Does Sensory Modality Matter? Not for Speech Perception

Lawrence D. Rosenblum

Address:
Department of Psychology
University of California, Riverside
900 University Drive
Riverside, California 92521
USA

Email:
lawrence.rosenblum@ucr.edu

Speech perception is possible through sound, sight, and touch, and the speech brain treats all of this input the same.

Speech by Touch

Rick Joy is deaf and blind. But you wouldn't know this from our conversation. When I ask him questions about his language training, Rick provides detailed answers on how deaf-blind individuals were taught to perceive and produce speech. He then adds "I'm probably one of the last of my kind." He explains that he only knows of eight remaining individuals in the United States who have been trained to understand speech the way he understands mine: by touching my face.

Rick is using the "Tadoma" method of perceiving speech by touching my lips, jaw, and neck with the fingers of his right hand (for demonstrations of Tadoma, see bit.ly/2WQq4E7). In this way, he is able to understand me about as well as when a hearing person is conversing in a noisy restaurant. I have to repeat myself every few minutes but this doesn't inhibit the flow of our conversation. He responds with his own speaking voice. Rick's Tadoma skills allowed him to excel in high school and become the first blind-deaf Eagle Scout. It also allowed him to graduate college and then design circuit boards for Hewlett-Packard for 30 years.

The Tadoma technique was taught to a young Helen Keller and other deaf-blind children in the early to mid-twentieth century. These days, there are fewer deaf-blind infants (thanks to the rubella vaccine), and for those who are, cochlear implants and other technical advances have made the use of Tadoma rare. Rick Joy is one of the last individuals known to have been formerly taught the technique. I ask Rick if this makes him feel special, and he answers "Not really. Tadoma is something anyone can learn if they have the time and patience. Of course, there's not much of that around these days." We both laugh.

Rick is absolutely correct about Tadoma. Research shows that subjects with normal hearing and vision can learn to use the technique nearly as well as Rick does if they are willing to dedicate 100 hours to practice (Reed et al., 1985). But it is the research on Tadoma *novices* that is most striking. Despite most of us having little, if any, experience touching faces for speech, we can all easily identify many consonants and vowels using the technique. Moreover, as soon as we touch a face for speech, we integrate the speech we feel with the speech we hear (Fowler and Dekle, 1991). For us novice users, touching the face of a talker can quickly enhance our understanding of noisy speech as well as make *lipread* speech easier to comprehend (Gick et al., 2008). This inherent utility of felt speech is evident in brain reactivity; simultaneously touching and listening to a talker speeds critical evoked responses in auditory brain areas (Treille et al., 2014). Taken together, research on Tadoma shows that despite few of us using the technique as well as Rick Joy, our brains are ready to use felt speech and use it similarly to how it uses heard speech: as information about articulation.

https://doi.org/10.1121/AT.2019.15.3.47

The more we understand about the perceptual brain, the more it seems agnostic about sensory modality. Brain areas once thought dedicated to a single sense are now known to react to multiple senses (for a review, see Rosenblum et al., 2016). Many who study multisensory perception now believe that the perceptual brain is more accurately characterized as being designed around tasks and behavioral function than around individual sensory systems (e.g., Reich et al., 2012). The research supporting this new conception comes from multiple areas of behavioral and neurophysiological perceptual science. However, much of what has come to be known as the Multisensory Revolution (e.g., Rosenblum, 2013) has been motivated by research on speech perception. The aforementioned research on Tadoma and felt speech has been part of this endeavor. But much more of this work has addressed our more usual way of perceiving speech: via audiovisual means.

We All Lipread

Research shows that regardless of our hearing, we use visible speech (lipread) information when it is available. We use visible speech to enhance our perception of auditory speech that is degraded by background noise (e.g., Bernstein et al., 2004b) or a heavy foreign accent (Arnold and Hill, 2001). We use visual speech as we acquire our first language(s) (e.g., Teinonen et al., 2008) and second languages (e.g., Hazan et al., 2005). In fact, not having access to visual speech during language development causes predictable delays for blind children, the remnants of which can be observed in adulthood (e.g., Delvaux et al., 2018).

