A review by
Douglas H. Whalen of Hearing by Eye II: Advances in the Psychology of Speechreading and Audition-Visual Speech. American Journal of Psychology, Vol. 115, No. 4, pp. 631-640.

1269

Book Reviews

DOMINIC MASSARO, EDITOR
University of California, Santa Cruz

Seeing Is Perceiving, Even When It Is Speech

Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech
Edited by Ruth Campbell, Barbara Dodd, and Denis Burnham. Hove, UK: Psychology Press, 1998. 319 pp. Cloth, $80.

Although the use of visual information for speech has been known to be effective for decades—silent movie actors were occasionally fired for having said rude things on camera even when their comments did not appear in the titles—it was not until the serendipitous discovery of McGurk and McDonald that we learned that vision affects speech even when the auditory signal is clearly present. This discovery has led to a productive exploration of just what it means to say that speech is an acoustic signal and to examine the types of information that can be used visually to influence speech perception.

Progress in this field has been aided by the partly deliberate, partly circumstantial concentration of a small group of researchers who have pursued this area vigorously and who meet regularly to exchange progress and ideas. This has allowed researchers, scattered across the globe, to push back the frontiers of knowledge at a rate that otherwise would have been impossible. Unfortunately, this circle of colleagues has been reduced by the untimely deaths of three of its stellar members: Harry McGurk, whose discovery started an entire line of research; Christian Benoit, co-organizer of the NATO meeting at Bonas, France, in 1995 that started the chain of collaboration; and Kerry Green, one of the most productive members of the group. Research in this area proceeds, with a tinge of sadness. Although not designed for the purpose, this volume is a fitting memorial for these three cherished colleagues by virtue of its superior contribution to the fields of psychology, linguistics, and speech science.

The book follows the original *Hearing by Eye* by 11 years and again focuses on "lipreading in hearing people rather than the use of lipreading in deafness" (p. ix). The main implication of the McGurk effect is that speech perception is either multimodal or amodal. Thus theories that assume only acoustic structures cannot explain an extremely robust and reproducible effect in normal listeners. Although no theory of speech perception is completely adequate, none should treated as beginning to be adequate unless it takes bimodal perception into account. As often happens, though, the evidence has existed in the literature without affecting the thinking of many speech researchers. The

appearance of this book not only pushes forward our thinking about audiovisual integration but also restates the case for inclusion of such effects in any successful theory of speech perception.

The chapters in this volume cover broad topics with implications for theories of speech perception, engineering issues, neuropsychology, speechreading in the deaf, and sign augmentation. Although no section will be equally interesting to all readers, the chapters in each represent the leading research in each area.

The late Kerry P. Green's chapter reports some of the most convincing evidence to date that auditory and visual sources of information are integrated at an early stage. Looking at low-level coarticulatory effects, he found that the coarticulatory information available in the visual signal affected identification in the gradient fashion we would have expected from a purely acoustic context. The chances that these effects result from the integration of features or categories already extracted by one modality or the other seem quite small. Green then reviews evidence that shows that even very young children integrate the two modalities but with somewhat different weights than adults. This may reflect the general state of maturation of the phonological system rather than a change in audiovisual integration per se. His results make it clear that there is a great deal more to be learned before the theoretical implications of the McGurk effect are fully understood but that the ultimate theory probably requires immediate input from the visual modality into the speech perception process.

Denis Burnham further explores the developmental issues and relates them to cross-language differences. He reviews evidence that very young children are sensitive to audiovisual mismatches, indicating that there is unlikely to be a large learning component to this integration: It is found at almost the same age at which the visual system is mature enough to process faces. The adult work on crosslanguage comparisons shows both general and language-specific effects. He raises the interesting but so far unexplored possibility that nearly equivalent gestures (such as lip rounding) may contribute differently to different language percepts because of small differences in the visual appearance of that gesture across languages. Here again, a broad range of important questions is raised for future work to address.

Lawrence D. Rosenblum and Helena M. Saldaña find evidence of the need for dynamic information in the use of point light displays. These displays place markers on various points of the face and film the production of speech at such low contrast that only the points appear. When these are static, they are not even recognized as faces. But when the dynamic information is present, the phonetic content can be read as well as normal video, and they influence the heard signal in the same way as full video. In contrast, still video frames of the normally illuminated face at critical articulatory points did not influence heard speech. Although the possibility exists that the conflicting information in the still frames is responsible for this lack of effect (because the vowel occurs with a closed vocal tract, which is not possible in the real world), the evidence that the dynamics is sufficient for the McGurk effect is a powerful one whose implications are still being worked out.

