Audiovisual Speech Perception and the McGurk Effect

Lawrence D. Rosenblum

Subject: Cognitive Science, Psycholinguistics Online Publication Date: Aug 2019
DOI: 10.1093/acrefore/9780199384655.013.420

Summary and Keywords

Research on visual and audiovisual speech information has profoundly influenced the fields of psycholinguistics, perception psychology, and cognitive neuroscience. Visual speech findings have provided some of most the important human demonstrations of our new conception of the perceptual brain as being supremely multimodal. This “multisensory revolution” has seen a tremendous growth in research on how the senses integrate, cross-facilitate, and share their experience with one another.

The ubiquity and apparent automaticity of multisensory speech has led many theorists to propose that the speech brain is agnostic with regard to sense modality: it might not know or care from which modality speech information comes. Instead, the speech function may act to extract *supramodal* informational patterns that are common in form across energy streams. Alternatively, other theorists have argued that any common information existent across the modalities is minimal and rudimentary, so that multisensory perception largely depends on the observer’s associative experience between the streams. From this perspective, the auditory stream is typically considered primary for the speech brain, with visual speech simply appended to its processing. If the utility of multisensory speech is a consequence of a supramodal informational coherence, then cross-sensory “integration” may be primarily a consequence of the informational input itself. If true, then one would expect to see evidence for integration occurring early in the perceptual process, as well in a largely complete and automatic/impenetrable manner. Alternatively, if multisensory speech perception is based on associative experience between the modal streams, then no constraints on how completely or automatically the senses integrate are dictated. There is behavioral and neurophysiological research supporting both perspectives.

Much of this research is based on testing the well-known McGurk effect, in which audiovisual speech information is thought to integrate to the extent that visual information can affect what listeners report hearing. However, there is now good reason to believe that the McGurk effect is not a valid test of multisensory integration. For example, there are clear cases in which responses indicate that the effect fails, while other measures suggest that integration is actually occurring. By mistakenly conflating the McGurk effect with speech integration itself, interpretations of the completeness and automaticity of multi-

sensory may be incorrect. Future research should use more sensitive behavioral and neurophysiological measures of cross-modal influence to examine these issues.

Keywords: multisensory, audiovisual, speech perception, lip-reading, articulation, supramodal

1. The Multisensory Revolution

Research on visual and audiovisual speech information has had important influences on how we understand the brain. For example, visual speech was the first stimulus to show cross-sensory induction of activation in a human cortical area historically associated with another sense (primary auditory cortex; e.g., Calvert et al., 1997). This, and other visual speech findings have provided some of most the important human demonstrations of our new conception of the perceptual brain as being supremely multimodal. This “multisensory revolution” (e.g., Rosenblum, 2013; Rosenblum, Dori, & Dias, 2016) has seen a tremendous growth in research on how the senses integrate, cross-facilitate, and share their experience with one-another (for reviews, see Ghazanfar & Schroeder, 2006; Rosenblum, Dias, & Dorsi, 2017; Rosenblum et al., 2016).

The ubiquity and apparent automaticity of multisensory speech has led many theorists to propose that the speech brain is agnostic with regard to sense modality: it might not know or care from which modality speech information comes (for a review, see Rosenblum et al., 2016). Instead, the speech function may act to extract informational patterns that are common in form across energy streams. In this sense, speech perception may function in a modality-neutral manner and use the natural coherence of common informational forms across the modalities to support the “merging” of streams (Bicevskis, Derrick, & Gick, 2016; Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Fowler, 2004; Rosenblum et al., 2016). While this *supramodal* perspective is gaining traction, it is not the most prominent theory. More prominent is the assumption that any common information existent across the modalities is minimal and rudimentary, so that multisensory perception largely depends on the observer’s associative experience between the streams (e.g., Shams, 2011). From this perspective, the auditory stream is typically considered primary for the speech brain, with visual speech piggy-backed on its processing (e.g., Diehl & Kleunder, 1989; Hickok, 2009; Magnotti & Beauchamp, 2017).

If the utility of multisensory speech is a consequence of a supramodal informational coherence, then cross-sensory “integration” may be primarily a consequence of the informational input itself (Rosenblum, 2005; Rosenblum et al., 2016). If true, then one would expect to see evidence for integration occurring early in the perceptual process, as well in a largely complete and automatic/impenetrable manner. Alternatively, if multisensory speech perception is based on associative experience between the modal streams, then no constraints on how completely or automatically the senses integrate are dictated. Section 4 discusses the research examining the degree to which multisensory integration is complete and section 5 discusses the research on the impenetrability of integration.

