FOR SPEECH PERCEPTION BY HUMANS OR MACHINES, THREE SENSES ARE BETTER THAN ONE

Lynne E. Bernstein

Spoken Language Processes Laboratory
House Ear Institute
Los Angeles, California

Christian Benoît

Institut de la Communication Parlée
Université Stendhal
Grenoble, France

ABSTRACT

A growing assemblage of researchers has, in recent years, adopted methods and theories that acknowledge and exploit the multisensory nature of speech perception. This paper, which is an introduction to the special session, "The Senses of Speech Perception," gives a brief **historical** review of research concerning the multiple senses of speech perception, discusses **major issues**, and suggests **directions for future research**. (Work supported in part by NIH grant DC00695.)

1. INTRODUCTION TO THE SPECIAL SESSION ON THE SENSES OF SPEECH PERCEPTION

Not many years ago, the speech perception tool kit was complete with instruments to measure, sample, process, synthesize, and output only acoustic speech signals. Presently, for a growing number of speech perception researchers, instrumentation includes systems that deliver and/or sample optical speech signals. For a few investigators, speech stimuli are generated by movement and/or vibration transducers. The motivations for this multimedia armamentarium are numerous. They can be found in speech perception **theories**, **clinical applications**, and **technological developments.** Whatever the motivation, the adoption of multimodal speech research methods acknowledges as a fundamental fact that speech perception is multisensory.

The assignment accepted by the lecturers in this special session was to present research on speech perception or recognition involving audition, vision, and/or touch, and to address the question, What do the facts of multisensory speech perception imply about the nature of the spoken language processing system? The self-assigned task of the present authors was to briefly point out the precedents in the literature for multisensory speech perception and highlight several important areas under the headings of theoretical, clinical, and technological issues.

2. HISTORICAL BACKGROUND

Current interest in multisensory speech perception can be attributed largely to findings reported by McGurk and MacDonald [1]. They discovered that when acoustic /bɑ/ was presented in synchrony with a face saying /gɑ/, subjects reported that they heard /dɑ/. This effect of vision on heard speech drew attention because it was unanticipated by then extant theories of speech perception and was even contrary to intuitions.

Other related studies followed the initial description of the McGurk illusion [e.g., 2,3], and several different theories were offered [4,5,6]. Green [2], in this session, presents an overview of research on the McGurk illusion. The paper by Sekiyama et al. [3], in this session, shows that the illusion is sensitive to native language, acoustic quality, language proficiency, and cultural factors.

Although highly influential, the McGurk illusion should not be mistaken as the seminal research on crossmodal speech effects. Precedence belongs with researchers who showed that word identification accuracy under noisy acoustic conditions improves when the listener can also see the talker [7,8,9,10, also see 11]. Precedence also belongs with clinical researchers who examined visual-alone and audiovisual speech perception in individuals with hearing losses. Initial research in this area predated McGurk, in some cases by more than 50 years [12,13,14,15]. Finally, research on sensory supplements or substitutes for use by individuals with profound hearing loss is the precedent for investigations of speech perception via touch and dates to the early twentieth century [16]. Reed [17], in this session, discusses Tadoma, a method used by a small number of deaf-blind people to perceive speech through their hands. Developments in artificial sensory substitution have afforded further evidence that speech can be perceived via touch [18,19,20,21,22].

Without question, ample evidence exists in the literature for multisensory speech perception. The ease with which people frequently comprehend speech by audition alone and the methodological rigors associated with controlling several different stimulus modalities are perhaps two reasons why speech perception research to date has been almost exclusively focused on auditory speech perception.

3. THEORETICAL ISSUES

One surprising fact about multisensory speech perception is that it frequently causes superadditivity. Superadditivity occurs when

speech perception accuracy with two sources of information is greater than predicted by the sum of accuracy measures for the individual sources. A good example of superadditivity occurs when the face of the talker is combined with an acoustic signal that presents the same talker's voice fundamental frequency (F0). By itself, the F0 signal is not intelligible. When F0 is combined with visible speech (for which average performance by adults with hearing is approximately 20-30% words correct in sentences) performance is typically enhanced by approximately 20-40 percentage points words correct [e.g., 23,24].

Theoretical explanations are needed for superadditivity, as well as for enhanced speech perception in noise, the McGurk illusion, and visual-tactile speech perception. Each of these involves intersensory information organization and integration.

