THE PERCEPTUAL BASIS FOR AUDIOVISUAL SPEECH INTEGRATION

Lawrence D. Rosenblum

Department of Psychology
University of California, Riverside
rosenblu@citrus.ucr.edu

ABSTRACT

One question central to understanding the perceptual basis of audiovisual speech integration concerns *where* in the process the audio and visual streams are combined. Over 20 years of research on this question have provided answers ranging from the integration occurring at the level of the informational input, to occurring only after segment matches are made independently for each modality. In this paper I will present a modality-neutral account of audiovisual speech integration. From this account, speech perception is inherently amodal, and the information for speech is considered kinematic primitives that can be instantiated in any modality. Modality is considered to be invisible to the speech function, and 'integration' occurs as a function of the input information itself. Four classes of support for a modality-neutral account will be presented including a) the primacy/ubiquity of multimodal speech; b) evidence for very early integration from the behavioral and c) neuropsychological data; and d) evidence for informational similitude existent across modalities.

1. INTRODUCTION

Theories of audiovisual speech integration have varied widely on the question of where the informational streams are combined. The theories have ranged from proposing that integration occurs at the informational input [1]; before feature extraction [2]; after feature extraction [3]; and after segment, or even word recognition [4]. This paper will present a view which posits that in an important sense, the integration occurs at the level of information, even before it enters the senses.

2. THE MODALITY-NEUTRAL ACCOUNT

The modality-neutral account of audiovisual speech perception follows directly from Summerfield's [1] proposition that the informational metric taken at the point of speech integration is best construed as an articulatorily-based, modality-independent form. In its current version, the modality-neutral account also borrows from theories of information and multimodal perception outside the domain of speech [5,6]. Briefly stated, speech information in the form of higher-order kinematic patterns, is thought to be instantiated as structure in multiple arrays (visual; auditory; haptic). In that the speech function is concerned with the information and not the energy array in which it is available, modality can be said to be invisible to speech perception. Furthermore, 'cross-modal' speech integration is not something that occurs in the perceiver, but occurs in—*and as a property of*—the information itself [5,6,7]. This renders the job of a sense organ to one of extracting this higher-order structure as it is instantiated in the energy range to which the organ is sensitive [6].

For the modality-neutral account to be viable, audiovisual speech perception must show at least four properties. First, if the speech perception function is designed for modality-neutral speech, then integration effects should be ubiquitous and observed even in observers with little experience. Secondly, if integration occurs as a function of the external information, then there should be evidence for cross-modal influences at the earliest observable stage (e.g., features). Thirdly, the modality-neutral account would predict neuropsychological mechanisms that are sensitive to cross-modal input at a relatively peripheral level. Finally, critical to a modality-neutral account is evidence for information similitude: i.e., salience of the same informational characteristics across modalities. The first three of these properties of audiovisual speech are discussed in the next section which is followed by a more detailed discussion of information similitude.

2.1 The ubiquity of multimodal speech perception

It is proposed that multimodal speech is the primary mode of speech perception: Visual speech perception is not simply piggybacked on the auditory speech function. This primacy of multimodal speech is supported by the vast research on cross-modal syllable integration. In the McGurk effect [8], visual speech information (e.g., lipread /va/) can integrate with, and even override auditory speech information (/ba/) so perceivers report 'hearing' what they see (a heard /va/). This effect works on observers who are informed that the audio and visual components are discrepant; when the audio and visual components are derived from speakers of different gender [9]; and even when subjects are unaware they are seeing a face [10]. The McGurk effect works on observers with various language backgrounds [11], as well as with pre-linguistic infants [12]. Finally, influences on 'heard' speech can be induced by articulatory information that is felt (through touching the face) rather than seen [13]. In that the subjects in this last study had rarely, if ever, attempted to perceive speech in this way, these findings add support for the primacy of multimodal speech integration.

2.2 Cross-modal influences occur very early in the process

Much of the research on multimodal speech offers evidence that audiovisual speech is integrated at a very early stage of the perceptual process [2]. Much of this research was conducted by Kerry Green and his colleagues. First, Green provides evidence that cross-modal influences occur at least by the stage of featural extraction. He showed that both visually perceived rate and place of articulation can influence the interpretation of the auditory feature of voice onset time (VOT) [14, 15]. These influences take a pattern similar to the influences induced by analogous unimodal auditory changes.