Before these issues are addressed, section 2 presents a general overview of basic multisensory speech phenomena.

2. The Importance of Multisensory Speech Perception and the Influence of the McGurk Effect

Everyone lip-reads. Research shows that regardless of one's hearing, perceivers lip-read (use visible articulation) to enhance perception of auditory speech that is degraded by background noise (e.g., Bernstein, Auer, & Takayanagi, 2004; Sumby & Pollack, 1954), a heavy foreign accent (Arnold & Hill, 2001), and challenging content (Reisberg, McLean, & Goldfield, 1987). Perceivers lip-read to help them acquire their first languages (e.g., Reisberg et al., 1987; Teinonen, Aslin, Alku, & Csibra, 2008), and second languages (Hardison, 2005; Hazan, Sennema, Iba, & Faulkner, 2005; Navarra & Soto-Faraco, 2007). The importance of lip-reading for language acquisition is evidenced by the compromised phonological development in blind children (e.g., Mills, 1987; Brouwer et al., 2015). Blind children are known to have subtle developmental delays in perception and production of segments that are more difficult to audibly distinguish, but easier to see (e.g., /m/ vs. /n/). In fact, remnants of this developmental difference can be observed in blind adults who continue to display subtle speech production and perception differences (e.g., Ménard, Cathiard, Troille, & Giroux, 2015; Ménard, Dupont, Baum, & Aubin, 2009; Ménard, Leclerc, & Tiede, 2014; Ménard et al., 2013).

The most striking example of visual speech perception is the McGurk effect (McGurk & MacDonald, 1976). In the effect's original demonstration, visible syllables of one type (e.g., /va-va/) are synchronously dubbed with audio syllables of a different type (/ba-ba/). Despite the difference in place of articulation, observers are typically unaware of the discrepancy and report "hearing" a syllable strongly influenced by what they see (/va-va/). Depending on the syllable combinations, the resultant percept can sometimes be a visually dominated segment (audio /ba/ + visual /va/ is "heard" as /va/) or a fusion or blend of the audible and visible syllables (audio /ba/ + visual /ga/ is "heard" as /da/). (There are also some syllable combinations for which no visual influence is observed; e.g., McGurk & MacDonald, 1976; Rosenblum, Schmuckler, & Johnson, 1997.)

The original McGurk report (McGurk & MacDonald, 1976) is one of the most cited studies in perceptual psychology (5500, as of this writing; Google Scholar Search), and the effect has been studied under myriad conditions (for reviews, see Alsius, Paré, & Munhall, 2018; Dias, Cook, & Rosenblum, 2017; Tiippana, 2014). Overall, some form of the effect has been shown to be robust in the context of different languages (e.g., Fuster-Duran, 1996; Massaro, Cohen, Gesi, Heredia, & Tzuzaki, 1993; Sams et al., 1998; Sekiyama & Tohkura, 1991), extreme audio and visual stimulus degradations (Andersen, Tiippana, Laarni, Kojo, & Sams, 2009; Rosenblum & Saldana, 1996; Thomas & Jordan, 2002), across different observers' age (e.g., Jerger, Damian, Tye-Murray, & Abdi, 2014), perceptual experience (Jerger et al., 2014; Sams et al., 1998; but see Nath & Beauchamp, 2012; Proverbio, Massetti, Rizzi, & Zani, 2016; Strand, Cooperman, Rowe, & Simenstad, 2014), and awareness

of audio-visual discrepancy (Bertelson & de Gelder, 2004; Bertelson, Vroomen, Wiegeraad, & de Gelder, 1994; Colin, Radeau, Deltenre, Demolin, & Soquet, 2002; Green & Kuhl, 1991; Massaro, 1987; Soto-Faraco & Alsius, 2007, 2009; Summerfield & McGrath, 1984). Importantly, while some visual influence has been observed under all of these conditions, variables can affect the observed *strength* of the effect (e.g., depending on observer age and gender; Irwin, Whalen, & Fowler, 2006; Jerger et al., 2014), and which particular segments merge (e.g., depending on native language; Sekiyama & Tohkura, 1991). There is also strong evidence for individual differences in the effect, with some observers showing little visual influence in their responses (for a review, see Strand et al., 2014). I discuss many of these factors in section 4.