Perhaps the most compelling demonstration of audiovisual speech is the McGurk effect (McGurk and MacDonald, 1976). There are myriad examples of the effect online (e.g., illusionsindex.org/i/mcgurk-effect and acousticstoday.org/speech-not-acoustic). In one example, a video of a face articulating the syllables "ba," "ga," "va," and "la" is synchronously dubbed with an audio recording of the repeated syllable "ba." Observers asked what they *hear* typically report "ba," "da," "va," and "tha" despite their ears receiving a clear "ba" four times. Thus, it seems that what we *hear* can be strongly affected by what we see.

I have been demonstrating the McGurk effect in this way to my classes for over 30 years. Still, the effect works on me as well as it ever has. Indeed, research shows that the effect works regardless of one's awareness of the audiovisual discrepancy (e.g., Bertelson and de Gelder, 2004). The effect also works in different languages (e.g., Sams et al., 1998) when there are extreme audio and visual stimulus degradations (e.g., Rosenblum and Saldana, 1996) as well as across observers of different ages and perceptual experience (e.g., Jerger et al., 2014; but see Proverbio et al., 2016). There are certainly individual differences in the *strength* of the effect depending on, for example, the involved segments (e.g., related to native language). Still, the vast majority of (neurologically typical) individuals show some form of the effect.

One of the more interesting aspects of the McGurk effect is how visual speech can influence what one experiences *hearing*. This phenomenology corresponds to the neurophysiology of visual speech perception. Seeing an articulating face can induce activity in *auditory* brain areas, even for novice lipreaders (e.g., Calvert et al., 1997; see Rosenblum et al., 2016, for a review). In fact, visual speech was the first stimulus to show cross-sensory activation of a primary sensory brain area in humans. Visual speech can also modulate more upstream (earlier) auditory mechanisms (Musacchia et al., 2006; Namasivayam et al., 2015). For audiovisual McGurk stimuli, a visual syllable "va" synchronized with an auditory "ba" induces auditory brain area activity consistent with the activity from hearing an auditory "va" (Callan et al., 2001). Based on this neurophysiology, it is not surprising that observers experience "hearing" what they are seeing.

The McGurk effect is one of the most studied phenomena in modern perceptual psychology. However, some of us have recently questioned its use as a tool to measure the strength of multisensory integration (for reviews, see Alsius et al., 2018; Rosenblum, 2019). There is strong evidence, for example, that when the McGurk effect appears to fail (a perceiver reports just hearing the audio component), dimensions of the channels are still integrated (e.g., Brancazio and Miller, 2005).

Still, the effect is useful for simply establishing that integration *has* occurred, and the effect can occur in some very surprising ways. Consider the aforementioned speech perception by touch. Research shows that touching a face articulate syllables while listening to different syllables can make the heard syllables "sound" like those being felt (Fowler and Dekle, 1991). Relatedly, a brief puff of air applied to an observer's skin (on the neck or arm) can integrate with synchronized heard syllables to make a "ba" sound more like a "pa" (e.g., Gick and Derrick 2009). In another touch example, a heard vowel ("ea" in "head") can

sound different (as "a" in "had") if it is synchronously timed with the gentle pulling up of the skin at the corner of a listener's mouth (Ito et al., 2009).

Besides demonstrating that the speech brain readily integrates all relevant articulatory information regardless of modality, these touch examples help make another point. It seems that speech information can be integrated regardless of one's experience with the modality through which it is conveyed. Very few of us have experience touching faces for speech, extracting speech information from puffs on our skin, or having our mouth pulled as we listen to speech. Still, observers seem to readily integrate that novel information for perception.

In this sense, these examples may pose a challenge to probabilistic accounts of perception that assume that the likelihood of cue integration depends on probabilities derived from associative experience (e.g., Altieri et al., 2011). These accounts may also have a difficult time accounting for a very recent audiovisual example of the McGurk effect. Watching ultrasound videos of tongue blade movements can influence heard speech and induce brain responses characteristic of typical audiovisual speech integration (e.g., Treille et al., 2018). It seems that the speech brain is primed to integrate all types of information for speech articulation, even without prior associative experience between the information streams.