Green also provided evidence for cross-modal context effects in which low-level coarticulatory influences appear. In the auditory speech literature it has long been known that a mismatch between a vowel and adjacent consonant, created through acoustic editing, will influence the interpretation of the consonant [e.g., 16]. Green and Gerdeman [17] were able to show that analogous influences on perceived consonants can occur when a vowel mismatch is established cross-modally through discrepant audiovisual information. Relatedly, Green and Norrix [18] have shown that a coarticulatory phonetic context established solely in one modality can influence featural extraction in the other modality. To explain these results, Green [2] suggests that audiovisual information is integrated early enough to influence the perceived coarticulatory state which, in turn, influences the interpreted adjacent segments.

This evidence for cross-modal influences at the level of featural and coarticulatory extraction is consistent with a modality-neutral account. The evidence shows a conflux of cross modal information at the earliest stage in the speech recognition process observable with a perceptual methodology.

2.3 The brain treats visual speech like auditory speech

Recent fMRI research has revealed that a silent lipreading task can induce primary auditory cortex activity similar to that induced by auditory speech [19, 20; but see 21]. These findings augment earlier results which used a mismatched negativity methodology to show that changes in visual speech information can change auditory cortex activity during audiovisual integration [22]. These findings provide neurophysiological support that visual speech has a very early influence on heard speech. More strongly, the findings could be interpreted as evidence that in an important way, modality is invisible to the speech perception function, even at a relatively early level.

It should be mentioned that this research has not been without controversy. Bernstein and her colleagues [21] have reported recent evidence that visual speech does not induce primary auditory cortex under all conditions. Clearly more research is needed to clarify these potentially provocative findings.

3. INFORMATIONAL SIMILITUDE

In his seminal paper, Summerfield [1] sketched an example of modality-neutral speech information. This information took the form of kinematic patterns that specified an oscillatory articulatory dynamics. More recently a literature has emerged showing close correspondences between visible speech and acoustic information [e.g., 23]. This research has shown strong correlations between visible speech kinematics and acoustic properties such as amplitude (rms) and spectral composition. This research is discussed in detail elsewhere in this volume [24].

The following sections will discuss work from our lab showing that important characteristics of the *general* nature of auditory speech information are also found for visual speech.

3.1 Salience of time-varying information

Research on both auditory and visual speech has shown that isolated time-varying aspects of the signals provide important speech information. In auditory speech research, there is evidence that signals which do not involve the traditional cues of formants, transitions, and noise bursts can still be understood as speech [25]. These signals are composed of a set of sine-waves synthesized to track the pitch and amplitudes of the center formant frequencies of an utterance. This sinewave speech can be understood well enough for listeners to transcribe sentences [25]. It is thought that sinewave speech is effective in isolating the time-varying dimensions of the signal.

With regard to visual speech, work in our laboratory has shown that isolated time-varying *visual* information for articulation is also salient [10, 26]. For these demonstrations, a point-light technique is implemented for which small illuminated dots are affixed to a darkened face. The face is then filmed speaking in the dark so that only the dots and their movements are visible in the resultant stimuli. Research has shown that while these images contain no standard facial features, they do provide visual speech information to the degree that they can enhance auditory speech in noise, and integrate with auditory speech in a McGurk-effect paradigm [10, 26]. Thus isolated time-varying articulatory information conveyed either visually or auditorily supports speech perception.

There is also research in both domains showing greater relative salience of time-varying over static speech information. In auditory speech, it seems that the parts of the signal that are least changing (vowel nuclei; consonantal burst targets) are less informative than parts that are more dynamic and influenced by coarticulation [27, 28]. For example, much of the vowel nucleus of a CVC syllable can be deleted without hindering vowel identification judgments [27].

Research in our laboratory has shown analogous findings for visual speech [29]. We have found evidence that the most salient parts of visible vowels presented in a CVC context lie at the most coarticulated portions of the syllables. Similar to auditory speech, much of the visual vowel nucleus can be deleted without a loss in vowel identification. Thus, the research on both auditory and visual speech suggests that not only is time-varying information useful, it might in fact be the most relevant dimension for speech perception. Potentially then, the speech system is sensitive to modality-neutral time-varying properties available across modalities.

3.2 Influences of indexical information

Cross-modal information similitude is also apparent in the relationship between speech and speaker properties. Auditory speech research has revealed that indexical properties of an utterance (those associated with specific speakers), play an important role in phonetic recovery [30, for a review]. There is evidence that speaker-specific information can facilitate linguistic recovery in the contexts of single vs. multiple speaker lists [31], word naming and identification in noise [32], recognition memory [33], and form-based priming [34]. Furthermore, there is evidence that speaker-specific phonetic (idiolectic) properties of the speech signal can be used for speaker recognition. Remez and his colleagues [30] used sinewave speech re-synthesis to isolate the phonetic dimensions of speakers' sentences (deleting voice quality; fundamental frequency). They found that listeners could recognize speakers from these stimuli in both matching and identification contexts. They argue that the observed contingencies of speech perception on speaker information might be based on the recovery of idiolectic dimensions used for both function.