The neurophysiological basis of the McGurk effect—and speech as a multisensory function—has been examined extensively. As stated, visual speech, on its own, was the first stimulus to show a cross-sensory influence on human primary sensory cortexes. Calvert and her colleagues used fMRI (functional magnetic resonance imaging) to observe that watching a talker's silently articulating face can induce activity in auditory cortex (Calvert et al., 1997). This finding has been replicated a number of times, with a number of different technologies (for a review, see Rosenblum et al., 2017; but see Beauchamp, Nath, & Pasalar, 2010). Visual speech can also induce activity in upstream areas including auditory midbrain (Musacchia, Sams, Nicol, & Kraus, 2006) and even cochlear functioning (suppressing otoacoustic emissions; Namasivayam, Wong, Sharma, & van Lieshout, 2015).

Returning to *audiovisual* speech, there is evidence that with McGurk stimuli, the visual segment can override the auditory segment to induce patterns of activity in auditory cortex similar to those induced if the "heard" auditory component was presented on its own (Callan, Callan, Kroos, & Vatikiotis-Bateson, 2001; Colin et al., 2002; Mottonen, Krause, Tiipana, & Sams, 2002; Sams et al., 1991). For example, an audio /va/ - visual /ba/ stimulus will induce functionally identical auditory cortex activity as an (perceptually equivalent) audio /va/ stimulus. This finding seems consistent with observers' phenomenological experience of "hearing" the segment they actually see. The finding may also suggest that visual and auditory speech information is handled similarly by the speech brain. The next section examines this possibility.

3. The Speech Function Treats Auditory and Visual Speech Similarly

In some obvious ways, there are clear differences between auditory and visual speech perception. The functions are based on different energy media and peripheral sensory organs, and show differences in some central neurophysiological mechanisms. The two modalities are also differently suited to inform about different aspects of speech articulation, with auditory speech being the richer signal for most (but not all; e.g., Mills, 1987) phonetic distinctions.

However, there is a surprising degree to which the auditory and visual streams are treated similarly (e.g., Rosenblum, 2005; Rosenblum et al., 2017). The speech function seems to extract similar—and specialized—informational dimensions from both streams. For example, both auditory and visual speech make use of *talker* information to facilitate phonetic perception (for reviews, see Nygaard, 2005; Rosenblum, 2005). Also, for both streams, the time-varying, transitional aspects of the signals seem most salient. Finally, both signals can serve to prime a speech production response in very similar ways. Each of these common characteristics of audio and visual speech is discussed in a later section.

It is intuitive that we are better able to understand the heard speech of a familiar, than unfamiliar, talker—especially if it is degraded by a loud environment or a poor cell phone connection (Borrie, McAuliffe, Liss, O'Beirne, & Anderson, 2013; Nygaard, 2005). What is more surprising is that even without formal lip-reading experience, we are better able to lip-read from a familiar talker (e.g., Lander & Davies, 2008; Schweinberger & Soukup, 1998; Yakel, Rosenblum, & Fortier, 2000). These findings can be interpreted that the speech function uses our familiarity with voice and face characteristics to facilitate speech perception. However, other research suggests that for both modalities, this familiarity may be with something deeper. This research shows that becoming familiar with a talker in one modality can facilitate speech perception in another. Thus, becoming familiar with a talker by lip-reading them (with no sound) for an hour facilitates perception of that talker's *auditory* speech, and vice versa (Rosenblum, Miller, & Sanchez, 2007; Sanchez, Dias, & Rosenblum, 2013). These findings could suggest that for both modalities, the speech function gains experience with a talker's *speaking style*—an amodal property—allowing that talker's speech to be more easily perceived through sound and sight. Regardless, the function treats the streams similarly in making use of indexical information to help recover phonetic information.

