SPECIAL SESSION: ISSUES IN AUDIOVISUAL SPOKEN LANGUAGE PROCESSING (WHEN, WHERE, AND HOW?)

Lynne E. Bernstein [1], Denis Burnham [2], Jean-Luc Schwartz [3]

[1] Department of Communication Neuroscience, House Ear Institute, Los Angeles, USA
[2] MARCS Auditory Laboratories, University of Western Sydney, Sydney, AUSTRALIA
[3] Institut de la Communication Parlée, UMR CNRS No 5009, Grenoble, FRANCE
lbernstein@hei.org

ABSTRACT

The many aspects of audiovisual (AV) speech processing have attracted a loyal following from a wide range of perspectives and disciplines. Beginning with the landmark NATO workshop in 1995 [50], a series of international special sessions and meetings has fostered the development of numerous lines of AV speech research. This paper is an introduction to the special session and an inventory of several developments in the area. The introduction is organized around three questions concerning human and machine performance—*when*, *where*, and *how* AV speech processing occurs—and a summary of applications, including AV speech recognition and synthesis, AV processing with cochlear implants, and AV effects observed in second language learners.

1. INTRODUCTION

Almost 50 years after Sumby and Pollack [51] reported on functional enhancements that visible speech produces when combined with audible speech in noise, many questions about human and machine AV speech processing remain unanswered. Fortunately, a critical mass of scientists and engineers is now attempting to find answers.

Many speech scientists regard AV speech phenomena as particularly challenging, given long-held assumptions about the primacy of the auditory system for processing speech stimuli. Other scientists concerned with perception but not necessarily with speech regard the McGurk effect [37] (in which auditory /ba/ dubbed onto the visual lip and face movements for /ga/ is perceived as *da*, or *tha*) as one of the quintessential examples of human multisensory integration. Perhaps, of particular importance today is the possibility that human communication between individuals with different languages can be improved when the dialogue is AV [Hazan et al., this session; Hardison, this session; Erdener and Burnham, this meeting].

In the domain of speech engineering, the availability of optical signals has led to efforts to enhance acoustic signal processing for speech recognition [Deligne et al., this session; 43]. The observed correspondence between optical and acoustic signals in naturally produced speech has led to new possibilities for creating realistic audiovisual speech synthesis [Bailly, this session; 2].

In the last few years, AV speech issues have progressively "invaded" and energized many aspects of speech research, for example, attention and memory [22], speech detection [27], phonetics [35], sociolinguistics [48], modularity [44], perceptual representations [46], prosody [17], emotion [19], and even the perception of lexical tone [11]. The study of AV speech processing has renewed or enhanced speech research and entered into the core of our modern conception of human speech.

2. THE QUESTIONS

2.1. *When* does A and V speech information combine?

This question is critical for explaining AV speech processing. It is frequently examined in adults' behavior in terms of perceptual identification of stimuli and response times (RTs). These measures suggest that auditory (A) and visual (V) information combines early. Interestingly, Sekiyama [this session] shows that RTs to incongruent McGurk stimuli are typically longer than to congruent stimuli. Evidence consistent with this has also been reported using electrical event-related potentials (ERPs) [4, see also 18]. Sekiyama interprets her RT and functional brain imaging results as indication that AV integration includes a process of optimal weighting of different modalities under given circumstances, rather than passive processing of sensory inputs.

Ponton et al. [this session] examine the time course of speech processing through the central nervous system. They suggest that the neurobiology that supports speech perception constrains the time (and cortical location) at which AV interactions that are specific to speech can occur. They suggest that early AV interactions are not specific to speech stimuli, likely arising at subcortical levels, whereas later ones likely are.

At the developmental level, *when* takes on a different meaning: When in early development does AV speech perception occur? And according to what developmental schedule does perception change? McGurk and MacDonald [37] reported that 3- to 5- and 7- to 8-year-old children had 59% and 52% visually-influenced responses, respectively, compared with adults' 92%. However, the three groups identified auditory stimuli equally accurately (near ceiling), suggesting that AV interaction increases as a function of development.

Talking faces constitute a major part of an infant's perceptual experience. A number of studies have reported evidence for the McGurk effect in early infancy [9,32]. According to Lewkowicz [this session], through the process of watching and listening while people talk to them, infants have the opportunity to acquire cognitive, linguistic, social, and emotional skills. The acquisition of these skills depends, in part, on infants' ability to integrate the audible and visible attributes of speech. Lewkowicz' studies examining infants' perception of AV speech show that responsiveness to various features of faces and voices changes

throughout early human development, and that intersensory integration of these features depends on the nature of the information, its source modality, and the infant's developmental age.