Jean-Luc Schwartz, Jordi Robert-Ribes, and Pierre Escudier provide a taxonomy of the various theories of audiovisual integration. They classify theories as

to whether the two modalities directly interact (DI), are separately identified (SI), have one modality that is dominant in recoding (DR), or are jointly turned into a motor recoding (MR). Three aspects of perception, then, lead to a tentative choice among the four possibilities. The fact that the modalities can be seen to conflict yet still influence one another argues against the DI type. The fact that pitch trains that are not heard as speech by themselves can still affect voicing judgments when presented with the video signal indicates that integration must be late in the process, casting doubt on SI models. The lack of dominance of audition in speech perception further seems to separate DR and MR theories, with the motor approach being most compatible with existing results. Hybrid models are also considered but deemed unnecessarily complicated. Although the models that these authors consider collapse the time dimension, it may be that this is unnecessarily restrictive: It is commonly the case that visual cues (and, for some theories, acoustic ones as well) are thought of in static terms, but this may not be enough to capture the differences between theories (as suggested by the Rosenblum and Saldaña results).

N. Michael Brooke examines the data reduction needed to make visual information useful in automatic speech recognition. Although the most efficient means of reduction would be to find the parameters that fully specify the speech information, that level of analysis is beyond our current knowledge. Instead, some simple parameterizations of images of the mouth are shown to be almost as useful in recognizing speech as the original images, allowing efficient use of vision without a fully developed theory of the visemes. However, the need to restrict the area of the visual image limits the usefulness of this approach for more general applications because cameras in use outside the laboratory will be following a moving target not well suited to automatic extraction of the mouth features.

Kevin G. Munhall and Eric Vatikiotis-Bateson report on a groundbreaking series of experiments examining the degree to which different portions of the face contain speech information. They have found that there is a great deal of redundancy in the movement of parts of the face far removed from the lips. Not only are correlations with the movement good, but a moderately intelligible speech signal can be synthesized from them. They have also studied patterns of gaze direction during speech, where they found that listeners look more at the mouth with increasing levels of noise. But even at the highest noise level, listeners still look at the eyes of the talker about half the time. The authors suggest that this results from the importance of eye contact during conversation but that visual retrieval of phonetic information is still possible because of the redundancy they found and the importance (and availability) of movement detection in the parafoveal field.

Jerker Rönnberg, Stefan Samuelsson, and Björn Lyxell review an interesting body of work showing that phonological working memory is correlated with individual differences in lipreading ability. Their work shows that only early heightened use of the visual modality, caused by early deafness, increases lipreading ability directly; simply needing to use the information and paying more attention to it (as for those who become deafened late in life) is not enough to enhance the use of visual cues. Context provides important clues, and the

best lipreading performance comes with typical sentences in familiar, "scripted" situations.

Timothy R. Jordan and Paul C. Sergeant report the results of two experiments showing that the visual contribution to speech perception (in both congruent and incongruent, McGurk stimuli) is robust even with greatly reduced size of the visual image. What differences they found occurred when the image was 5% or 2.5% of the original size; reductions to 10% were equivalent to full size. Given the wide range of visual images used in today's electronic world, it is important to know that the visual information is still usable at vastly different scales.

Ruth Campbell attempts to settle apparent contradictions in the neurological evidence for the location of audiovisual speech integration. Early indications were that the right hemisphere was largely responsible for speechreading, but later results implicated the left or even bilaterality. A plausible account is suggested in which the majority of visual speech processing occurs in the right hemisphere, with communication being interrupted by some left hemisphere lesions, accounting for the complicated results in the aphasia literature.

Beatrice De Gelder, Jean Vroomen, and Anne-Catherine Bachoud-Levi present a detailed case study of the effects of severe visual impairment that nonetheless spared some aspects of speechreading. After a stroke, this patient was unable to recognize familiar faces and objects yet seemed to have normal potentials in the early visual pathways. Her categorization of still photographs of vowels was barely above chance, but dynamic productions were categorized much better. The effect on audiovisual speech integration was complicated; in some tests, one category or another dominated the judgments depending only on which test it was. Further tests found good visual recognition but poor auditory and bimodal perception. A final test showed fairly good memory for silently mouthed sequences of numbers. This seems to be a case in which visual movement recognition was spared (unlike static vision), but the use of that information for speech recognition was unlike that found in unimpaired listeners. As the authors conclude, case studies such as this indicate that there is much more going on (or at least much more is possible) in audiovisual perception than current theories assume.

Lynne E. Bernstein, Marilyn E. Demorest, and Paula E. Tucker present important evidence that speechreading ability is stronger in the deaf rather than in the hearing, as had previously been reported. The reason for the previous report probably is a matter of sample size because working with the deaf entails a larger set of problems in subject selection than with the hearing. The attributes that correlate with speechreading ability have more to do with amount and success with using English (both spoken and written) than with similar experience of American Sign Language (ASL). Almost all listeners improved in their speechreading with hearing aids, even if the level of attainment was not very high. Still in question, however, is what the people who perform well above average do differently from others with similar backgrounds who do far less well.

