition of fundamental frequency to the sight of lips alone increased the rate of correct repetition in words/minute by an average of 83% for 5 subjects. Interestingly, experienced observers "...report a surprisingly complete degree of integration. Subjectively, the pulse train ceases to sound like a buzz; it acquires vowel color and other acoustical attributes" (Summerfield, 1987, p.16, fn.3). Similar impressions of observers' actually hearing what they have seen (and may even know themselves to have seen) are reported for the well-known "McGurk effect" (McGurk & MacDonald, 1976).

2.1.2. *The Effect Of Seen Changes In Speech Rate On Auditorily Specified Phonetic Percepts.* A second example reverses the role of sight and sound in the previous example: a phonetically ambiguous optic signal is combined with a phonetically clear acoustic signal. Green & Miller (1985) used a cross-splicing technique to construct three natural speech voice onset time (VOT) continua, ranging from /bi/ to /pi/. The continua differed in overall syllable duration, mimicking differences in speech rate; the syllables within a continuum differed in the duration of the aspiration preceding voice onset. The authors also prepared video tapes of a speaker uttering /bi/ and /pi/ at rates reliably judged to be "fast" or "slow", and determined that these syllables, when lipread, were completely ambiguous as to their voicing status. Finally, having established that the phoneme boundaries along the auditory continua varied as a function of their auditorily specified rates, the authors dubbed fast and slow video tokens onto the moderate rate auditory continuum, and tested observers for a possible effect of visually specified rate on the phoneme boundary. The result was a small, but significant effect of visual rate on the auditorily based phonetic judgements.

2.1.3. *Listening By Touch.* A final illustration of precategorical, cross-modal integration of continuous speech information exploits the McGurk effect (McGurk & MacDonald, 1976), with a novel twist. In the standard McGurk procedure subjects watch a video of a subject uttering, for example, the sequence of syllables, /ba, va, ða, da/, while hearing the synchronized auditory sequence, /ba, ba, ba, ba/. Subjects typically have the compelling experience of hearing the syllables that they see.

Fowler & Dekle (1991), in an attempt to eliminate the possible effects of experience with audiovisual speech, tested for a haptic McGurk effect. Subjects listened to syllables randomly drawn from a synthetic /ba/ - /ga/ continuum, while simultaneously holding their index finger against the upper lip, their second finger against the lower lip, of a speaker who was silently mouthing either /ba/ or /ga/ in synchrony with the auditorily presented syllables. Subjects were asked to indicate on each trial both what they heard and what they felt (with their fingers). In a second condition of the experiment, subjects watched a video screen on which the printed syllables BA or GA were flashed in synchrony with the synthetic acoustic syllables, and subjects were now asked to say both what they heard and what they saw.

If auditory and visual information were combined categorically, we would expect some interference between read and heard syllables. In the event, there was none. By contrast, there was systematic, mutual acoustic-haptic interference, such that the phoneme boundary significantly shifted as a function of the felt syllable, and judgements of the felt syllable significantly shifted as a function of the heard syllable's position on the continuum.

2.2 INFANTS

2.2.1. *Perceptual Preference Studies.* Perhaps the earliest work is that of Dodd (1979) who showed that 4-month-old infants watched the face of a woman reading nursery rhymes more attentively if her voice was synchronized with her facial movements than if it was delayed by 400 ms. Synchrony alone is not enough, however, to elicit a preference: infants also require structural correspondence between acoustic and optic signals. Kuhl & Meltzoff (1982, 1984) showed that 4-5 month old infants looked longer at the face of a woman synchronously articulating the vowel they were hearing (either [i] or [a]) than at the same face synchronously articulating the other vowel. Moreover, when the acoustic signals synchronized with the woman's movements were pairs of pure tones centered at the woman's fundamental frequency (200 Hz) and matched in amplitude envelope over time, duration, and temporal alignment to the original vowels, the preference disappeared. Evidently, it was a match between a mouth shape and a particular spectral structure that the infants wanted to see.

