

Speech Perception

Lawrence D. Rosenblum

Speech Perception as a Multimodal Phenomenon

Lawrence D. Rosenblum

*University of California, Riverside*

Address correspondence to Lawrence D. Rosenblum, Department of Psychology,

University of California, Riverside, Riverside, CA 9252; e-mail:

rosenblu@citrus.ucr.edu.

**ABSTRACT**—Speech perception is inherently multimodal. Visual speech (lip-reading) information is used by all perceivers and readily integrates with auditory speech. Imaging research suggests that the brain treats auditory and visual speech similarly. These findings have led some researchers to consider that speech perception works by extracting amodal information that takes the same form across modalities. From this perspective, speech integration is a property of the input information itself. Amodal speech information could explain the reported automaticity, immediacy, and completeness of audiovisual speech integration. However, recent findings suggest that speech integration can be influenced by higher cognitive properties such as lexical status and semantic context. Proponents of amodal accounts will need to explain these results.

**KEYWORDS**—speech; audiovisual; multimodal; lip reading

We all read lips. We read lips to better understand someone speaking in a noisy environment or speaking with a heavy foreign accent (for a review, see Rosenblum, 2005). Even with clear speech, reading lips enhances our comprehension of a speaker discussing a conceptually dense topic. While wide individual differences exist in lip-reading skill, evidence suggests that all sighted individuals from every culture use *visual speech information*. Virtually any time we are speaking with someone in person, we use information from seeing movement of their lips, teeth, tongue, and non-mouth facial features, and have likely been doing so all our lives. Research shows that, even before they can speak themselves, infants detect characteristics of visual speech, including whether it corresponds to heard speech and contains one or more language. Infants, like adults, also automatically integrate visual with auditory speech streams.

Speech perception is inherently *multimodal*. Despite our intuitions of speech as something we hear, there is overwhelming evidence that the brain treats speech as something we hear, see, and even feel. Brain regions once thought sensitive to only auditory speech (primary auditory cortex, auditory brain stem), are now known to respond to visual speech input (Fig. 1; e.g., Calvert et al., 1997; Musacchia, Sams, Nicol, & Kraus, 2005). Visual speech automatically integrates with auditory speech in a number of different contexts. In the McGurk effect (McGurk & MacDonald, 1976), an auditory speech utterance (e.g., a syllable or word) dubbed synchronously with a video of a face articulating a discrepant utterance induces subjects to report “hearing” an utterance that is influenced by the mismatched visual component. The “heard” utterance can take a form in which the visual information overrides the auditory (audio “ba” + visual “va” = heard

“va”) or in which the two components fuse to create a new perceived utterance (audio “ba” + visual “ga” = heard “da”).

Even *felt* speech, accessed either through touching a speaker’s lips, jaw, and neck or through the kinesthetic feedback from one’s own speech movements, readily integrates with heard speech (e.g., Fowler & Dekle, 1991; Sams, Mottonen, & Sihvonen, 2005). The former of these effects occurs with naïve subjects who have no experience perceiving speech through touch. This finding suggests that our skill with multimodal speech perception is likely based not on learned cross-modal associations but, rather, on a more ingrained sensitivity to lawfully structured speech information.

It is also likely that human speech evolved as a multimodal medium (see Rosenblum, 2005, for a review). Most theories of speech evolution incorporate a critical influence of visuofacial information, often bridging the stages of manuo-gestural and audible language. Also, multimodal speech has a traceable phylogeny. Rhesus monkeys and chimpanzees are sensitive to audible–facial correspondences of different types of calls (alarm, coo, hoot). Brain imaging shows that the neural substrate for integrating audiovisual utterances is analogous across monkeys and humans (Ghazanfar, Maier, Hoffman, & Logothetis, 2005). Finally, there is speculation that the world’s languages have developed to take advantage of visual as well as auditory sensitivities to speech. Languages typically show a complementarity between the audibility and visibility of speech segments such that segment distinctions that are harder to hear (“m” vs. “n”) are easier to see and vice versa.