The Senses Share Their Experience

There is another context in which specific associative experience may be unnecessary for the modalities to help one another: speech learning. As mentioned, new language learners benefit from seeing as well as hearing someone speak. This *multisensory training benefit* also extends to help our auditory comprehension when later just listening to the new language (e.g., Hazan et al., 2005). Multisensory stimuli are also useful for training listeners with mild hearing impairments to better hear degraded speech (Montgomery et al., 1984).

Multisensory training also helps us learn to audibly recognize a talker's voice (e.g., Schall and von Kriegstein, 2014). Thus, if you are having difficulty distinguishing talkers on your favorite podcast, research suggests that you would greatly benefit from watching them speak for a short period. A small amount of audiovisual experience would then enhance your ability to distinguish the talkers by hearing alone.

The multisensory training benefit also allows one to understand what a new talker is *saying* but in a particularly interesting way. It has long been known that listeners are able to better understand the speech of familiar versus unfamiliar talkers (for a review, see Nygaard, 2005). Predictably, this familiar talker advantage is even greater if one has audiovisual experience with the talker (e.g., Riedel et al., 2015). More surprising is that we are better able to lipread a familiar talker, even if our silent lipreading is not very good, and that familiarity is gained over just 30 minutes (e.g., Yakel et al., 2000). But even more surprising is that the experience one gets from silently lipreading a talker will then allow them to better *hear* that talker's voice (Rosenblum et al., 2007).

This is a particularly interesting instance of the multisensory training benefit. In this experiment, participants never experienced the talker *bimodally*; they never simultaneously saw and heard the talker speak. Instead, their familiarity with the talker through lipreading seemed to *transfer across modalities*, allowing them to then hear that talker better.

Related research shows that transfer of talker familiarity can also work in the opposite direction so that initial auditory experience with a talker makes them easier to lipread later on (Sanchez et al., 2013). Additionally, experience with *recognizing talkers* in one modality can transfer to allow better recognition of those talkers in the other modality (Simmons et al., 2015).

How might talker experience transfer across modalities despite perceivers never having bimodal experience with the talker? It may be that perceivers are learning something about the talker's articulatory style (e.g., idiolect). Because articulatory style can be conveyed audibly and visibly, learning to attend to a talker's idiosyncrasies in one modality may allow a perceiver to attend to, and take advantage of, those same idiosyncrasies in the other modality. This conjecture is based on a number of other findings.

First, despite our intuitions, talkers do look like they sound. This is apparent from research showing that perceivers can successfully match a talker's voice to their (silently) articulating face, even if the face and voice are saying different words (e.g., Lachs and Pisoni, 2004). Furthermore, perceivers are able to perform these matches when the voice is reduced to a signal of simple sine waves and the face is reduced to a video of white points moving against a black

background (Lachs and Pisoni, 2004). These sine-wave speech and point-light stimuli are, no doubt, very odd for perceivers. However, despite lacking what is typically thought of as useful audio and visual information for identifying talkers (e.g., fundamental pitch, voice timbre, lips and other facial features), these stimuli can convey both usable speech and talker information (e.g., Remez et al., 1997; Rosenblum et al., 2002).

It is thought that although severely degraded, these stimuli *do* retain talker-specific phonetic information, including articulatory style. If true, then the perceivers may be able to match faces to voices by attending to the remaining articulatory-style information present in these odd stimuli. Moreover, it may be this talker-specific phonetic information, available in both modalities, that the perceivers are learning as they become familiar with a talker. If so, then this may help explain the crossmodal talker facilitation effects. The perceivers may become familiar with, and adept at using, talker-specific phonetic information based on experience with one modality. When they are then presented the same talker-specific information in the other modality, they can use that experience to better recognize that talker and what they are saying.

From this perspective, speech learning involves becoming more adept at attending to speech properties that are *amodal*: articulatory properties that can be conveyed through multiple modalities. This is a striking claim that prompts multiple questions. For example, what form might these amodal informational parameters take if they can be conveyed in light and sound? The answer may be what has come to be known as *supramodal information*.