Walton & Bower (in press) replicated this finding for the vowels /a/ and /u/ in an operant conditioning study of 4 1/2 month old infants. The infants learned to control presentation of the vowel sounds, paired with visual presentation of either a matched or a mismatched facial gesture, by sucking on a non-nutritive nipple. They then sucked to call up matched pairs significantly more often than they did to call up mismatched pairs. In a second experiment, these investigators asked whether infants prefer a match because it is familiar or because, unlike a mismatch, it is articulatorily "possible" (or natural). In the same operant conditioning paradigm, they tested 6-8 month old infants, growing up in Texas, by presenting a single facial gesture, the rounded lips appropriate for both English /u/ and French /y/, paired with one of three sounds, differing in their presumed familiarity to the infants: English /u/, English /i/, or French /y/; they also presented the three sounds alone without the visual gesture. The infants displayed no preference among the sounds presented without the gesture, but significantly preferred the

articulatorily matched pairs (familiar English /u/ or unfamiliar French /y/ with rounded lips) to the mismatched pair (familiar English /i/ with rounded lips). Evidently, it is the physical correspondence between lips and sound, not their familiarity, that infants prefer.

Preliminary evidence that infant capacity to recognize acoustic - optic correspondences in speech is a left hemisphere function comes from a study by MacKain, Studdert-Kennedy, Spieker & Stern (1983). These investigators showed that 5-6 month old infants looked significantly longer at the face of a woman repeating a disyllable they were hearing (e.g. /zuzi/) than at the synchronized face of the same woman repeating another disyllable (e.g. /vava/) -- but only when they were looking to their right sides. Fourteen of the eighteen infants in the study preferred more matches on their right sides than on their left. In a follow-up investigation of familial handedness, MacKain and her colleagues learned that six of the infants had left-handed first or second order relatives. Of these six, four were the infants who preferred more left-side than right-side matches.

These results can be understood in the light of studies by Kinsbourne and his colleagues. Kinsbourne (1972) found that right-handed adults tended to shift their gaze to the right, while solving verbal problems, to the left, while visualizing spatial relations; left-handers tended to shift gaze in the same direction for both types of task, with each direction roughly equally represented across the subject group. Lempert and Kinsbourne (1982) showed that the effect was reversible for right-handed subjects on a verbal task: Subjects who rehearsed sentences, with head and eyes turned right, recalled the sentences better than subjects who rehearsed, while turned left. Thus, attention to one side of the body may facilitate processes for which the contralateral hemisphere is specialized.

Extending this interpretation to the infants of MacKain et al. (1983), we may infer that infants with a preference for matches on the right side, rather than the left, were revealing a left hemisphere capacity for recognizing acoustic-optic correspondences in speech. If, further, the metric specifying these correspondences is the same as that specifying the auditory-motor correspondences necessary for imitation (as might reasonably be assumed), we may conclude that 5- to 6-month-old infants already possess a speech perceptuo-motor link in the left hemisphere.

2.2.2. *Imitation Studies.* As an incidental finding of their study of infant perceptual preference, cited above, Kuhl & Meltzoff (1982) reported that 10 of their 32 4-5 month old infants "...produced sounds that resembled the adult female's vowels. They seemed to be imitating the female talker, 'taking turns' by alternating their vocalizations with hers" (p.1140). Such imitations are never, so far as I know, reported for studies of unimodal, auditory speech perception by infants; nor did the infants of Kuhl & Meltzoff (1982) vocalize when the sounds paired with the woman's face were pure tone controls. Nonetheless, since Kuhl & Meltzoff did not vary acoustic and optic displays independently, we cannot be sure whether the infants were imitating the sound, the mouth movements, or both.