Together, these findings suggest a *multimodal primacy* of speech. Nonauditory recognition of speech is not simply a function piggybacked on auditory speech

perception; instead the relevant operations and associated neurophysiology of speech are likely designed for multimodal input. The multimodal primacy of speech is consistent with recent findings in general perceptual psychology showing the predominance of cross-modal influences in both behavioral and neurophysiological contexts (Shimojo & Shams, 2001, for a review). This has led a number of researchers to suggest that the perceptual brain is designed around multimodal input. Audiovisual speech is considered a prototypic example of the general primacy of multimodal perception, and the McGurk effect is one of the most oft-cited phenomena in this literature. For these reasons, multimodal speech research has implications that go well beyond the speech domain.

### **AMODAL THEORIES OF MULTIMODAL SPEECH PERCEPTION**

Findings supporting the primacy of multimodal speech have influenced theories of speech integration. In “amodal” or “modality neutral” accounts, speech perception is considered to be blind to the modality specifics of the input from the very beginning of the process (e.g., Rosenblum, 2005). From this perspective, the physical movements of a speech gesture can shape the acoustic and optic signals in a similar way, so that the signals take on the same overall form. Speech perception then involves the extraction of this common, higher-order information from both signals, rendering integration a consequence and property of the input information itself. In other words, for the speech mechanism, the auditory and visual information is functionally never really separate. While the superficial details of the acoustic and optic signals—along with the associated peripheral physiology—are distinct, the overall informational form of these signals is the same. This fact would obviate any need for the speech function to translate or actively bind one modality’s information to another’s prior to speech-segment recognition.

Fortunately for speech perception, the optic and acoustic structures most always specify the same articulatory gesture. However when faced with McGurk-type stimuli, amodal speech perception could extract whatever informational components are common across modalities, which could end up either spuriously specifying a “hybrid” segment or a segment closer to that specified in one or the other of the two modalities.

### **SUPPORT FOR AMODAL ACCOUNTS**

Support for amodal accounts comes from the aforementioned evidence for the neurophysiological and behavioral primacy of multimodal speech perception. If the modalities are functionally never separate, then evidence that the system is designed around multimodal input would be expected. For similar reasons, amodal theories predict evidence for an automaticity, completeness, and immediacy of audiovisual speech integration. Support for these predictions has come from research using the McGurk effect (see Rosenblum, 2005, for a review). It turns out that the effect works even when the audio and visual components are made conspicuously distinct by spatial or temporal separation, or by using audio and visual components taken from speakers of different genders. These facts provide evidence for the automaticity of speech integration. The McGurk effect also occurs when subjects are told of the dubbing procedure or are told to concentrate on the audio channel, suggesting that perceivers do not have access to the unimodal components once integration occurs: Integration seems functionally complete.

There is also evidence that audiovisual speech integrates at the earliest observable stage, before phonemes or even phoneme features are determined. Research shows that visible information can affect auditory perception of the delay between when a speaker initiates a consonant (e.g., separating their lips for “b” or for “p”) and when their vocal

chords start vibrating. This *voice-onset time*, is considered a critical speech feature for distinguishing a voiced from a voiceless consonant (e.g., “b” from “p”; Green, 1998). Relatedly, the well-known perceptual compensation of phoneme features based on influences of adjacent phonemes (coarticulation) occurs even if the feature and adjacent phoneme information are from different modalities (Green, 1998). Thus, cross-modal speech influences seem to occur at the featural level, which is the earliest stage observable using perceptual methodologies. This evidence is consistent with neurophysiological evidence that visual speech modulates the auditory brain’s peripheral components (e.g., the auditory brainstem; Musacchia et al., 2005) and supports the amodal theory’s claim that the audio and visual streams are functionally integrated from the start.

### **MODALITY-NEUTRAL SPEECH INFORMATION**

Additional support for amodal theories of speech comes from evidence for similar informational forms across modalities—that is, evidence for *modality-neutral* information. Macroscopic descriptions of auditory and visual speech information reveal how utterances that involve reversals in articulator movements structure corresponding reversals in both sound and light. For example, the lip reversal in the utterance “aba” structures an amplitude reversal in the acoustic signal (loud to soft to loud) as well as a corresponding reversal in the visual information for the lip movements (Summerfield, 1987). Similar modality-neutral descriptions have been applied to quantal (abrupt and substantial) changes in articulation (shifts from contact of articulators to no contact, as in “ba”) and repetitive articulatory motions. More recently, measurements of speech movements on the front of the face have revealed an astonishingly close correlation

between movement parameters of visible articulation and the produced acoustic signal's amplitude and spectral parameters (Munhall & Vatikiotis-Bateson, 2004).