Supramodal Information

On the surface, auditory and visual speech information seem very different. Although auditory speech information is necessarily revealed over time, visual speech is often construed as more spatial in nature (visible lip shapes, jaw position). But research with sine-wave and point-light speech stimuli (along with other work) has revealed another way of considering the information (for a review, see Rosenblum et al., 2016). Recall that both types of stimuli retain only the more global, time-varying dimensions of their respective signals, yet are still effective at conveying speech and talker information. When considered in this way, the salient informational forms in each modality are more similar.

Consider the higher order information for a very common speech production occurrence: reversal of the articulators as in the production of "aba." As the jaw and lower lip rise, close the mouth, and then reverse, there is an accompanying reversal in optical (visible) structure (Summerfield, 1987). Importantly, this visible reversal is also accompanied by a reversal in the amplitude and spectral structure of the resultant acoustic signal. Furthermore, the articulation, optic, and acoustic reversals all share the same temporal parameters. Thus, at this level of abstraction, the audible and visible time-varying information takes the same form: a form known as *supramodal information* (e.g., Rosenblum et al., 2016). The brain's sensitivity to supramodal information may account for many multisensory speech phenomena.

Supramodal information may also account for the surprisingly high correlations observed between the signals (e.g., Munhall and Vatikiotis-Bateson, 2004). Detailed measures of facial movements have been shown to be highly correlated with amplitude and spectral changes in the acoustic signal. Part of the reason for the high correlations is the degree to which visible movements can inform about deeper, presumably "hidden," vocal tract actions. Parameters of vocal intonation (vocal pitch changes) are actually correlated with, and therefore visible through, head-nodding motions. Similarly, the deep vocal tract actions of intraoral pressure changes and lexical tone (vowel pitch changes to mark words, as in Mandarin) can be perceived by novice lipreaders (Burnham, et al., 2000).

The strong correlations between visible and audible speech signals have allowed usable visible speech to be animated directly from the acoustic signal (e.g., Yamamoto et al., 1998) and audible speech to be synthesized from the parameters of visible speech movements (e.g., Yehia et al., 2002).

Returning to speech perception, the supramodal information thesis states that the brain can make use of this higher level information that takes the same form across modalities. In fact, research shows that when supramodal information for a segment *is* available in both modalities, the speech function seems to take advantage (e.g., Grant and Seitz, 2000). As intimated, the supramodal information thesis might help explain how speech and talker learning can be shared across modalities, without bimodal experience. If listening to a talker involves attuning to supramodal talker-specific properties available in the acoustic signal, then later lipreading the talker becomes easier because those same supramodal properties can be accessed by the visual system. A similar conception may help explain *multisensory training benefits* overall as well as our ability to match talking voices and faces.

The supramodal thesis also seems compatible with the modal flexibility of the brain (e.g., Rosenblum et al., 2016). As stated, auditory brain areas respond to visual, and even haptic, speech (for reviews, see Treille et al., 2014; Rosenblum et al., 2016).

The supramodal account may also help explain some *general* commonalities observed across the auditory and visual speech functions. First, sine-wave and point-light speech show that the brain can use dynamic, time-varying information in both modalities (Remez et al., 1997; Rosenblum et al., 2002). A second general commonality is that talker information can interact with, and even inform, speech perception in both modalities. As stated, we perceive speech better from familiar talkers whether listening or lipreading, despite having little formal experience with the latter. There is also neurophysiological evidence for a single brain area that brings together audiovisual talker and phonetic information (e.g., von Kriegstein et al., 2005).

We Always Imitate

There is third general way in which auditory and visual speech perception are interestingly similar; they both act to shape the phonetic details of a talker's response. During conversation, talkers will inadvertently imitate subtle aspects each other's speech intonation, speed, and vocal intensity (e.g., Giles et al., 1991). Talkers will also subtlety imitate one another's more microscopic aspects of speech: the phonetic details (e.g., Pardo, 2006). These details include vowel quality (e.g., Pardo, 2006) and the talker-specific delay in vocal cord vibration onset for segments such as "p" (Shockley et al., 2004).