There is also evidence that the talker-specific aspects of each stream may be similarly accessed in a more idiosyncratic way. It turns out that a substantial amount of "classic" talker information can be removed from each stream without preventing talker recognition. For auditory speech, the typical talker information of fundamental frequency, and voice quality dimensions (breathiness; raspiness) can be removed, in a technique that leaves only three undulating sinewaves (which track center formant frequencies) remaining in the signal (Remez, Fellowes, & Rubin, 1997). Research shows that this *sinewave speech* can also provide phonetic information (e.g., Remez, Rubin, Pisoni, & Carrell, 1981). For visual speech, the typical facial feature and configural information can be removed so that the visual information is reduced to a series of dots moving along with articulation (e.g., Rosenblum, Johnson, & Saldaña, 1996). Despite this reduction, observers can recognize both phonetic and talker information in these *point-light displays* (Rosenblum et al., 2002; Rosenblum, Smith, & Niehus, 2007). For both sinewave and point-light stimuli, it is thought that what is retained is talker-specific phonetic information which can serve to inform about both what is being said and whom is saying it (e.g., Rosenblum et al., 2016). This *idiolectic* information would be amodal, and therefore available in both modalities. This notion is supported by evidence that observers can match a talker's sinewave speech to their point-light speech, even when these stimuli are from different utterances (Lachs

& Pisoni, 2004). That sinewave and point-light speech may capture common amodal idiolectic information is also supported by recent findings showing that learning to better recognize a particular point-light talker transfers across modalities so that the talker's sinewave speech is then more easily recognized (Simmons, Dias, & Dorsi, & Rosenblum, 2015).

Neurophysiological research also shows that the speech-speaker connection exists across modalities. There is evidence that when asked to report the auditory speech of a talker, observers will show activation in an area associated with face movements (posterior superior temporal sulcus—pSTS; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). This area is activated to a greater degree if very brief audiovisual exposure to the talker is provided (von Kriegstein & Giraud, 2006). Relatedly, when observers are asked to *identify* a voice, they show activation in an area associated with face recognition (fusiform face area; von Kriegstein et al., 2008; von Kriegstein & Giraud, 2006). Again, even brief bimodal exposure to a talker will enhance this activity (von Kriegstein et al., 2008). Together, this neurophysiological research is consistent with the behavioral findings showing that (a) both auditory and visual speech provide closely connected speech-speaker information—perhaps in the form of talker-specific phonetic information; (b) this information can be used by both modalities to facilitate speech and speaker recognition; and (c) this information may take an amodal form allowing the modalities to share what is gleaned from unimodal experience with a talker.

The salience of both sinewave and point-light speech signals exemplifies a second commonality in how audio and visual speech is used. Both signals isolate the time-varying dimensions of their modal information, showing that the speech function can use this type of information on its own. In fact, other research shows that the time-varying portions of each signal may be more salient than the steady-state, "canonical" portions (e.g., Strange, Jenkins, & Johnson, 1983; Yakel et al., 2000). There is also evidence that cortical areas around posterior superior temporal sulcus respond similarly to point-light (but not static, photographic) visual speech and auditory speech supporting a cross-modal sensitivity to time-varying dimensions of the signals (Allison, Puce, & McCarthy, 2000; Haxby, Hoffman, & Gobbini, 2002; Puce, Allison, Bentin, Gore, & McCarthy, 1998; Santi, Servos, Vatikiotis-Bateson, Kuratate, & Munhall, 2003; Zhu & Beauchamp, 2017).

A third way in which the speech function uses auditory and visual speech similarly is as a prime for a production response (for a review, see Dias & Rosenblum, 2015). Behavioral research shows that during interlocution (and laboratory simulations), a talker's uttered response will spontaneously incorporate subtle aspects of the person's speech the talker has just heard (e.g., Pardo, 2006; Pardo, 2013), or lipread (e.g., Dias & Rosenblum, 2015; Miller et al., 2010). This phenomenon of *speech alignment* shows that we inadvertently imitate the speech of the individual with whom we are speaking. While this imitation likely has some social-psychological purpose, its proximate cause may be related to perceived and produced speech using a common metric that is related to articulatory dynamics (e.g., Fowler, 2004). Neurophysiologically, there is substantial evidence for spontaneous reactivity in motor cortical areas when perceiving auditory speech (for reviews, see

Rosenblum et al., 2016; Smalle, Rogers, & Mottonen, 2015). Both premotor and motor cortical areas are activated when passively listening to speech (for a review, see Iacaboni, 2008). Other evidence shows that the same areas of precentral gyrus are activated when either producing or perceiving the same *specific* phonetic segments. These activation patterns seem to also manifest in associated articulatory musculature (with TMS priming of motor areas; e.g., Nuttall, Kennedy-Higgins, Hogan, Devlin, & Ada, 2016). While there is ongoing debate over whether motor area involvement is *necessary*, or even *facilitative*, for speech perception (for a review, see Rosenblum et al., 2016), it is clear that motor system reactivity does occur.