The importance of AV speech early in development can also be seen with developmental data on the *production* of speech. In babbling, the predominance of easy-to-see bilabials increases in deaf children [49] but decreases in blind children [39]. Lipreading is engaged in speech acquisition, as demonstrated by the difficulty of blind children to acquire the /m-n/ contrast [38], which is difficult to hear and easy to see.

2.2. *Where* does auditory-visual speech processing occur?

The question of where speech combines can be answered in terms of linguistic structure and in terms of functional neuroanatomy of the cerebral cortex. As suggested earlier, laboratory results suggest that A and V information combines at the sub-segmental level. Methods using functional magnetic resonance imaging (fMRI) hold out the possibility of localizing the neuroanatomical site of AV stimulus interaction. The beginnings of progress along these lines have been made. Visual speech information activates some areas of the temporal cortex also activated by auditory speech [5,13,15]. However, much work is needed to understand fully the neural substrate for AV speech processing. Sekiyama [this session] presents brain imaging results that suggest differential cortical responses when the auditory part of AV speech is more or less intelligible.

The cortical location of AV speech perception is addressed by Ponton et al. [this session]. They outline the neuroanatomy of the A and V cortical pathways for AV speech. They present possible explanations for AV speech interactions from the perspective of the neurophysiology. In particular, they focus on explanations for cortical AV speech enhancement effects. One observation they make is that neuroanatomy affords several opportunities for interaction at several different levels of speech stimulus processing.

One question is whether perceptual integration occurs initially in infants involving the same cortical and subcortical areas as it does in adults. While some of the data on the development of different types of AV perception in infancy may throw some light on this issue [see 10, and Lewkowicz, this session], these issues remain relatively unexplored.

2.3. *How* does AV speech combine perceptually?

In 1987, Summerfield asserted that, "Any comprehensive account of how speech is perceived should encompass AV speech perception." Summerfield's essay [52] provided a systematic exploration of possible mechanisms for AV speech integration. His seminal thinking has been carried forward in numerous studies on how auditory, visual, and even tactile information combine [e.g., 24,28,36,46]. For example, Green and Miller [29] built upon the finding that voice onset time identification crossovers for a /bi-pi/ continuum are influenced by vowel duration. They showed this effect when the temporal factor was varied only in the visual stimuli. This finding was interpreted as evidence for a common metric explanation for AV speech perception, consistent with Summerfield's overall conclusions.

Another approach to understanding how A and V information combine perceptually has been to study relationships between the respective speech signals. Bateson and colleagues initiated important studies on intermodal dependencies. Bateson [this session] summarizes their findings on the physiological and functional mappings between head and face motions and components of the speech acoustics (F0, amplitude, and spectral properties). This work has shown that A and V speech afford a coherent stimulus with predictable relationships between signals [cf., Bailly, this session]. Research is going beyond such initial demonstrations to assess specific aspects of the contribution of the face and head to AV speech perception, including the spatial frequencies of visual speech, and talker-specific visual motion.

Rosenblum [this session] reviews explanations for AV speech perception from the ecological psychology perspective. Rosenblum's experiments, including ones using novel point-light stimuli have emphasized the primacy of multimodal perception and its generality across speech versus non-speech stimuli. His research has emphasized the importance of viewing speech signals in terms of their articulatory origin: AV speech perception is the perception of articulatory gestures, whose acoustic and optical signal properties are perceived in an amodal format.

Massaro [36] has extensively studied how the Fuzzy Logical Model of Speech Perception (FLMP) can explain AV speech perception. His is an information processing model with a computational implementation. He proposes a three-part model comprising independent modality-specific feature evaluation, integration, and decision. Feature evaluation is conceptualized as a process in which sensory/perceptual systems evaluate modality-specific stimulus features versus ideal features. Feature evaluation is continuous (values between 0 and 1) and varies with the stimulus. Feature evaluation is independent within modalities, and feature values are integrated multiplicatively during integration to determine the extent to which category identification is supported by the stimulus. The selection of a particular perceptual category, the response decision, follows the integration using a decision rule that weights the perceptual evidence. The integration process is generic to all types of feature combinations.