This *phonetic convergence* not only occurs in the context of live conversation but also in the lab when participants are asked to listen to words and then simply say the words they hear out loud (e.g., Goldinger, 1998). Despite never being asked to explicitly mimic, or even "repeat," participants will inadvertently articulate their words in a manner more similar to the words they hear. There are a number of possible reasons for phonetic convergence including facilitation of social bonding (e.g., Pardo, 2006); easing speech understanding when faced with background noise (Dorsi et al., in preparation); and/or a by-product of the known link between speech perception and production (e.g., Shockley et al., 2004).

Importantly, there is now evidence that phonetic convergence can be induced by *visible* speech in perceivers with no formal lipreading experience (Miller et al., 2010). Visible speech can also enhance convergence because research shows that having visual as well as audible access to a talker's articulations will increase one's degree of imitation (e.g., Dias and Rosenblum, 2016). Finally, evidence shows that audible and visible speech integrate before inducing convergence in a listener's produced speech (Sanchez et al., 2010).

The fact that both auditory and visual speech behave similarly in inducing convergence is consistent with the neuroscience. As intimated, one explanation for convergence is the hypothesized connection between speech perception and production (e.g., Shockley et al., 2004). Convergence may partly be a by-product of the speech production system being enlisted for, and thus *primed by*, perception of the idiosyncrasies of a perceived word. The question of motor system involvement in speech perception has been ongoing since the 1960s (for a review, see Fowler et al., 2015). Although it is unclear whether motor involvement is necessary or just facilitatory (e.g., Hickok et al., 2009), it is known that speech motor brain areas are typically primed during speech perception (for a review, see Rosenblum et al., 2016).

Importantly, it is also known that motor areas of the brain are primed during *visual speech perception* regardless of one's formal lipreading experience (e.g., Callan et al., 2003). Motor brain involvement also seems *enhanced* when perceiving audiovisual versus audio-alone or video-alone speech (e.g., Callan et al., 2014; but see Matchin et al., 2014). This finding is consistent with the enhanced phonetic convergence observed for audiovisual, versus audio speech (e.g., Dias and Rosenblum, 2016).

Thus, both the behavioral and neurophysiological research reveal a commonality in the ability of auditory and visual speech information to induce a convergent production response. This characteristic joins time-varying and talker-relevant dimensions as general forms of information commonalities across the modalities. These facts, together with the close correlations between the detailed visible and acoustic dimensions, provide support for the speech brain being sensitive to a supramodal form of information.

Future Questions

The supramodal account proffers that much of multisensory speech perception is based on a speech function sensitive to higher order information that takes the same form across modalities. Although this may seem an unconventional theory of multisensory perception, we believe that it is consistent with much of the behavioral and neurophysiological data.

Future research can be designed to test additional aspects of the theory. Fortunately, the theory makes some very specific predictions. For example, if multisensory perception is actually a consequence of common supramodal information contained in both light and sound, then "integration" functionally occurs at the *level of the stimulus input*. If this is true, evidence of integration should be observed at the earliest stages. Potentially, early integration is already evidenced by (1) visual modulation of early auditory brain areas and (2) crossmodal influences of low-level speech features, such as the voice onset timing distinguishing "p" from "b." However, other researchers have argued that the modalities stay separate up through the determination of words (Bernstein et al., 2004a). Future research will need to examine additional evidence for early versus later integration of the channels.

Relatedly, if as the supramodal approach claims, integration is a function of the input itself, then integration should be "impenetrable" to other cognitive influences (e.g., higher level linguistics, attention). However, a number of studies have shown higher level *lexical* influences on the strength of the McGurk effect (e.g., Brancazio, 2004), contrary to the prediction of the supramodal account. As intimated above, however, the McGurk effect is not a straight forward tool for measuring integration. Very recent research suggests that lexical influences may actually bear on postintegration categorization of segments (Dorsi, 2019). Still, more research is needed to determine the degree to which multisensory integration is impenetrable to outside cognition.

