Perception & Psychophysics
1978, Vol. 24 (3), 253-257

Visual influences on speech perception processes

JOHN MacDONALD and HARRY McGURK
University of Surrey, Guildford, Surrey, GU2 5XH, England

An experiment is reported, the results of which confirm and extend an earlier observation that visual information for the speaker's lip movements profoundly modifies the auditory perception of natural speech by normally hearing subjects. The effect is most pronounced when there is auditory information for a bilabial utterance combined with visual information for a nonlabial utterance. However, the effect is also obtained with the reverse combination, although to a lesser extent. These findings are considered for their relevance to auditory theories of speech perception.

Most previous studies of the role of vision in the perception of speech have been of one of two types. Into one group fall those investigations of vision as an *alternative* mode to hearing in speech perception. Thus, for example, Conrad (1977) and Pelson and Prather (1974) compared the lip-reading abilities of deaf and hearing subjects, while Binnie, Montgomery, and Jackson (1974) assessed the visual discriminability of the associated lip movements for a variety of syllabic utterances. A second group of studies has investigated the *compensatory* or *complimentary* role of vision upon the auditory perception of speech under conditions of noise. Dodd (1977) provides a recent example of studies of this type. In both kinds of study, it has been traditionally assumed that, in the case of normally hearing subjects, speech perception in face-to-face contexts is a unimodal (i.e., auditory) process and that the role of vision is essentially independent and additive. Evidence is now becoming available which profoundly challenges this assumption.

In an earlier study, using videorecording techniques, we presented normally hearing children and adults with auditory information for one CV utterance simultaneously with visual information for an alternative utterance (McGurk & MacDonald, 1976). The subjects were instructed to watch the speaker and repeat what she said. Somewhat unexpectedly, the subjects in this study experienced neither intermodal conflict nor domination of one modality by the other. Rather, their responses revealed an interactive relationship between seeing and hearing. For example, when voicing for the utterance /ba-ba/ was dubbed onto lip movements for the utterance /ga-ga/ 80% of preschool children and 98% of adults reported hearing /da-da/; the reverse dubbing process also produced illusions, with subjects typically reporting that they heard the speaker saying /gab-ga/ or /bag-ba/. Similarly, voice for /pa-pa/ with lips for /ka-ka/ elicited /ta-ta/ as the dominant response; here the reverse dub resulted in such responses as /pak-pa/ and /kap-ka/.

The previous literature on speech perception affords few clues as to the process or processes which underlie these phenomena. An analogous fusion of two conflicting auditory stimuli was reported by Cutting (1976), who, in a dichotic listening experiment, observed that with /ba/ presented to one ear and /ga/ to the other many subjects perceived a centrally located /da/. He advanced an acoustic averaging hypothesis to account for his results, arguing that when the relevant acoustic values for /ba/ and /ga/ are averaged, the resulting values are close to those for /da/. It is unlikely that even a modified form of this hypothesis would account for our observations. First, subjects would need to transform the visual input into the appropriate acoustic form before the averaging process could begin. We know from lip-reading studies that individual consonants within the labial and nonlabial catagories are difficult to discriminate visually. Thus, we would expect a relatively wide range of fused responses to any one stimulus configuration, particularly when the visual information was for a nonlabial consonant; instead, we observe that particular auditory-visual combinations yield only one or two types of response. Secondly, an averaging hypothesis would predict comparable responses for, say, ga-lips/ba-voice and ba-lips/ga-voice combinations; this prediction is not fulfilled by our data.

A visual dominance explanation is equally unsatisfactory. Due to the difficulty of discriminating

This research was supported by a grant from the U.K. Leverhulme Trust Fund to author McGurk. We should like to acknowledge our gratitude to Ms. S. Ballantyne who acted as our model. Reprint requests should be addressed to Harry McGurk, Department of Psychology, University of Surrey, Guildford, Surrey, GU2 5XH, England.

visually between stop consonants within the labial and nonlabial categories, such a hypothesis entails that there should be frequent substitutions between /b/ and /p/ for labial visual inputs and that there should be a range of responses whenever the visual stimulus is nonlabial. As noted above, such predictions were not fulfilled by our data.

Our own post facto interpretation of our findings was speculative. Acknowledging the visual similarities between lip movements for /ga/ and /da/, we also assumed that the acoustic waveform for /ba/ contained features in common with that for /da/ but not with /ga/. Thus, in a ba-voice/ga-lips presentation, there would be visual information for /ga/ and /da/ and auditory information with features common to /da/ and /ba/. By responding to the common information in both modalities, a subject would arrive at the unifying percept /da/; similar reasoning would account for the /ta/ response to pa-voice/ka-lips presentations. By the same token, it was argued that with ga-voice/ba-lips or ka-voice/pa-lips combinations the modalities would be in conflict, having no shared features. In the absence of domination of one modality by the other, the listener would have no way of deciding between the two sources of information and would therefore oscillate between them, variously hearing /bagba/, /pakpa/ and so on. Unfortunately, such speculation, although it accounted for the immediate findings, goes only a little beyond the level of description and has no predictive power whatever.