Developmental research with infants can contribute to explaining *how* A and V speech information combines. Lewkowicz [this session] reports results that suggest that there is a developmental time course for the mechanism(s) responsible for processing AV speech relationships. An early studied showed that by 3 months infants can demonstrate sensitivity to the correspondence between A and V speech [33]. Auditory speech facilitates infants' face perception [7]. These in turn are influenced by the ambient language [10], implying that developmental experience is a factor in AV how AV perception develops. However, 3- 5- and 7-month-old full-term, but not pre-term, infants are sensitive to the voice-mouth correspondence [41], implying maturation also.

3. MULTISENSORY SPEECH APPLICATIONS

Engineers interested in machine speech processing would like to know *when*, *where*, and *how* acoustic and optical signals can be combined to accomplish various speech processing tasks. Algorithmic solutions need not be constrained by human neural architecture.

3.1. AV speech recognition, enhancement and compression

Beginning with Petajan [40], AV Speech Recognition (AVSR) has become a major challenge for robust speech recognition and human-machine dialog [see, Deligne et al., this session; Heckmann et al., Lucey et. al, Rogozan, Wigger et al., and Wojdel et al., this

meeting]. Thanks to the complementarity and redundancy of the visual channel, it appears that AVSR outperforms ASR for degraded speech and also Lombard [30] and impaired speech [42]. Apart from the design of the visual front-end (detection and characterization of visual cues), which is an important problem, the basic problem for realizing AVSR systems is the nature of the fusion mechanisms discussed below in terms of our basic questions.

Where refers to the fusion level: here, late vs. early fusion generally implies decision vs. feature fusion. In the first case, two separate classifiers make decisions, respectively, on the A and V channels, and the decisions are then fused. In the second case, both channels must be integrated in some way, to provide input to a single classifier. This distinction is reminiscent of human speech perception [47]. Several proposals have been made recently to apply notions from early integration to audio speech enhancement. Feature fusion might lead to being able to estimate improved audio features from combined degraded audio + video [26, Deligne et al., and Sodoyer et al., this meeting]. This could provide a technological counterpart of the "very early" AV scene analysis process demonstrated by Schwartz et al. [this meeting]. Lastly, AV redundancy can also be exploited for predicting the image from the sound [reviewed in 6], the sound from the image [54], or compressing AV stimuli in telecommunication applications [25].

How to combine information involves basically the choice of an optimal fusion process able to deal efficiently with the "quality" of information [43]. While Massaro's FLMP [36] provides a cognitive basis to the Bayesian multiplicative fusion model applied to separate classifiers, a number of proposals have suggested weighting the audio and video channels in relation to control parameters such as SNR.

The fusion problem also must confront the *when* question, because the A and V channels are in some respects asynchronous: Face motion commonly anticipates some audible speech gestures [1]. Synchronizing the flow of decisions is therefore a complex and important problem, and a number of proposals have been made in the framework of the preferred statistical machine in ASR and AVSR [43].

3.2. AV speech synthesis, talking heads and cued speech

In the field of speech synthesis, the realism of talking faces is gradually being improved [2; Cosi et al., and Zelezny et al., this meeting]. In this session, Bailly presents the "ground truths" needed to develop efficient models. He distinguishes between model-based and image-based approaches, and he advocates for data-driven techniques that incorporate a priori knowledge of the articulatory degrees-of-freedom of speech organs.

Evaluation of synthesis could involve intelligibility of speech in noise. To date, intelligibility of talking heads is still near half the intelligibility benefit provided by natural faces [4]. Evaluation could exploit the McGurk effect. In both cases, the quality of both lip shapes and movements seems of primary importance: poor quality leads to little intelligibility gain, while natural dynamics do compensate from lack of shape information [46].

Talking heads probably should also be provided with hands for manual communication systems used by some deaf individuals [34]. Following Duchnowski et al. [23], Attina et al. [this meeting] present an articulatory study describing the lip-hand co-ordinations in French cued speech. Lastly, talking heads need emotions to be smarter communication agents and also for "well-formed" internal states. Emotional multisensory agents are a major goal for the coming years.

4. HUMAN SOCIAL AND CLINICAL APPLICATIONS

AV speech recognition, synthesis and processing will probably be increasingly introduced into dialog systems, mobile and internet applications, including entertainment, e-learning, etc.