Barbara Dodd, Beth McIntosh, and Lynn Woodhouse provide longitudinal evidence from 11 children that success with speechreading correlates with early success in syntax and semantics of the spoken language. All the children test-

ed were being raised in a "whole communication" environment in which most sentences addressed to them were both spoken and signed (in English), so all had lifelong experience with (essentially) spoken English. The exact nature of the correlation with syntactic ability is not clear, but an enormous range of ability in speechreading is again evident. If this ability can be taught, and the present correlation results from the actual speechreading rather than from some predisposition for being better at decoding language, then further improvements in the English language abilities of the deaf could be expected. It could turn out that speechreading ability will be found to be as good a predictor of reading ability in the deaf as phonemic awareness is in the hearing.

Marc Marschark, Dominique LePoutre, and Linda Bement point out unexpected parallels between the use of mouth gestures in deaf signers and the use of visual information in conjunction with spoken language. Some signs in ASL have common oral gestures associated with them, and these oral gestures can, in some cases, replace the manual ones. Although the stability of such gestures is questioned, it is clear that both the manual and facial visual spaces are used. This may arise so naturally because of the common sources of spoken and manual language (although we may never have a definitive answer to the ultimate origins of human language). Alternatively, the use of mouthed speech may result from the increase in oralist orientation in education, but its presence even in ASL brings up intriguing relations with the incorporation of vision in the spoken modality.

Michael Oerlemans and Peter Blamey explore issues raised by the tactile conveyance of language information. There is evidence that feeling the articulators moving, as in the Tadoma method of speech enhancement, is integrated into a speech percept at an early stage of processing. The various devices that translate portions of the speech signal into vibrations or electrical stimulation are somewhat successful at inducing speech percepts. The authors point out that the kind of information that is successfully transmitted by these devices is (not surprisingly) the same kind of information as is transmitted by the speech signal itself, so they do not add as much information as the visual signal does when paired with the speech signal.

Jacqueline Leybaert, Jésus Alegria, Catherine Hage, and Brigitte Charlier add the interesting technique of cued speech to the debate. Cued speech is a method devised for aiding lip reading, using hand shapes that the speaker makes in conjunction with spoken language. The hand shapes are ambiguous between three or so phonemes (e.g., /k/, /v/, and /z/) that are easy to distinguish visually (so information of voicing, for example, is carried by the handshape). Children who grow up being taught this method (and who use it at home as well as at school) perform better than those who learn it later in life. The results are interesting in that the information provided is somewhat categorical rather than being similar to the speech itself (as with the tactile devices discussed by Oerlemans and Blamey). The authors point out that this result causes some difficulty for the motor theory of speech perception, but they wrongly assume that it is compatible with Fowler's theory of direct perception; the inclusion of featurally specified rather than gesturally specified information is equally unexpected in her theory. A theory that does not care at what level the

information exists, such as Massaro's fuzzy logical model of perception, may be the most compatible. This area clearly deserves further work to elucidate its implications.

Overall, this book is an outstanding contribution to the field and a worthy sequel to the first *Hearing by Eye.* The implications for speech theory and psychology in general are still being explored in the literature. Our world is a multimodal one, and our theories must take that into account. Just as the barn owl does not care whether it hears or sees its prey as long as it gets to eat, we are constantly taking in information about our surrounding without regard to its sensory source. We should be more surprised when we find that the senses are kept apart because what we want to know about is the world, not our sense data. This collection will be an active part of illuminating this viewpoint for a decade (when the next volume should, by rights, appear) and beyond.

Notes

Preparation of this review was supported by NIH grants HD-01994, DC-02717, and DC-00403 to Haskins Laboratories. Thanks to Julia R. Irwin and Carol A. Fowler for comments on earlier drafts.

D. H. Whalen
Haskins Laboratories
270 Crown Street
New Haven, CT 06511
E-mail: whalen@haskins.yale.edu

Diversity in Education: It Is All in the Language

Language Diversity and Education

By David Corson. Mahwah, NJ: Erlbaum, 2000. 256 pp. Paper, $29.95.

Given what we now know about ethnic, cultural, linguistic, socioeconomic, and gender diversity, no teacher candidate should complete his or her training without a course on diversity in education. Furthermore, no course on diversity in education is complete without considering the role language variance plays in establishing and perpetuating peoples' differences. With the publishing of Corson's *Language Diversity and Education,* we now have a single text that introduces educators to this all important and yet often neglected aspect of diversity by focusing on language differences for all these diversity types.

Although language differences of our immigrant, bilingual students are an obvious concern for educators, and although language differences between the genders have become somewhat part of America's popular culture in the past decade with Tannen's bestseller and the *Mars/Venus* popularity, fewer educators are fully informed of the issues surrounding nonstandard varieties of English. Still fewer educators are even aware of the language variances at the discourse level that exist for various cultural and socioeconomic groups. Readers of this text will become fully aware of the issues of language diversity for each of these four populations.