Other research shows how correlations in cross-modal information are perceptually useful and promote integration. It is known that the ability to detect the presence of auditory speech in a background of noise can be improved by seeing a face articulating the same utterance. Importantly, this research shows that the amount of improvement depends on the degree to which the visible extent of mouth opening is correlated with the changing auditory amplitude of the speech (Grant & Seitz, 2000). Thus, cross-modal correspondences in articulatory amplitude facilitate detection of an auditory speech signal.

Perceivers also seem sensitive to cross-modal correlations informative about more subtle articulator motions. Growing evidence shows that articulatory characteristics once considered invisible to lip reading (e.g., tongue-back position, intra-oral air pressure) are actually visible in subtle jaw, lip, and cheek movements (Munhall & Vatikiotis-Bateson, 2004). Also, the prosodic dimensions of word stress and sentence intonation (distinguishing statements from questions), typically associated with pitch and loudness changes of heard speech, can be recovered from visual speech. Even the pitch changes associated with lexical tone (salient for Mandarin and Cantonese), can be perceived from visual speech (Burnham, Ciocca, Lauw, Lau, & Stokes, 2000). These new results not only suggest the breadth of visible speech information that is available but are encouraging that the visible dimensions closely correlated with acoustic characteristics have perceptual salience.

There are other commonalities in cross-modal information that take a more general form. Research on both modalities reveals that the *speaker* properties available in the signals can facilitate speech perception. Whether listening or lip-reading, people are better at perceiving the speech of familiar speakers (Rosenblum, 2005, for a review). For both modalities, some of this facilitating speaker information seems available in the specified *phonetic* attributes: that is, in the auditory and visual information for a speaker's *idiolect* (idiosyncratic manner of articulating speech segments). Research shows that usable speaker information is maintained in auditory and visual stimuli that have had the most obvious speaker information (voice quality and pitch, facial features and feature configurations) removed, but maintain phonetic information. For auditory speech, removal of speaker information is accomplished by replacing the spectrally complex signal with simple transforming sine waves that track speech formants (intense bands of acoustic energy composing the speech signal) (Remez, , Fellowes, & Rubin, 1997). For visual speech, a facial point-light technique, in which only movements of white dots (placed on the face, lips, and teeth) are visible, accomplishes the analogous effect (Rosenblum, 2005). Despite missing information typically associated with person recognition, speakers can be recognized from these highly reduced stimuli. Thus, whether hearing or reading lips, we can recognize speakers from the idiosyncratic way they articulate phonemes. Moreover, these reduced stimuli support *cross-modal speaker matching*, suggesting that perceivers are sensitive to the modality-neutral idiolectic information common to both modalities.

Recent research also suggests that our familiarity with a speaker might be partly based on this modality-neutral idiolectic information. Our lab has shown that becoming

familiar with a speaker through silent lip reading later facilitates perception of that speaker's auditory speech (Fig. 2; Rosenblum, Miller, & Sanchez, 2007). This cross-modal transfer of speaker familiarity suggests that some of the information allowing familiarity to facilitate speech perception takes a modality-neutral form.

In sum, amodal accounts of multimodal speech perception claim that, in an important way, speech information is the same whether instantiated as acoustic or optic energy. This is not to say that speech information is equally available across modalities: A greater range of speech information is generally available through hearing than vision. Still, the information that *is* available takes a common form across modalities, and as far as speech perception is concerned, the modalities are never really separate.

### **ALTERNATIVE THEORIES OF MULTIMODAL SPEECH PERCEPTION**

While amodal accounts have been adopted by a number of audiovisual speech researchers, other researchers propose that the audio and visual streams are analyzed individually, and maintain that they are separated up through the stages of feature determination (e.g., Massaro, 1998) or even through word recognition (Bernstein, Auer, & Moore, 2004). These late-integration theories differ on how the evidence for early integration is explained, but some propose influences of top-down feedback from multimodal brain centers to the initial processing of individual modalities (Bernstein et al., 2004).