Kirk et al. [this session] present research on children and adults with cochlear implants. A measure of AV enhancement was calculated relative to the maximum possible performance in the auditory condition. Among results were that children who were better at recognizing isolated words through A alone also were better at combining AV stimulation and also had superior speech intelligibility. Adults with cochlear implants were found to be better perceivers in the visual condition and to take more advantage of visual information in the AV condition. Kirk et al. conclude that training programs for individuals with impaired hearing could profitably benefit from emphasis on AV speech. Clearly, visible speech is also an important input for individuals with less significant hearing losses than those of individuals with cochlear implants.

AV speech is important for language learning and inter-language communication. Hazan et al. (this session) presents a number of results on AV, visual, and auditory perception of L2 contrasts, showing a lack of effect for visual cues to L2 contrasts that are non-contrastive in L1, e.g., labiodental identification by Spanish learners of English, even in the lipreading-alone condition where attention should be focused on the visual modality [see also Erdener and Burnham, this meeting]. This study suggests some barriers to use of AV speech.

In contrast, Hardison [this session] reports on findings of a three-week perceptual training experiment involving Japanese and Korean learners of English. Variability of speech stimuli was a major factor in the training, which involved auditory and AV speech. Results showed greater improvement with AV training. Several transfer effects were obtained, including ones for novel stimuli and a new talker.

5. CONCLUSIONS

It is fascinating to consider how AV speech has progressively "invaded" many aspects of the study of speech communication, renewed or enhanced a number of paradigms, models and applications, and entered into core modern concepts about human speech. We anticipate that this will only increase in the future as the tools and concepts needed to understand, manipulate, and build complex systems increase.

6. REFERENCES

[1] Abry, C., et al., in [53].

[2] Bailly, G., et al., in [53].

[3] Benoît, C., et al., [50].

[4] Bernstein, L. E., C. W. Ponton, and E. T. Auer, Jr., "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex," *Proceedings of AVSP'01*, 50-55, 2001.

[5] Bernstein, L. E., E. T. Auer, Jr., J. K. Moore, C. W. Ponton, M. Don, and M. Singh, "Visual speech perception without primary auditory cortex activation," *NeuroReport*, 13:311-315, 2002.

[6] Brooke, M., in [53].

[7] Burnham, D., "Visual recognition of mother by young infants: facilitation by speech," *Perception*, 22:1133-1153, 1993.

[8] Burnham, D., "Language specificity in the development of auditory-visual speech perception," pp. 27-60 in [16].

[9] Burnham, D., and B. Dodd, in [50].

[10] Burnham, D., and B. Dodd, in *Adv. Infancy Res.,* (C. Rovee-Collier, Ed.) 12:170-187, 1998.

[11] Burnham, D., S. Lau, H. Tam, and C. Schoknecht, "Visual Discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers," *Proceedings of AVSP'01*, 155-160, 2001.

[12] Callan, D., A. Callan, and E. Vatikiotis-Bateson, "Neural areas underlying the processing of visual speech information under conditions of degraded auditory information," *Proceedings of AVSP'01*, 45-49, 2001.

[13] Calvert, G. A., R. Campbell, and M. J. Brammer, "Evidence from functional magnetic resonance imaging of crossmodal binding the human heteromodal cortex," *Current Biology*, 10:649-657, 2000.

[14] Campbell, R., in [50].

[15] Campbell, R., "How brains see speech: the cortical localisation of speechreading in hearing people," pp. 177-194 in [16].

[16] Campbell R., B. Dodd, and D. Burnham (Eds.), *Hearing by Eye (II): The Psychology of Speechreading and Auditory-visual Speech*, East Sussex, UK: Psychology Press, 1998.

[17] Cavé, C. et al., *Proceedings of ICSLP'96*, 2175-2178, 1996.

[18] Colin, C., M. Radeau, and P. Deltenre, "The mismatch negativity (MMN) and the McGurk effect," *Proceedings of AVSP'01*, 56-61, 2001.

[19] DeGelder, B., and J. Vroomen, "The perception of emotions by ear and eye," *Cognition and Emotion*, 14:289-31, 2000.

[20] Dodd, B., and R. Campbell, (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*, Lawrence Erlbaum Associates, London, 1987.

[21] Dodd, B., B. McIntosh, and L. Woodhouse, "Early lipreading ability and speech and language development of hearing-impaired pre-schoolers," pp. 229-242 in [16].

[22] Driver, J., "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," *Nature*, 381:66-68, 1996.