Speaker information has also been shown to influence recovery of *visual* speech in an analogous way. Visual speaker information can influence visible vowel identification [35], single vs. multiple speaker lists [36], memory for lipread words [37], and the McGurk effect [38]. Relatedly, stimulus manipulations known to disproportionately disrupt face recognition can also inhibit visual speech perception [39, 40]. Finally, in experiments analogous to those of Remez, et al. [10] we have shown that isolated visible articulatory information can be used for speaker recognition [41, 42]. Using the point-light technique to isolate visible articulation, we found that observers can recognize speakers from these stimuli in both matching and identification tasks. Proposing that observers perform this task using the idiolectic information available in visible speech, we tested whether the analogous information in auditory speech would support cross-modal matching. We found that observers can make speaker matches from auditory speech sentences to articulating point-light faces with some skill [43]. Potentially then, observers are sensitive to modality-neutral idiolectic properties available in both the auditory and visual signals.

4. CONCLUSIONS

The four properties of multimodal speech discussed above by no means provide exclusive support for the modality-neutral account. They are simply provided as a partial list of minimal criteria for the account to be true. Evidence against these properties would provide evidence against the modality-neutral account [e.g., 4]. At the same time, other properties of multimodal speech would be expected from the modality-neutral account including cross-modal transfer of training (e.g., to specific speakers), and that an intra-stimulus switch of modality should have negligible interfering effects on speech recovery. Tests of these and other predictions will further our understanding of where in the process the audio and visual streams are combined.

5. REFERENCES

[1] Summerfield, Q. "Some preliminaries to a comprehensive account of audiovisual speech perception." In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53-83). London: Lawrence Erlbaum Associates, Inc., 1987.

[2] Green, K. P. "The use of auditory and visual information during phonetic processing: Implications for theories of speech perception." In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech* (pp. 3-25). London, England: Lawrence Erlbaum Assoc. 1998.

[3] Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Assoc, Inc.

[4] Bernstein, L.E., Auer, E.T., Moore, J.K. "Modality-Specific Perception of Auditory and Visual Speech". To appear in B. Stein (Ed.) *The Handbook of Multimodal Processing.* In press.

[5] Gibson, J. J. *The ecological approach to visual perception*. Boston: Houghton-Mifflin, 1979.

[6] Stoffregen T. A. & Bardy, B. G. "On specification and the senses." *Behavioral & Brain Sciences, 24(2)*, 195-261. 2001.

[7] Rosenblum, L.D. & Gordon, M.S. "The generality of specificity: Some lessons from audiovisual speech." *Behavioral & Brain Sciences, 24(2)*, 239-240. 2001

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

[9] Green, K. P., Kuhl, P. K., Meltzoff, A. M., & Stevens, E. B. "Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect" *Perception and Psychophysics, 50,* 524-536. 1991.

[10] Rosenblum, L.D. and Saldaña, H.M. (1996). "An audiovisual test of kinematic primitives for visual speech perception." *Journal of Experimental Psychology: Human Perception and Performance. 22(2),* 318-331. 1996.

[11] Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics, 21,* 445-478.

[12] Rosenblum, L.D., Schmuckler, M.A., & Johnson, J.A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59(3),* 347-357.

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

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

[15] Green, K. P., & Kuhl, P. K. "The role of visual information in the processing of place and manner features in speech perception." *Perception & Psychophysics, 45,* 34-42. 1989.

[16] Whalen, D.H. "Effects of vocalic formant transitions and vowel quality on the English [s]-[S] boundary." *Journal of the Acoustical Society of America, 69, 275-282*. 1981.

[17] Green, K.P. & Gerdman, A. "Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels." *Journal of Experimental Psychology: Human Perception and Performance,* 21, 1409-1426. 1995.

[18] Green, K.P. & Norrix, L. "The Perception of /r/ and /l/ in a Stop Cluster: Evidence of Cross-Modal Context Effects". *Journal of Experimental Psychology: Human Perception & Performance. 27 (1):* p. 166-177. 2001.

[19] Calvert, G.A., Bullmore, E., Brammer, M.J., Campbell, R., Iversen, S.D., Woodruff, P., McGuire, P., Williams, S., David, A.S. "Silent lipreading activates the auditory cortex." *Science, 276,* 593-596. 1997.