Returning to behavioral research, alignment to auditory speech has been observed for over twenty years (e.g., for a review, see Pardo, Urmanche, Wilman, & Wiener, 2017). More recent research shows that alignment can also occur to visual speech, even for observers with no formal lip-reading experience (for a review, see Dias & Rosenblum, 2015). Talkers will inadvertently produce words that are (acoustically) similar to the specific words they have just lip-read (Miller, Sanchez, & Rosenblum, 2010). In fact, the degree of alignment to lip-read speech seems as great as that to heard speech (Miller et al., 2010). Relatedly, adding visual speech information to auditory speech enhances the degree to which talkers align (Dias & Rosenblum, 2011, 2015) Finally, presenting temporally incongruent (in voice onset time) audio and visual speech shapes talkers' production responses so that they reflect the *integrated* steams. This finding has been interpreted as evidence that audiovisual integration occurs before it influences spontaneous production responses (Sanchez, Miller, & Rosenblum, 2010; and see below).

There is also neurophysiological research showing an induction of motor areas by visual and audiovisual speech, similar to that of auditory speech (e.g., Callan et al., 2004; Calvert & Campbell, 2003; and for a review, see Rosenblum et al., 2016). Interestingly, there is some evidence for *greater* motor area reactivity for audiovisual—versus auditory or visual-alone—speech stimuli (e.g., Skipper, Nusbaum, & Small, 2005) that may be segment-specific based on the integrated streams (Skipper, van Wassenhove, Nusbaum, & Small, 2007). This and related findings have led some researchers to argue that motor area involvement is critical for the actual *integration* of the auditory and visual speech streams (Skipper et al., 2007; but see Rosenblum et al., 2016). Regardless, audiovisual and visual speech seem to show a similar pattern in inducing activity in motor areas—a pattern consistent with the behavioral findings on speech alignment.

In sum, there are unique ways in which visual speech seems to be treated very similarly to auditory speech. The speech function seems to use a common strategy for both streams in incorporating talker-specific and dynamic aspects of the signals allowing for the priming of a production response influenced by these dimensions. The neurophysiological basis of these characteristics seems to make use of common mechanisms for the auditory and visual streams (e.g., Calvert et al., 1997; and for a review, see Rosenblum et al., 2017).

Based on these and other considerations, some researchers have proposed that multisensory speech perception works through a functional mechanism that is agnostic with regard to sensory modality (e.g., Chandrasekaran et al., 2009; Fowler, 2004; Rosenblum et al., 2016; Summerfield, 1987). From this *supramodal* (or *modality-neutral*) account, the speech function extracts a common form of higher-order information existent in multiple energy arrays (acoustic; optic). As an example, the reversing of lip and jaw motions during an /aba/ articulation is accompanied by a reversal of the acoustics signal's amplitude and spectral structure. At the same time, this articulatory reversal structures the optic signal in such a way that it reveals a reversal with the same form and rate changes as that in the acoustic signal (Summerfield, 1987). In this sense, the information for reversal is supramodal. The speech perception function would then be tasked with extracting this supramodal rate and form information existing in both media.

More formal examples of supramodal information have been proposed. These examples include observation of high correlations between vocal tract configurations, acoustic signal, and visible mouth movements that capture up to 85% of the variance (Munhall & VatikiotisBateson, 2004; Yehia, Kuratate, & Vatikiotis-Bateson, 2002; Yehia, Rubin, & Vatikiotis-Bateson, 1998). In fact, these correlations seem to be especially high for acoustic energy in the 2–3 kHz range, despite it being the range where the supposed *less* visible gestures (tongue and pharynx positions) play their largest role (Chandrasekaran et al., 2009). This fact is consistent with the bourgeoning research showing that presumably "hidden" articulatory dimensions (e.g., lexical tone and intraoral pressure) are, in fact, visible from a talking face (e.g., Burnham, Ciocca, Lauw, Lau, & Stokes, 2000; Munhall & Vatikiotis-Bateson, 2004).