Summerfield [25] argued that integration of speech information occurs prior to linguistic categorization, and he suggested alternative common metrics as terms in the integration calculus. His inventory of common metrics for intersensory integration is similar to the inventory of positions taken by theorists in speech perception concerning what are the objects of speech perception. Possibilities include the sounds of speech [26], the articulatory gestures [5], the talker's linguistic intentions [4], and the linguistic units of language [27]. Green [2], in this session, points out that the McGurk illusion is frequently interpreted as evidence that the common metric is articulatory, and he reviews other phenomena that support either an articulatory or auditory metric.

Integration implies that the perceiving system somehow registers that diverse information belongs to the same event. Remez [28], in this session, argues that intersensory integration shares characteristics with intrasensory integration. He proposes that the organization of speech information is not achieved by a cognitive decision process nor by similarity principles of Gestalt psychology but rather relies on perception of informational coherency across sensory systems. Acceptance of this proposal implies a search for the perceptual principals [see 29] and/or formal expressions for intersensory coherency.

Vatikiotis-Bateson et al. [30], in this session, discuss several quantitative analyses that appear to reflect coherency across data from video recordings of talkers' faces, 3D facial marker positions, speech acoustics, and EMG recordings from perioral muscle activity. Reported high correlations between 3D facial marker positions and RMS amplitude of the speech acoustic signals are evidence that speech signals afford characteristics that could induce the intersensory coherency Remez [28] predicts. A possible implication of the Vatikiotis-Bateson et al. results is that a common metric is not required for intersensory integration, if temporal alignment is maintained across sources of information.

Pisoni et al. [31], in this session, investigated whether audiovisual speech processing enhances and/or diminishes memory capacity. They assume the integration process and ask how it affects higher-level processing. Their results suggest that intersensory integration is achieved at a cost to memory span, although it enhances long-term storage. Whether/how these costs and benefits affect spoken language comprehension remains to be discovered.

4. CLINICAL ISSUES

Although experiments involving individuals with hearing and/or visual deficits, or with brain lesions are typically consigned to the primarily clinical literature, much basic knowledge is to be learned about multisensory speech perception from such individuals. By investigating speech perception in the absence of auditory and/or visual experience, and speech perception by individuals with specific brain damage, it is possible to dissociate modality-specific from modality-independent characteristics of speech perception and language processing.

One source of modality-specific effects is the localization of processing in different brain structures as a function of the site of sensory stimulation. Campbell [32], in this session, presents evidence that visual-alone speech perception engages somewhat different visual areas of the brain than does audiovisual speech perception. Her studies with brain-lesioned patients suggest the possibility that the architecture for processing visible speech differs from that for acoustic speech.

Reed [17], in this session, reports on studies of experienced deaf-blind users of Tadoma. Reed and colleagues [33] estimated that Tadoma performance with connected speech was equivalent to listeners' performance with sentences in speech-to-noise conditions of 0 dB, that is, approximately 70% words correct in sentences. Their results are seen as an existence proof for speech perception via touch alone. Bernstein, Demorest and Tucker [34] have studied highly accurate visual speech perception in adults with profound congenital hearing losses. Similarly to the expert Tadoma users, subjects' accuracy rates for words in sentences hovered near 80% words correct. That several of the individuals studied had little if any auditory experience can be seen as an existence proof for speech perception via vision alone.

Results from congenitally deaf and deaf-blind individuals strongly suggest that speech perception is potentially modality independent. The paper by Pisoni et al. [31] is, however, a caution against a premature leap in this direction, given their findings that modality-specific information is carried forward into long-term memory. Furthermore, the fact that lipreading proficiency varies more widely in both hearing and deaf populations [34] than does auditory speech perception suggests that modality does matter in ways that need to be specified.

5. TECHNOLOGICAL ISSUES

It has long been a human dream to create machines able to speak and/or to understand speech. In the period of ancient Greece, priests used talking statues as oracles to cheat on and better convince their audience. More recently, several speaking machines have been devised, first as mechanical devices, then as electronic devices, and today mostly with computers. Indeed, the new era of multimedia renews the challenge of multimodal communication. Although the keyboard and mouse remain the most widely used modes of human-machine communication, it is

obvious that spoken communication is more natural and requires less practice. Multimodal human-machine spoken communication is a big challenge today, and technological spin-offs are naturally expected from a better understanding of how humans produce and understand multimodal speech. In return, synthetic speaking faces and automatic multimodal speech recognizers are tools that allow i) quantitative evaluation of theoretical human speech production and perception models, and ii) the generation of highly controlled speech stimuli unavailable from human talkers.