Finally, although work has been conducted to discover supramodal information across audio and visual channels, similar principles may apply to the haptic channel as well. As discussed, the haptic channel seems to induce the same perceptual and neurophysiological cross-sensory modulations as audio and visual speech. It is less clear how an informational form in the haptic stream could be supramodal with the other channels (but see Turvey and Fonseca, 2014). Future research can address this question to explain the miraculous abilities of Rick Joy to provide his speech brain with articulatory information from a most surprising source.

References

Alsius, A., Paré, M., and Munhall, K. G. (2018). Forty years after hearing lips and seeing voices: the McGurk effect revisited. *Multisensory Research* 31(1-2), 111-144.

Altieri, N., Pisoni, D. B., and Townsend, J. T. (2011). Some behavioral and neurobiological constraints on theories of audiovisual speech integration: A review and suggestions for new directions. *Seeing and Perceiving* 24(6), 513-539.

Arnold, P., and Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology* 92(2), 339-355.

Bernstein, L. E., Auer, E. T., Jr., and Moore, J. K. (2004a). Convergence or association? In G. A. Calvert, C. Spence, and B. E. Stein (Eds.), *Handbook of Multisensory Processes*. MIT Press, Cambridge, MA, pp. 203-220.

Bernstein, L. E., Auer, E. T., Jr., and Takayanagi, S. (2004b). Auditory speech detection in noise enhanced by lipreading. *Speech Communication* 44(1), 5-18.

Bertelson, P., and de Gelder, B. (2004). The psychology of multi-sensory perception. In C. Spence and J. Driver (Eds.), *Crossmodal Space and Crossmodal Attention*. Oxford University Press, Oxford, UK, pp. 141-177.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. Journal of Experimental Psychology: *Human Perception and Performance* 30(3), 445-463.

Brancazio, L., and Miller, J. L. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception & Psychophysics* 67(5), 759-769.

Burnham, D., Ciocca, V., Lauw, C., Lau, S., and Stokes, S. (2000). Perception of visual information for Cantonese tones. *Proceedings of the Eighth Australian International Conference on Speech Science and Technology*, Australian Speech Science and Technology Association, Canberra, December 5-7, 2000, pp. 86-91.

Callan, D. E., Callan, A. M., Kroos, C., and Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception reveled by independent component analysis: A single sweep EEG case study. *Cognitive Brain Research* 10(3), 349-353.

Callan, D. E., Jones, J. A., and Callan, A. (2014). Multisensory and modality specific processing of visual speech in different regions of the premotor cortex. *Frontiers in Psychology* 5, 389. https://doi.org/10.3389/fpsyg.2014.00389.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., and Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport* 14(17), 2213-2218. https://doi.org/10.1097/00001756-200312020-00016.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., and David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science* 276(5312), 593-596. https://doi.org/10.1126/science.276.5312.593.

Delvaux, V., Huet, K., Piccaluga, M., and Harmegnies, B. (2018). The perception of anticipatory labial coarticulation by blind listeners in noise: A comparison with sighted listeners in audio-only, visual-only and audiovisual conditions. *Journal of Phonetics* 67, 65-77.

Dias, J. W., and Rosenblum, L. D. (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, & Psychophysics* 78, 317-333. https://doi.org/10.3758/s13414-015-0982-6.

Dorsi, J. (2019). *Understanding Lexical and Multisensory Context Support of Speech Perception*. Doctoral Dissertation, University of California, Riverside, Riverside.

Fowler, C. A., and Dekle, D. J. (1991). Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology* 17(3), 816-828.

Fowler, C. A., Shankweiler, D., and Studdert-Kennedy, M. (2015). "Perception of the speech code" revisited: Speech is alphabetic after all. *Psychological Review* 123(2), 125-150.

Gick, B., and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature* 462(7272), 502-504. https://doi.org/10.1038/nature08572.

Gick, B., Johannsdottir, K. M., Gibraiel, D., and Muhlbauer, J. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *The Journal of the Acoustical Society of America* 123(4), 72-76. https://doi.org/10.1121/1.2884349.