The supramodal account holds a number of important implications for theories of multisensory speech perception. If the relevant information is functionally the same across the two streams, then in an important way, the streams are never really separate. If this is true, then sensory "integration" is a function of the information *itself*, not a goal of the perceptual process (e.g., Rosenblum et al., 2016). While it is acknowledged that this is not the typical way of understanding multisensory perception, the aforementioned evidence for (a) multisensory reactivity in primary sensory cortices; which (b) treats the sensory channels in a similar way; (c) allowing for a cross-sensory sharing of experience, is supportive of the account. In addition, the approach does have much in common with popular task-machine/metamodal accounts of multisensory perception (e.g., Pascual-Leone & Hamilton, 2001; Reich, Maidenbaum, & Amedi, 2012; Ricciardi, Bonino, Pellegrini, & Pietrini, 2014; Striem-Amit et al., 2011) and is supported by evidence for those accounts, as well (for a review, see Rosenblum et al., 2017).

Importantly, the supramodal theory makes additional predictions that can be readily tested. For example, because integration is considered a function of the modality-neutral informational form, it should be revealed as occurring as early in the perceptual and neurophysiological process as methodologies can detect. The research literature that addresses this question has been extensively reviewed elsewhere (e.g., Rosenblum, 2005) and will not be the focus of this article. Instead, this article will review the literature addressing

two other assumptions that have come to distinguish the supramodal approach: integration should be functionally *complete* at that early stage and *impenetrable* from outside cognition.

However, most other theories of multisensory speech assume that integration is not a consequence of amodal information but must occur through standard cognitive processes. Consequently, these integration processes are thought to be inferential, statistical, and fallible, resulting in integration that is often incomplete and susceptible to outside influences. This fact provides a testable distinction between supramodal and cognitive accounts. As discussed in section 4, a great deal of research over the last 10 years has been designed to address the completeness and impenetrability issues.

4. How Complete Is Audiovisual Speech Integration?

One of the most enduring questions about audiovisual speech is how completely the audio and visual channels are combined. Of course, the answer to this question will depend on the operational definitions of “combined” and “completely,” and both of these terms have been interpreted in somewhat different ways in the literature (for discussions of these topics, see Brancazio & Miller, 2005). However, here, we will take the term “combined” to simply refer to any behavioral evidence for bimodal or cross-modal influences on speech perception. “Completeness” will be defined as the degree to which any remnants of the unimodal information show a perceptual influence.

It is also important to distinguish the concept of perceptual completeness from that of perceptual *clarity*—especially in the context of the McGurk effect. There is a good deal of research showing that the speech perceived from incongruent audiovisual presentations is often not as clear or strong as speech derived from audiovisually congruent, or auditory alone, segments (e.g., Brancazio, 2004; Brancazio, Best, & Fowler, 2006; Green & Kuhl, 1991; Jerger, Damian, M. F., Tye-Murray, N., & Abdi, 2017; Massaro & Ferguson, 1993; Rosenblum & Saldana, 1992). Unsurprisingly, conflicting audiovisual information is likely to make the resulting perceived segment less canonical, and this has been shown through both matching judgments and response reaction time measures (which lengthen with more ambiguity). This fact may mean that there will be times when the streams *do* combine and still produce a segment that is perceived as closer to the one in the auditory stream alone. These occurrences would then appear to be failures of the McGurk effect, and be erroneously interpreted as a failure of integration (e.g., Brancazio, 2004; Brancazio & Miller, 2005). This is a critical point for both theory and method, and will be discussed in detail later in this section.

For example, when subjects are asked to shadow audio /aba/ + visual /aga/, they sometimes say—and report perceiving—the auditorily dominated “aba” (Gentilucci & Cattaneo, 2005; and see Sato, Buccino, Gentilucci, & Cattaneo, 2010). Typically, this type of response would be interpreted as the McGurk effect failing, and that the visual speech in-

formation did not combine with auditory information. However, in such instances, subjects' verbal "aba" response will often show subtle articulatory movements that reflect aspects of the ostensibly ignored visual /aga/ (Gentilucci & Cattaneo, 2005). This outcome has been explained by assuming that the streams are never integrated but can individually influence the motoric response (Gentilucci & Cattaneo, 2005).

However, this type of covert effect could instead reflect the influence of the *combined* modalities. As stated, incongruent audiovisual segments are known to result in perceived segments that are less canonical. Thus, there will be times when the resultant combined segment is perceptually categorized—and identified—as being the same as the auditory component. However, the ambiguity inherent to the combined segment might induce changes in the (presumably) more sensitive motoric measures that reflect the combined segment. In this sense, identifying an audiovisually incongruent stimulus as in accord with the auditory-alone component does not imply a lack of fusion.