Reconsideration of the study by Binnie et al. (1974) recently led us to develop a more predictive hypothesis to account for our observations. They had found that the lip movements for such CV utterances as /da, ga, ta, or ka/ were visually difficult to discriminate from each other; similarly, lip movements for /ba, pa, or ma/ were frequently confused with each other. On the other hand, labial and nonlabial consonants were never confused. Thus, frontal (labial) place of articulation is visually distinct from middle and back while, visually at least, the two latter are not readily distinguished from each other. Binnie et al. also found that the feature of place of articulation is more efficiently detected by vision than is the feature of manner of articulation. On the other hand, they confirmed Miller and Nicely's (1955) finding that voicing and nasality of consonantal utterances are readily perceived auditorially, even under conditions of low signal/noise ratios. Consideration of these conclusions, together with examination of our earlier findings, has resulted in the tentative development of a manner-place hypothesis to account for the responses we observed following exposure to conflicting auditory-visual inputs. Basically, the hypothesis says that, in face-to-face communication between normally hearing people, manner of articulation of consonantal utterances is detected by ear (e.g., whether the utterance is voiced or voiceless, oral or nasal, stopped or continuant, etc.); place of articulation, on the other hand, is detected by eye. The hypothesis argues that, at an as yet unknown level of processing, information from the two sources is combined and synthesized, resulting in the "auditory" perception of a best fit solution. Thus, manner (auditory) information for a voiced, stopped, oral utterance together with place (visual) information for middle/back articulation will result in the perception of /ga/ or /da/. Similarly, manner information for a voiced, continuant, nasal utterance with place information for middle/back articulation will result in the perception of /na/.

The experiment reported here was designed to assess the generality of the original observatons across all possible auditory-visual combinations of the six stop consonants /p, b, t, d, k, g/ plus /m and n/, and to evaluate the predictive power of the manner-place hypothesis outlined above.

METHOD

Stimuli

A young woman was filmed while she fixated a television camera lens and spoke a series of CV utterances. Each utterance was repeated three times with an interval of approximately .5 sec between repetitions. Each utterance comprised one of the six plosive stops /p, b, t, d, k, g/ or a nasal /m, n/ plus the vowel /a/. Dubbing operations were performed on these recordings to produce a new video film comprising all possible auditory-visual combinations of the original stimuli, a total of 56 combinations in all. The dubbing was carried out so as to ensure, within the temporal constraints of telerecording equipment, that there was auditory-visual coincidence of the release of the consonant in each repetition of each utterance. The 56 auditory-visual combinations were then grouped into four sequences such that reciprocal combinations (e.g., ba-lips/ga-voice; ga-lips/ba-voice) were assigned to different sequences; moreover, each sequence contained at least one example of the lip movements associated with each of the original eight consonants. To each of these sequences were added a further eight auditory-visual combinations comprising, in each instance, the voicing for each CV utterance dubbed onto its own lip movements by the same procedures as were employed in the preparation of the other stimuli. Finally, the order of presentation of stimuli within each sequence was randomized and a 10-sec gap of blank video tape was created between each series of three repetitions of each auditory-visual composite; the series of three repetitions represented one trial. A single film therefore comprised 22 trials, each trial separated by a 10-sec gap, and there were four such films.

Subjects

The sample comprised 44 graduate and undergraduate students aged between 18 and 24 years; there were approximately equal numbers of males and females. None had known hearing defects; a few wore spectacles to compensate for minor deficiencies of vision.

Procedure

Each subject was randomly assigned to an individual viewing of one of the four video films described above. The films were

presented on a 19-in. TV monitor via a National Panasonic V3160E edit videorecorder. The subjects viewed the TV screen at eye level and audio-visual reproduction was of good quality. The subjects were instructed that during each trial they were to watch the female model until she had finished speaking and then to repeat what they had heard her say. Responses were recorded verbatim by the experimenter.

RESULTS

Preliminary analysis revealed no consistent differences between responses elicited by comparable stimuli presented on each film, and data from all four films were therefore combined for further analysis. Overall results are summarized in Table 1, where responses are listed in each cell with percentage rates indicated alongside responses.

Table 1 falls naturally into four quadrants commensurate with a binary split on each modality into labial /p, b, m/ and nonlabial /t, d, k, g, n/ consonants. Along the main diagonal are presented the subjects' responses to each CV utterance when dubbed onto its own lip movements. For purposes of analysis, a correct response was defined as an accurate repetition of the auditory stimulus. Inspection of the diagonal cells indicates an average error rate of only 1.5%, with a range between 0% and 4%. This result establishes the validity of the dubbing procedure and was therefore used as the basis for all further analyses.