Giles, H., Coupland, N., and Coupland, I. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, and N. Coupland (Eds.), *Contexts of Accommodation: Developments in Applied Sociolinguistics*. Cambridge University Press, Cambridge, UK, pp. 1-68.

Goldinger, S. (1998). Echoes of echoes? Shadowing words and nonwords in an episodic lexicon. *Psychological Review* 105, 251-279.

Grant, K. W., and Seitz, P. F. P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America* 108(3), 1197-1208. https://doi.org/10.1121/1.422512.

Hazan, V., Sennema, A., Iba, M., and Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication* 47(3), 360-378. https://doi.org/10.1016/j.specom.2005.04.007.

Hickok, G., Holt, L. L., and Lotto, A. J. (2009). Response to Wilson: What does motor cortex contribute to speech perception? *Trends in Cognitive Sciences* 13(8), 330-331. https://doi.org/10.1016/j.tics.2009.06.001.

Ito, T., Tiede, M., and Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America* 106(4), 1245-1248. https://doi.org/10.1073/pnas.0810063106.

Jerger, S., Damian, M. F., Tye-Murray, N., and Abdi, H. (2014). Children use visual speech to compensate for non-intact auditory speech. *Journal of Experimental Child Psychology* 126, 295-312. https://doi.org/10.1016/j.jecp.2014.05.003.

Lachs, L., and Pisoni, D. B. (2004). Specification of cross-modal source information in isolated kinematic displays of speech. *The Journal of the Acoustical Society of America* 116(1), 507-518. https://doi.org/10.1121/1.1757454.

Matchin, W., Groulx, K., and Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: Evidence from articulatory suppression, the McGurk effect, and fMRI. *Journal of Cognitive Neuroscience* 26(3), 606-620.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746-748.

Miller, R., Sanchez, K., and Rosenblum, L. (2010). Alignment to visual speech information. *Attention, Perception, & Psychophysics* 72(6), 1614-1625. https://doi.org/10.3758/APP.72.6.1614.

Montgomery, A. A., Walden, B. E., Schwartz, D. M., and Prosek, R. A. (1984). Training auditory-visual speech reception in adults with moderate sensorineural hearing loss. *Ear and Hearing* 5(1), 30-36. https://doi.org/10.1097/00003446-198401000-00007.

Munhall, K. G., and Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. A. Calvert, C. Spence, and B. E. Stein (Eds.), *Handbook of Multisensory Processes*. MIT Press, Cambridge, MA, pp. 177-188.

Musacchia, G., Sams, M., Nicol, T., and Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*,168(1-2), 1-10. https://doi.org/10.1007/s00221-005-0071-5.

Namasivayam, A. K., Wong, W. Y. S., Sharma, D., and van Lieshout, P. (2015). Visual speech gestures modulate efferent auditory system. *Journal of Integrative Neuroscience* 14(1), 73-83.

Nygaard, L. C. (2005). The integration of linguistic and non-linguistic properties of speech. In D. Pisoni and R. Remez (Eds.), *Handbook of Speech Perception*. Blackwell, Malden, MA, pp. 390-414.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119(4), 2382-2393. https://doi.org/10.1121/1.2178720.

Proverbio, A. M., Massetti, G., Rizzi, E., and Zani, A. (2016). Skilled musicians are not subject to the McGurk effect. *Scientific Reports* 6, 30423. https://doi.org/10.1038/srep30423.

Reed, C. M., Rabinowitz, W. M., Durlach, N. I., Braida, L. D., Conway-Fithian, S., and Schultz, M. C. (1985). Research on the Tadoma method of speech communication. *The Journal of the Acoustical Society of America* 77(1), 247-257. https://doi.org/10.1121/1.392266.

Reich, L., Maidenbaum, S., and Amedi, A. (2012). The brain as a flexible task machine: Implications for visual rehabilitation using noninvasive vs. invasive approaches. *Current Opinion in Neurobiology* 25(1), 86-95.

Remez, R. E., Fellowes, J. M., and Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance* 23(3), 651-666. https://doi.org/10.1037/0096-1523.23.3.651.