Legerstee (1990) addressed this question for the vowels /a/ and /u/ in 3-4 month old infants. She elicited both vocal and purely motor imitations by presenting matched and mismatched acoustic-optic pairs. Infants produced significantly more /a/ sounds when auditory /a/ was presented than when auditory /u/ was presented, and significantly more /a/ sounds when it was matched than when it was mismatched with the articulating face; the same, *mutatis mutandis*, for /u/. Scoring the infants' mouth movements (wide open for /a/, pursed open for /u/), with or without concomitant vocalization, yielded a higher overall probability of imitation with

essentially the same pattern of results. We can conclude that, at the age of 3-4 months, the combination of visual with auditory information facilitates an imitative response, whether vocal or purely gestural. The visual component is not necessary, however, since blind children learn to talk with minimal delay in phonological development (Mills, 1987; Mulford, 1988); nor is the visual component sufficient, since deaf children have notable difficulties in learning to talk.

3. Conclusions

Under appropriate conditions observers integrate acoustic, optic and haptic patterns into a unified, precategorical phonetic form. The evident isomorphism of the three modalities has its origin in a common source, the speaker's articulatory gestures. By adopting the gesture as a perceptual primitive we ground the infant's early phonological development in its prelinguistic capacities for facial and vocal imitation. Such capacities may well have evolved, at least in part, under selection pressures for speech, but are not in themselves linguistic. Thus, we are absolved from the tautology of deriving a property of language from a supposed linguistic capacity.

Acknowledgements

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

4.0 References

Browman, C. P. and Goldstein, L. (1990) 'Gestural specification using dynamically-defined articulatory structures', Journal of Phonetics 18, 299-320.

Campbell, R., Landis, T., and Regard, M. (1986). 'Face recognition and lipreading: A neurological dissociation', Brain, 109, 509-521.

Dodd, B. (1979) 'Lipreading in infants: Attention to speech presented in- and out-of-synchrony', Cognitive Psychology 11, 478-484.

Ellis, A. W. (1989) 'Neurocognitive processing of faces and voices', in A. W. Young and H. D. Ellis (eds.), Handbook of Research on Face Processing, North-Holland Publishers, Amsterdam, pp. 207-215.

Fowler, C. A. and Dekle, D. J. (1991) 'Listening with eye and hand: Cross-modal contributions to speech perception', Journal of Experimental Psychology: Human Perception and Performance 17, 816-828.

Green, K. P. and Miller, J. L. (1985) 'On the role of visual rate information in phonetic perception', Perception & Psychophysics 38, 269-276.

Kinsbourne, M. (1972) 'Eye and head turning indicates cerebral lateralization', Science 176, 539-541.

Kuhl, P. K. and Meltzoff, A. N. (1982) 'The bimodal perception of speech in infancy', Science 218, 1138-1144.

Kuhl, P. K. and Meltzoff, A. N. (1984) 'The intermodal representation of speech in infants', Infant Behavior and Development 7, 361-381.

Legerstee, M. (1990) 'Infants use multimodal information to imitate speech sounds', Infant Behavior and Development 13, 343-354.
Lempert, H. and Kinsbourne, M. (1982) 'Effect of laterality of orientation on verbal memory', Neuropsychologia 20, 211-214
MacKain, K. S., Studdert-Kennedy, M., Spieker, S., and Stern, D. (1983) 'Infant intermodal speech perception is a left hemisphere function', Science 219, 1347-1349.
McGurk, H. and MacDonald, J. (1976) 'Hearing lips and seeing voices', Nature 264, 746-748.
Mills, A. E. (1987) 'The development of phonology in the blind child', in B. Dodd and R. Campbell (eds.), Hearing by Eye: The Psychology of Lip-Reading, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 145-161.
Mulford, R., (1988) 'First words of the blind child', in M. D. Smith and J. L. Locke (eds.), The Emergent Lexicon, Academic Press, New York, pp. 293338.
Rosen, S. M., Fourcin, A. J., and Moore, B. C. J. (1981) 'Voice pitch as an aid to lipreading', Nature 281, 150-152.
Summerfield, Q. (1987) 'Some preliminaries to a comprehensive account of audio-visual speech perception', in B. Dodd and R. Campbell (eds.), Hearing by Eye: The Psychology of Lip-Reading, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 3-51.
Walton, G. W. and Bower, T. G. R. (in press) 'Amodal representation of speech in infants', Infant Behavior and Development.