Automatic recognition of acoustic speech has been extensively studied over the last two decades. In contrast, automatic lipreading has been minimally investigated, with the exception of a few pioneering efforts [35,36,37]. Today, we see the emergence of interest in automatic recognition of audiovisual speech, due to several factors. First, performance of acoustic speech recognizers is plateauing, whereas usage is more widespread, particularly in noisy environments, in which optical information could dramatically compensate for poor quality acoustic signals. At the same time, real-time image processing is now accessible on many computers: small light-weight micro-cameras are increasingly common, so that it is now feasible to design automatic lipreaders.

As discussed by Brooke [37], there are two main areas of research in the design of audio-visual speech recognizers -- extraction of optical characteristics and integration of information -- both addressing basic questions relevant to the human ability to process speech bimodally. One approach to extracting optical cues uses techniques such as lip contour detection, by analysis of lip luminance and/or chrominance, associated with the adjustment of a lip model or of deformable templates. The other uses stochastic techniques to associate raw face images with labial features or visemes [see 38]. Also, lip dynamics can be evaluated with techniques such as optical flow analysis. Successful automatic lipreaders will certainly integrate all these techniques together in a hybrid approach that will weight processors of color, lighting, shape, and movement.

Optical-acoustic integration is a major challenge for machine speech processing. Following Summerfield [25], several models have been proposed and evaluated for automatic integration [40,41]. It is possible that speech information transmitted through various channels (e.g., a microphone, a camera, or a "face-glove") to a machine must be recoded into an "amodal" representation. Robert-Ribes et al. [40] suggested the motor (or articulatory) space as a common metric. A prerequisite to a common metric is to determine whether integration should occur early or late, that is, whether the optical and acoustical flows are processed and decoded separately or appended as a single vector. Adjoudani and Benoît [41] demonstrated better performance when outputs from the optical and acoustical decoders were first processed independently and then weighted depending on their estimated reliability. This late integration of weighted modalities is the only one to date that passes the basic test for any multimodal speech recognizer, namely, that multimodal performance is higher than that observed with any unimodal processor, as is the case with human perceivers.

The studies in this session focus on the remarkable ability of humans to process speech through multiple senses. Human perception is a fundamental framework for developing anthropomorphic machines, and machines are essential tools to model and test the integration of multisensory information.

6. REFERENCES

1. McGurk, H., and MacDonald, J. (**1976**). “Hearing lips and seeing voices: A new illusion,” Nature **264**, 746-748.

2. Green, K. P. (**1996**). “Studies of the McGurk effect: Implications for theories of speech perception,” *these Proceedings.*

3. Sekiyama, K., Tokhura, Y., and Umeda, M. (**1996**). "A few factors which affect the degree of incorporating lip-read information into speech perception", *these Proceedings*.

4. Liberman, A. M., and Mattingly, I. G. (**1985**), "The motor theory of speech perception revised,” Cognition **21**, 1-36.

5. Fowler, C. A. (**1986**). "An event approach to the study of speech perception from a direct realist perspective," J. Phonet. **14**, 3-28.

6. Massaro, D. W. (**1987**). *Speech perception by ear and eye: A paradigm for psychological inquiry* (Lawrence Erlbaum Associates, London).

7. Sumby, W. H., and Pollack, I. (**1954**). “Visual contribution to speech intelligibility in noise,” J. Acoust. Soc. Am. **26**, 212-215.

8. Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). “The intelligibility of speech as a function of the context of the test materials,” J. Exp. Psychol. **41**, 329-335.

9. O'Neill, J. J. (**1954**). “Contributions of the visual components of oral symbols to speech comprehension,” J. Speech Hear. Disord. **19**, 429-439.

10. Erber, N. P. (**1969**). “Interaction of audition and vision in the recognition of speech stimuli,” J. Speech Hear. Res. **14**, 848-857.

11. Benoît, C., Mohamadi, T., and Kandel, S. (**1994**). "Effects of phonetic context on the audio-visual intelligibility of French,” J. Speech Hear. Res. **37**, 1195-1203.

12. Jeffers, J., and Barley, M. (**1971**). *Speechreading (Lipreading)* (C. C. Thomas, Springfield, IL).

13. Erber, N. P., and McMahan, D. A. (**1976**). “Effects of sentence context on recognition of words through lipreading by deaf children,” J. Speech Hear. Res. **19**, 112-119.

14. Nitchie, E. B. (**1916**). "The use of homophenous words," Volta Rev. **18**, 85-93.

15. Utley, J. (**1946**). “A test of lip reading ability,” J. Speech Hear. Disord. 11, 109-116.

16. Gault, R. H. (**1924**). “Progress in experiments on tactual interpretation of oral speech,” Soc. Psychol. **14**, 155-159.