Riedel, P., Ragert, P., Schelinski, S., Kiebel, S. J., and von Kriegstein, K. (2015). Visual face-movement sensitive cortex is relevant for auditory-only speech recognition. *Cortex* 68, 86-99. https://doi.org/10.1016/j.cortex.2014.11.016.

Rosenblum, L. D. (2013). A confederacy of the senses. *Scientific American* 308, 72-75.

Rosenblum, L. D. (2019). Audiovisual speech perception and the McGurk effect. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Oxford, UK.

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. https://doi.org/10.1037/0096-1523.22.2.318.

Rosenblum, L. D., Dias, J. W., and Dorsi, J. (2016). The supramodal brain: Implications for auditory perception. *Journal of Cognitive Psychology* 28, 1-23.

Rosenblum, L. D., Miller, R. M., and Sanchez, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. *Psychological Science* 18(5), 392396. https://doi.org/10.1111/j.1467-9280.2007.01911.x.

Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nodarse, B. C., and Niehus, R. P. (2002). Visual speech information for face recognition. *Perception & Psychophysics* 64(2), 220-229. https://doi.org/10.3758/BF03195788.

Sams, M., Manninen, P., Surakka, V., Helin, P., and Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication* 26(1-2), 75-87.

Sanchez, K., Dias, J. W., and Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception & Psychophysics* 75, 1359-1365. https://doi.org/10.3758/s13414-013-0534-x.

Sanchez, K., Miller, R. M., and Rosenblum, L. D. (2010). Visual influences on alignment to voice onset time. *Journal of Speech, Language, and Hearing Research* 53, 262-272.

Schall, S., and von Kriegstein, K. (2014). Functional connectivity between face-movement and speech-intelligibility areas during auditory-only speech perception. *PLoS ONE* 9(1), e86325. https://doi.org/10.1371/journal.pone.0086325.

Shockley, K., Sabadini, L., and Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics* 66(3), 422-429.

Simmons, D. C., Dias, J. W., Dorsi, J., and Rosenblum, L. D. (2015). Crossmodal transfer of talker learning. *The Journal of the Acoustical Society of America* 137, 2416.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*, Lawrence Erlbaum Associates, Inc., London, UK, pp. 53-83.

Teinonen, T., Aslin, R. N., Alku, P., and Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognitio*, 108(3), 850-855. https://doi.org/10.1016/j.cognition.2008.05.009.

Treille, A., Vilain, C., and Sato, M. (2014). The sound of your lips: Electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. *Frontiers in Psychology* 5, 1-8. https://doi.org/10.3389/fpsyg.2014.00420.

Treille, A., Vilain, C., Schwartz, J. L., Hueber, T., and Sato, M. (2018). Electrophysiological evidence for audio-visuo-lingual speech integration. *Neuropsychologia* 109, 126-133.

Turvey, M. T., and Fonseca, S. T. (2014). The medium of haptic perception: A tensegrity hypothesis. *Journal of Motor Behavior* 46(3), 143-187.

Von Kriegstein, K., Kleinschmidt, A., Sterzer, P., and Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience* 17(3), 367-376. https://doi.org/10.1162/0898929053279577.

Yakel, D. A., Rosenblum, L. D., and Fortier, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics* 62, 1405-1412.

Yamamoto, E., Nakamura, S., and Shikano, K. (1998). Lip movement synthesis from speech based on hidden Markov models. *Speech Communication* 26(1-2), 105-115. https://doi.org/10.1016/S0167-6393(98)00054-5.

Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion, and speech acoustics. *Journal of Phonetics* 30(3), 555-568.

BioSketch

Lawrence Rosenblum is a professor of psychology at the University of California, Riverside (Riverside). He studies multisensory speech and talker perception, and his research has been supported by the National Science Foundation (Alexandria, VA) and National Institutes of Health (Bethesda, MD). He is the author of the book *See What I'm Saying: The Extraordinary Powers of Our Five Senses*. His research has been published widely in scientific journals and has been featured in *The Economist, Scientific American,* and *The New York Times* as well as on international radio and television.