The binomial distribution (Siegel, 1956) was used in the analysis of off-diagonal cells with 98% and 2% being taken as the parameter values for p and q (rounded from 98.5 and 1.5, respectively, for computational ease). With these values, an accuracy rate of 80% or less is significantly different from chance ($p < .05$).

Very few errors indeed occurred in response to the voice-lip combinations represented by the off-diagonal cells in the upper left quadrant of Table 1. This result confirms the virtual interchangeability, from the viewpoint of vision, between different places of articulation within the range /da, ta, ga, ka, na/. According to the manner-place hypothesis, auditory presentation of /da, ta, ga, ka, na/ with any of these lip movements should result in non-illusory perception of the auditory stimulus, and that

Table 1
Response Type and Percentage of Response to Match and Mismatch Auditory Visual Utterances

	VS														
AS	da		ta		ga		ka		na		ba		pa		ma	
da	da	100.0	da	100.0	da	100.0	da	100.0	da	100.0	da	58.0	da	82.0	da	83.0
											ba	17.0	bda	18.0	bda	17.0
											bda	17.0				
											pda	8.0				
ta	ta	100.0	ta	98.0	ta	100.0	ta	100.0	ta	100.0	ta	58.0	ta	58.0	ta	100.0
			kta	2.0							pta	25.0	pta	36.0		
											pa	8.5	pa	9.0		
											bta	8.5				
ga	ga	100.0	ga	100.0	ga	100.0	ga	100.0	ga	100.0	ga	83.0	ga	100.0	ga	91.0
											bga	17.0			bga	9.0
ka	ka	100.0	ka	100.0	ka	100.0	ka	100.0	ka	100.0	ka	82.0	ka	82.0	ka	75.0
											pka	18.0	pa	9.0	pa	8.3
													pka	9.0	pka	8.3
															bka	8.3
na	na	100.0	na	100.0	na	100.0	na	91.0	na	96.0	na	64.0	mna	50.0	na	55.0
							la	9.0	mna	2.0	mna	18.0	na	25.0	mna	27.0
									la	2.0	ma	9.0	ma	25.0	ma	9.0
											bna	9.0			bna	9.0
ba	da	27.0	da	82.0	da	64.0	ga	58.0	da	58.0	ba	100.0	ba	100.0	ba	100.0
	ga	27.0	ba	18.0	ga	27.0	da	25.0	ga	25.0						
	a	27.0			ba	9.0	ba	17.0	ba	17.0						
	ba	19.0														
pa	pa	50.0	pa	50.0	pa	55.0	pa	70.0	pa	64.0	pa	83.0	pa	96.0	pa	100.0
	ta	40.0	ka	50.0	ta	27.0	ta	10.0	ka	27.0	ba	17.0	ba	2.0		
	ka	10.0			ka	18.0	ka	10.0	pka	9.0			bpa	2.0		
							tha	10.0								
ma	na	91.0	na	82.0	na	92.0	na	91.0	na	80.0	ma	91.0	ma	100.0	ma	98.0
	ma	9.0	ma	9.0	nma	8.0	ma	9.0	ma	20.0	ba	9.0			bma	2.0
			la	9.0												

Note–AS = auditory stimulus; VS = visual stimulus (lip movements).

is precisely what happens.

The off-diagonal cells in the lower right quadrant of Table 1 also contain few errors, illustrating that, when voicing for /ba, pa, or ma/ is presented with any of the labial lip movements from within the same set, veridical percepton of the auditory stimulus ensues. This result also conforms to the manner-place hypothesis.

However, it is the results from the two remaining quadrants of Table 1 which are critical for the hypothesis under consideraton. Here, combination of information for manner of articulation from the auditory component with place information from the visual component should result in auditory illusions.

When the combinations are of labial sounds with nonlabial lip movements (lower left quadrant of Table 1), then all examples yield highly significant error rates from 30% to 100% with a mean rate of 73% ($p < .001$). Moreover, inspection of the errors reveals that, with relatively few expections, their nature conforms to the prediction of the manner-place hypothesis.

With combinations involving simultaneous presentation of nonlabial sound and labial lip movements (upper right quadrant of Table 1), errors are again in evidence, with the rates for different combinations ranging from 0% to 75%. The average error rate for this segment of the table is 25% ($p < .05$). Not only is there a lower rate of errors than for the other combinations, but such errors as do occur are different from what was anticipated on the basis of the manner-place hypothesis. As in our earlier experiments (McGurk & MacDonald, 1976) the most frequent error involved interpolation of a labial consonant, presumably picked up visually, before the middle or back consonant presented in the auditory modality. Significantly, with few exceptions, this interpolated consonant was voiced if the auditory stimulus was voiced, unvoiced if the auditory stimulus was unvoiced, and nasal if the auditory stimulus was nasal. This much, at least, is in conformity with our hypothesis.