[20] MacSweeney, M., Amaro, E., Calvert, G. A. & Campbell, R. "Silent speechreading in the absence of scanner noise: An event-related fMRI study." *Neuroreport: For Rapid Communication of Neuroscience Research, 11,* 1729-1733. 2000.

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

[22] Sams, M., Aulanko, R., Hamalainen, M., Hari., R., Lounasmaa, O.V., Lu, S-T., & Simola, J. (1991). "Seeing speech: visual information from lip movements modifies activity in the human auditory cortex." *Neuroscience Letters, 127,* 141-145. 1991.

[23] Munhall, K.G. & Vatikiotis-Bateson, E. "The moving face during speech communication." In R. Campbell & B. Dodd (Eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech* (pp. 123-139). London, England: Lawrence Erlbaum Assoc. 1998.

[24] Vatikiotis-Bateson, E. "The stimulus as basis for audiovisual integration." *Proceedings of the 7th International Conference on Spoken Language Comprehension.* 2002.

[25] Remez, R., Rubin, P., Pisoni, D., & Carrell, T. "Speech perception without traditional speech cues." *Science, 212,* 947-950. 1981.

[26] Rosenblum, L.D., Johnson, J. A., & Saldaña, H.M. "Visual kinematic information for embellishing speech in noise." *Journal of Speech and Hearing Research 39(6),* 1159-1170. 1996.

[27] Jenkins, J. J., Strange, W. & Edman. T. R. "Identification of vowels in "vowelless" syllables." *Perception & Psychophysics, 34,* 441-450. 1983.

[28] Strange, W., Jenkins, J. J. & Johnson, T. L. "Dynamic specification of coarticulated vowels." *Journal of the Acoustical Society of America, 74,* 695-705. 1983.

[29] Yakel, D. A. *Effects of time-varying information on vowel identification accuracy in visual speech perception.* Doctoral dissertation, University of California, Riverside, USA. 2000.

[30] Remez, R. E., Fellowes, J. M. & Rubin, P. E. "Talker identification based on phonetic information." *Journal of Experimental Psychology: Human Perception & Performance 23 (3)*, 651-666. 1997.

[31] Mullennix, J. W., Pisoni, D. B., & Martin, C. S. Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85(1),* 365-378. 1989.

[32] Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. "Speech perception as a talker-contingent process." *Psychological Science, 5(1),* 42-46. 1994.

[33] Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. "Episodic encoding of voice attributes and recognition memory for spoken words." *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(2),* 309-328. 1993.

[34] Saldaña, H. M., & Rosenblum, L. D. "Voice information in auditory form-based priming." *Journal of the Acoustical Society of America, 95(5),* 2870. 1994.

[35] Schweinberger, S.R. & Soukup, G.R. "Asymmetric relationships among perceptions of facial identity, emotion, and facial speech." *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1748-1765. 1998.

[36] Yakel, D.A., Rosenblum, L.D., & Fortier, M.A. "Effects of talker variability on speechreading." *Perception & Psychophysics, 62,* 1405-1412. 2000.

[37] Sheffert, S.M. & Fowler, C.A. (1995). "The effects of voice and visible speaker change on memory for spoken words." *Journal of Memory and Language, 34,* 665-685. 1995.

[38] Walker, S., Bruce, V., & O'Malley, C. "Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect." *Perception & Psychophysics, 57,* 1124-1133. 1995.

[39] Jordan, T.R., & Bevan, K. "Seeing and hearing rotated faces: Influences of facial orientation on visual and audio-visual speech recognition." *Journal of Experimental Psychology: Human Perception and Performance., 23,* 388-403. 1997.

[40] Rosenblum, L.D., Yakel, D.A., & Greene, K.G. "Face and mouth inversion affects on visual and audiovisual speech perception." *Journal of Experimental Psychology: Human Perception and Performance. 26(3),* 806-819. 2000.

[41] Rosenblum, L.D., Yakel, D.A., Baseer, N., Panchal, A., Nordarse, B.C. & Niehus, R.P. "Visual speech information for face recognition." *Perception & Psychophysics, 64(2),* 220-229. 2002.

[42] Rosenblum, L.D.,., & Smith, N. "Look who's talking: Recognizing friends from visible articulation." To be submitted to *Psychological Science.* In preparation.

[43] Rosenblum, L.D., Nichols, S, Lee, J., "Matching voices to visible speech movements: Evidence for cross-modal idiolectic information." To be submitted to *Journal of Experimental Psychology: Human Perception & Performance.* In preparation.