In fact, some very recent findings hint that speech integration might not be as automatic and immediate as amodal perspectives would claim. These new results have been interpreted as revealing higher-cognitive, or “upstream,” influences on speech integration—an interpretation consistent with late-integration theories. For example,

lexical status (whether or not an utterance is a word) can bear on the strength of McGurk-type effects. Visual influences on subject responses are greater if the influenced segment (audio “b” + visual “v” = “v”) is part of a word (*valve*) rather than nonword (*vatch*; Brancazio, 2004). Similarly, semantic context can affect the likelihood of reporting a visually influenced segment (Windmann, 2004). Attentional factors can also influence responses to McGurk-type stimuli. Observers presented stimuli composed of speaking-face videos dubbed with sine-wave speech will only report a visual influence if instructed to hear the sine waves *as* speech (Tuomainen, Andersen, Tiippana, & Sams, 2005). Other results challenge the presumed completeness of audiovisual speech integration. When subjects are asked to shadow (quickly repeat) a McGurk-type utterance (audio “aba” + visual “aga” = shadowed “ada”), the formant structure of the production response shows remnants of the individual audio and visual components (Gentilucci & Cattaneo, 2005).

These new results might challenge the presumed automaticity and completeness of audiovisual speech integration, and could be interpreted as more consistent with late-integration than with amodal accounts. However, other explanations for these findings exist. Perhaps the observed upstream effects bear not on integration itself but, instead, on the recognition of phonemes that are already integrated (which, if composed of incongruent audio and visual components, can be more ambiguous and thus more susceptible to outside influences). Further, evidence that attending to sine-wave signals as speech is necessary for visual influences might simply show that while attention can influence whether amodal speech information is detectable, its recovery, once detected, is automatic and impervious to outside influences. Future research will be needed to test

these alternative explanations. At the least, these new results will force proponents of amodal accounts to more precisely articulate the details of their approach.

### **FUTURE DIRECTIONS**

As I have suggested, multimodal speech perception research has become paradigmatic for the field of general multimodal integration. In so far as an amodal theory can account for multimodal speech, it might also explain multimodal integration outside of the speech domain. There is growing evidence for an automaticity, immediacy, and neurophysiological primacy of nonspeech multimodal perception (Shimojo & Shams, 2001). In addition, modality-neutral descriptions have been applied to nonspeech information (e.g., for perceiving the approach of visible and audible objects) to help explain integration phenomena (Gordon & Rosenblum, 2005). Future research will likely examine the suitability of amodal accounts to explain general multimodal integration.

Finally, mention should be made of how multimodal-speech research has been applied to practical issues. Evidence for the multimodal primacy of speech has enlightened our understanding of brain injuries, autism, schizophrenia, as well as the use of cochlear implant devices. Rehabilitation programs in each of these domains have incorporated visual-speech stimuli. Future research testing the viability of amodal accounts should further illuminate these and other practical issues.

### **Recommended Reading**

Bernstein, L.E., Auer, E.T., Jr., & Moore, J.K. (2004). (See References). Presents a “late integration” alternative to amodal accounts as well as a different interpretation of the neurophysiological data on multimodal speech perception.

Brancazio, L. (2004). (See References). This paper presents experiments showing lexical influences on audiovisual speech responses and discusses multiple explanations.

Calvert, G.A., & Lewis, J.W. (2004). Hemodynamic studies of audiovisual interactions. In G.A. Calvert, C. Spence, & B.E. Stein (Eds.), *The handbook of multisensory processing*, 483–502. Cambridge, MA: MIT Press. Provides an overview of research on neurophysiological responses to speech and nonspeech cross-modal stimuli.

Fowler, C.A. (2004). Speech as a supramodal or amodal phenomenon. In G.A. Calvert, C. Spence, & B.E. Stein (Eds.), *The handbook of multisensory processing*, 189–202, Cambridge, MA: MIT Press. Provides an overview of multimodal speech research and its relation to speech production and the infant multimodal perception literature; also presents an argument for an amodal account of cross-modal speech.

Rosenblum, L.D. (2005). (See References). Provides an argument for a primacy of multimodal speech and a modality-neutral (amodal) theory of integration.