[23] Duchnowski, P., L. Braida, D. Lum, M. Sexton, J. Krause, and S. Banthia, "Automatic generation of cued speech for the deaf: status and outlook," *Proceedings of AVSP'98*, 161-166, 1998.

[24] Fowler, C. A., and D. J. Dekle, "Listening with eye and hand: cross-modal contributions to speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, 17:816-828, 1991.

[25] Girin, L. et al., *Proceedings of IEEE Workshop on Multimedia Signal Processing*, 1998.

[26] Girin, L., J. L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. Am.*, 109:3007-3020, 2001.

[27] Grant, K. W., and P. F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *J. Acoust. Soc. Am.*, 108:1197-1208, 2000.

[28] Green, K., "The use of auditory and visual information during phonetic processing: implications for theories of speech perception," pp. 3-26 in [16].

[29] Green, K. P, and J. L. Miller, "On the role of visual rate information on phonetic perception," *Perception & Psychophysics*, 38:269-276, 1985.

[30] Huang and Chen, *Proceedings of IEEE Workshop on Multimedia Signal Processing*, 2001.

[31] Jiang, J., A. Alwan, P. Keating, E. T. Auer, Jr., and L. E. Bernstein, "On the correlation between face movements, tongue movements, and speech acoustics," *EURASIP Journal on Applied Signal Processing,* in press.

[32] Johnson, J. A., L. D. Rosenblum, and M. A. Schmuckler, *J. Acoust. Soc. Am.*, 97:3286, 1995.

[33] Kuhl, P. K., and A. N. Meltzoff, "The bimodal perception of speech in infancy," *Science*, 218:1138-1141, 1982.

[34] Leybaert, J., J. Alegria, C. Hage, and B. Charlier, "The effect of exposure to phonetically augmented lipspeech in the prelingual deaf," pp. 283-301. in [16].

[35] Lisker, L., and M. Rossi, "Auditory and visual cueing of the [+/-rounded] feature of vowels," *Language and Speech*, 35:391-417, 1992.

[36] Massaro, D.W., *Speech Perception by Ear and Eye,* London: Laurence Erlbaum Associates, 1987.

[37] McGurk H., and J. MacDonald, "Hearing lips and seeing voices," *Nature*, 264:746-748, 1976.

[38] Mills, A. E., "The development of phonology in the blind child," pp. 145-161 [20].

[39] Mulford, R., in *The Emergent Lexicon,* (M. D. Smith, and J. L. Locke, Eds.), NY: Academic Press, pp. 293-338, 1998.

[40] Petajan, E. D., Doctoral Thesis, Univ. of Illinois, 1984.

[41] Pickens, J., et al., *Infant Behav. Devel.*, 17: 447-455, 1994.

[42] Potamianos, G., and C. Neti, "Automatic speechreading of impaired speech," *Proceedings of AVSP'01*, 177-182, 2001

[43] Potamianos, G., et al., in [53].

[44] Radeau, M., "Auditory-visual spatial interaction and modularity," *Current Psychology of Cognition*, 13:3-51, 1994.

[45] Robert-Ribes, J., et al., in [50].

[46] Rosenblum L. D., and H. M. Saldana, "An audiovisual test of kinematic primitives for visual speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, 22:318-331, 1996.

[47] Schwartz, J.-L., J. Robert-Ribes, and P. Escudier, "Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception," pp. 85-108 in [16].

[48] Sekiyama K., and Y. Tohkura, "McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," *J. Acoust. Soc. Am.*, 90:1797-1805, 1991.

[49] Stoel-Gammon, C., "Prelinguistic vocalizations of hearing-impaired and normally hearing subjects: a comparison of consonantal inventories," *Journal of Speech and Hearing Disorders*, 53:302-315, 1988.

[50] Stork, D., and M. Hennecke, (Eds.), *Speechreading by Humans and Machines*, NATO ASI Series. Springer, 1996.

[51] Sumby, W. H., and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.*, 26:212-215, 1954.

[52] Summerfield, Q., "Some preliminaries to a comprehensive account of audio-visual speech perception," pp 3-52 in [20].

[53] Vatikiotis-Bateson, E., et al., (Eds.), *Audiovisual Speech Processing*. MIT Press, in press.

[54] Yehia, H., P. Rubin, and E. Vatikiotis-Bateson, "Using speech acoustics to drive facial motion," *Speech Communication*, 26:23-43, 1998.