In summary, the results of this experiment serve to confirm and extend the generality of our previous observations (McGurk & MacDonald, 1976). It is evident that these auditory illusions are not mere transitory phenomena elicited by one or two peculiar auditory-visual combinations. Rather, they appear to be illustrative of a general effect of vision upon speech perception in face-to-face situations. Moreover, the results of the experiment confirm the predictive validity of the manner-place hypothesis with respect to the nature of the illusions elicited by labial-voice/nonlabial lips presentations; it is less satisfactory with respect to nonlabial sound/labial lips combinations and therefore will clearly require modification and refinement.

DISCUSSION

No contemporary theory of speech perception, whether passive" or "active" assigns a role to the influence of visual information upon the process of decoding the acoustic stimulus. A moot question, therefore, is the extent to which such theories may be modifiable so as to accommodate such a role.

Consider the kind of passive models of speech perception based on the notion of automatically registering "feature detectors" (e.g., Abbs & Sussman, 1971; Eimas, Cooper, & Corbit, 1973). Cortical detector cells are assumed to respond to complex, multidimensionsl features of the acoustic wave form; individual detectors fire only in response to a unique set of parameter values on specific dimensions. Thus, the model holds that acoustic-phonetic transformations are passively registered. Such a model seems particularly resistant to modification so as to incorporate a role for visual stimulation; it is difficult to conceptualize, within the model, how categorically different detectors would be fired under identical conditions of acoustical stimulation. This, however, is what would be required to account for the data reported above. This is not to dispute the existence of feature detectors. Such detectors may play a role in the mediation of speech perception but whatever the nature of that role, it is not a sufficient one.

There are two principal, highly similar variants of the active orientation towards speech perception. The motor theory argues that the speech signal is initially subject to acoustic analysis to extract such features of the spectral structure as frequency, intensity, and duration. Further decoding into phonetic units takes place via a process whereby the neural signals generated by the sensory input are brought into correspondence with internally generated neural commands which, if implemented, would have produced the articulatory gestures leading to that sensory input. Hence, speech perception and speech production are intimately related and perception is mediated by production. (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

The analysis by synthesis model proposed by Stevens and House (1972) similarly argues for a preliminary analysis of the acoustic stimulus. At this initial level, however, some phonetic information is directly decoded via invariant acoustic features assumed to be in direct relationship with phonetic segments. This allows a hypothesis to be generated, at a neural or muscular level, about the complete acoustic signal. As with the motor theory, the

neural activity corresponds to the articulatory gestures necessary to produce the hypothesized sounds; repeated testing takes place until a match is found with the stored sensory input.

It is noteworthy that both of the above variants already assume a role for intermodel integration (i.e., between proprioception and audition) in speech perception. We propose, therefore, that either variant could readily accommodate a role for vision, that in the generation of neuronal commands relating to the articulatory gestures associated with the spoken word, account is taken of information about place of articulation visually available from watching the speaker. That this process sometimes leads to illusion attests to the active, constructivist nature of the speech perception process.

REFERENCES

ABBS, J. H., & SUSSMAN, H. M. Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. *Journal of Speech and Hearing Research*, 1971, **14**, 23-26.

BINNIE, C. A., MONTGOMERY, A. A., & JACKSON, P. L. Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research*, 1974, **17**, 619-630.

CONRAD, R. Lip reading by deaf and hearing children. *British Journal of Educational Psychology*, 1977, **47**, 60-65.

CUTTING, J. E. Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, 1976, **83**, 114-140.

DODD, B. The role of vision in the perception of speech. *Perception*, 1977, **6**, 31-40.

EIMAS, P. D., COOPER, W. E., & CORBIT, J. D. Some properties of linguistic feature detectors. *Perception & Psychophysics*, 1973, **13**, 247-252.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Perception of the speech code. *Psychological Review*, 1967, **74**, 431-461.

MCGURK, H., & MACDONALD, J. Hearing lips and seeing voices. *Nature*, 1976, **264**, 746-748.

MILLER, G. A., & NICELY, P. E. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 1955, **27**, 338-353.

PELSON, R. O., & PRATHER, W. F. Effects of visual message-related cues, age and hearing impairment on speech reading performance. *Journal of Speech and Hearing Research*, 1974, **17**, 518-525.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

STEVENS, K. N., & HOUSE, A. S. Speech perception. In J. Tobias (Ed.), *Foundations of modern auditory theory* (Vol. 11). New York: Academic Press, 1972. Pp. 3-62.

(Received for publication March 7, 1978;
accepted June 22, 1978.)