***Acknowledgments***—This research was supported by the National Institute on Deafness and Other Communication Disorders Grant 1R01DC008957-01. The author would like to thank Rachel Miller, Mari Sanchez, Harry Reis, and two anonymous reviewers for helpful comments.

## REFERENCES

- Bernstein, L.E., Auer, E.T., Jr., & Moore, J.K. (2004). Audiovisual speech binding: Convergence or association. In G.A. Calvert, C. Spence, & B.E. Stein (Eds.), *Handbook of multisensory processing* (pp. 203–223). Cambridge, MA: MIT Press.
- Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, 30, 445–463.
- Burnham, D., Ciocca, V., Lauw, C., Lau, S., & Stokes, S. (2000). Perception of visual information for Cantonese tones. In M. Barlow & P. Rose (Eds.), *Proceedings of the Eighth Australian International Conference on Speech Science and Technology* (pp. 86–91). Canberra: Australian Speech Science and Technology Association.
- Calvert, G.A., Bullmore, E., Brammer, M.J., Campbell, R., Iversen, S.D., Woodruff, P., et al. (1997). Silent lipreading activates the auditory cortex. *Science*, 276, 593–596.
- Fowler, C.A., & Dekle, D.J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, 17, 816–828.
- Gentilucci, M., & Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. *Experimental Brain Research*, 167, 66–75.
- Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., & Logothetis, N.K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience*, 25, 5004–5012.

- Gordon, M.S. & Rosenblum, L.D. (2005). Effects of Intra-Stimulus Modality Change on Audiovisual Time-to-Arrival Judgments. *Perception & Psychophysics*, 67, 580–594.
- Grant, K .W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, 108, 1197–1208.
- Green, K.P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell & B. Dodd (Eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech* (pp. 3–25). London, UK: Erlbaum.
- Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
- McGurk, H., & MacDonald, J.W. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
- Munhall, K., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraint on audiovisual speech perception. In G.A. Calvert, C. Spence, & B.E. Stein (Eds.), *The handbook of multisensory processing* (pp. 177–188). Cambridge, MA: The MIT Press.
- Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2005). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, 168, 1–10.

- Remez, R.E., Fellowes, J.M., & Rubin, P.E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception & Performance*, *23*, 651–666.
- Rosenblum, L.D. (2005). The primacy of multimodal speech perception. In D. Pisoni & R. Remez (Eds.), *Handbook of speech perception* (pp. 51–78). Malden, MA: Blackwell.
- Rosenblum, L.D., Miller, R.M., & Sanchez, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. *Psychological Science*, *18*, 392–396.
- Sams, M., Mottonen, R. & Sihvonen. (2005). Seeing and hearing others and oneself talk. *Cognitive Brain Research*, *23*, 429– 435.
- Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Current Opinion in Neurobiology*, *11*, 505–509.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53–83). London: Erlbaum.
- Tuomainen J., Andersen T.S., Tiippana K., & Sams M. (2005). Audio-visual speech perception is special. *Cognition*, *96*, B13–B22.
- Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language*, *50*, 212–230.

**Fig. 1.** Functional magnetic resonance imaging (fMRI) scans depicting average cerebral activation of five individuals when listening to words (blue voxels) and when lip-reading a face silently mouthing numbers (purple voxels; adapted from Calvert et al., 1997). The yellow voxels depict the overlapping areas activated by both the listening and lip-reading tasks. The three panels represent the average activation measured at different vertical positions, and the left side of each image corresponds to the right side of the brain. The images reveal that the silent lip-reading task, like the listening task, activates primary auditory and auditory-association cortices.

**Fig. 2.** Data from an experiment testing the influence of lip-reading from a specific talker on the ability to later hear speech produced by either that same talker or a different talker, embedded in varying amounts of noise (adapted from Rosenblum, Miller, & Sanchez, 2007). Sixty subjects screened for minimal lip-reading skill first lip-read 100 simple sentences from a single talker. Subjects were then asked to identify a set of 150 auditory sentences produced by either the talker from whom they had just lip-read or a different talker. The heard sentences were presented against a background of noise that varied in signal-to-noise ratios: +5 dB (decibels), 0 dB, and -5 dB. For all levels of noise, the subjects who heard sentences produced by the talker from whom they had previously lip-read were better able to identify the auditory sentences than were subjects who heard sentence from a different talker.