17. Reed, C. M. (**1996**). “The implications of the Tadoma method of speechreading for spoken language processing,” *these Proceedings.*

18. Sherrick, C. E. (**1984**). “Basic and applied research on tactile aids for deaf people: Progress and prospects,” J. Acoust. Soc. Am. **75**, 1325-1342.

19. Summers, I. R. (Ed.) (**1992**). *Tactile Aids for the Hearing Impaired* (London, Whurr).

20. Bernstein, L. E., Demorest, M. E., Coulter, D. C., and O'Connell, M. P. (1991). “Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects,” J. Acoust. Soc. Am. **90**, 2971-2984.

21. Waldstein, R. S., and Boothroyd, A. (**1995**). “Speechreading supplemented by single-channel and multi-channel tactile displays of voice fundamental frequency,” J. Speech Hear. Res. **38**, 690-705.

22. Weisenberger, J. M., Broadstone, S. M. and Saunders, F. (1989), “Evaluation of two multichannel tactile aids for the hearing impaired,” J. Acoust. Soc. Am. **86**, 1764-1775.

23. Breeuwer, M., and Plomp, R. (**1986**). “Speechreading supplemented with auditorily presented speech parameters,” J. Acoust. Soc. Am. **79**, 481-499.

24. McGrath, M. and Summerfield, Q. (**1985**). "Intermodal timing relations and audio-visual speech recognition by normal hearing adults,” J. Acoust. Soc. Am. **77**, 676-685.

25. Summerfield, Q. (**1987**). "Some preliminaries to a comprehensive account of audio-visual speech perception,” in *Hearing by eye: The psychology of lipreading*, edited by B. Dodd and R. Campbell, (Lawrence Erlbaum Associates, London) pp. 3-51.

26. Diehl, R., and Kluender, K. (**1989**). “On the objects of speech perception,” Ecol. Psychol. **1**, 121-144.

27. Remez, R. E. (**1996**). “Critique: Auditory form and gestural topology in the perception of speech,” J. Acoust. Soc. Am. **99**, 1695-1698.

28. Remez, R. E. (**1996**). "Perceptual organization of speech in one and several modalities: Common functions, common resources," *these Proceedings.*

29. Meredith, M. A., & Stein, B. E. (**1986**). “Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration,” J. Neurophysiol. **56**, 640-662.

30. Vatikiotis-Bateson, E., Munhall, K. G., Kasahara, Y., Garcia, F., and Yehia, H. (**1996**). “Characterizing audiovisual information during speech,” *these Proceedings.*

31. Pisoni, D. B., Saldana, H. M., and Sheffert, S. M. (**1996**). “Multi-modal encoding of speech in memory: A first report,” *these Proceedings.*

32. Campbell, R. (**1996**). “Seeing speech in space and time: Psychological and neurological findings,” *these Proceedings.*

33. Reed, C. M., Rabinowitz, W. M., Durlach, N. I., Delhorne, L. A., Braida, L. D., Pemberton, J. C., Mulcahey, B. D., and Washington, D. L. (**1992**). “Analytic study of the Tadoma method: Improving performance through the use of supplementary tactual displays,” J. Speech Hear. Res. **35**, 450-465.

34. Bernstein, L. E., Demorest, M. E., and Tucker, P. E. (**1996**). “Speech perception without hearing,” manuscript in preparation.

35. Petajan, E. (**1984**). *Automatic Lipreading to Enhance Speech Recognition*, PhD thesis (University of Illinois at Urbana).

36. Brooke, M., and Petajan, E. (**1986**). “Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics,” *Proc. Int. Conf. Speech I/O, Techniques & Applications* (London) pp. 104-109.

37. Brooke, N. M. (**1996**). “Using the visual component in automatic speech recognition,” *these Proceedings.*

38. Finn, K. E. (**1986**). *An Investigation of Visible Lip Information to be Used in Automated Speech Recognition*, PhD thesis (Georgetown University, Washington, DC).

39. Stork, D., and Hennecke, M., Editors (**1996**). *Speechreading by Humans and Machines, NATO-ASI Series F: Computer and System Sciences, Vol. 150* (Springer-Verlag, Berlin).

40. Robert-Ribes, J., Schwartz, J. L., and Escudier, P. (**1995**). “A comparison of models for fusion of the auditory and visual sensors for speech perception,” Artificial Intelligence Rev. **9**, 323-346.

41. Adjoudani, A., and Benoît, C. (**1995**). "Audio-visual speech recognition compared across two architectures,” *Proceedings of the Eurospeech’95 Conference, Vol. 2* (Madrid, Spain), pp. 1563-1566.