A similar explanation can be applied to another shadowing study in which the McGurk effect *does* occur, but remnants of the (presumed) unimodal components are observable in the articulatory response (Gentilucci & Cattaneo, 2005). When subjects are asked to shadow a stimulus composed of audio /aba/ and visual /aga/, they will often articulate (and perceive) a fused /ada/. However, analysis of the articulation will show movements (and acoustic outcomes) that reflect aspects of the individual /aba/ and /aga/ components. These findings have been interpreted as evidence that either the components are never fully integrated or that the details of the shadowing response tap into a stage before integration (Gentilucci & Cattaneo, 2005).

Again however, these findings may simply be based on a shadowing response that reflects the ambiguity in the perceived *combined* segment. A shadowed response would likely reflect this ambiguity, such that articulatory deviations may sometimes resemble aspects of the individual audio and visual components. Thus, despite appearing as being influenced by the unimodal components, the articulations may actually reflect the ambiguous *combined* segment. If this fact is true, there should be instances for which articulatory deviations do not resemble the actual unimodal components. Unfortunately, no study to date has examined whether the articulatory kinematics of shadowed responses to McGurk stimuli more accurately reflect the combined *but ambiguous* segment (/ada/), or the unimodal components (/aba/ and /aga/).

There are, however, studies testing "covert" motor system reactivity to McGurk stimuli. These studies have, in fact, provided evidence that it is actually the *fused* segment that influences the motor system. Skipper et al. (2007) used fMRI to observe that when being presented an audio /pa/ with a visual /ka/, activity in motor areas resembled that of production for a /ta/—the fused syllable typically perceived. Similar findings have been reported using TMS priming on the motor brain to reveal articulator EMG activity that follows visually influenced perception (Sato et al., 2010; Sundara, Namasivayam, & Chen, 2001). These findings certainly suggest that the streams are combined by the time a motor response is initiated. The findings also challenge the interpretation (Gentilucci & Cat-

taneo, 2005) that motor responses to incongruent audiovisual speech contain remnants of the separate audio and visual streams in their articulation.

There is an additional, critical implication of all of these results: the McGurk effect may not be an accurate index of integration (see also Alsius et al., 2018). As stated, it is quite possible that the streams actually do combine in many instances for which the McGurk effect fails. In fact, compelling evidence for this possibility has been provided in an elegant study conducted by Brancazio and Miller (2005; see also MacDonald, Andersen, & Bachmann, 2000). These authors made use of a prior finding that the visible articulatory rate of a consonant (/pi/) can influence the perceived voice onset time (VOT) of a synchronously presented auditory consonant. This influence was evident in how subjects categorized auditory stimuli along a /bi/ to /pi/ continuum which varied in VOT. Specifically, seeing a more rapid articulation of /pi/ would make it more likely that an ambiguous token along the continuum would be perceived as /pi/. Seeing a slower articulation of /pi/, on the other hand, would induce that same ambiguous token to sound like /bi/.

Upon finding an analogous effect with visible /ti/'s paired with auditory tokens from a /ti/-/di/ continuum, the authors attempted to induce a classic McGurk effect with these stimuli. For this purpose, they combined tokens from their auditory /pi/-/bi/ continuum with the visible tokens of /ti/ spoken at different rates. A classic McGurk effect would be said to occur in instances of a (visually dominated) /ti/ or /di/ response. Brancazio and Miller (2005) did find this type of McGurk effect on about half of the trials. However, even when the McGurk effect failed to occur, listeners categorized the tokens along the VOT continuum in a way dependent on the rate of the spoken visual token. In other words, whether or not subjects provided a classic McGurk effect response, their perception of auditory VOT (and associated categorization of the tokens) was still influenced by the visible rate of the /ti/ syllable. This suggests that *combining the audio and visual streams still occurs even for cases in which a classic McGurk effect fails* (see also MacDonald et al., 2000). The finding is also consistent with the aforementioned interpretation that observing covert evidence for a visual influence when a McGurk effect fails (e.g., Gentilucci & Cattaneo, 2005; Sato et al., 2010) may actually reflect an influence that is based on a *combined* segment.

Methodologically, finding evidence for the fusing of audio and visual information in the face of a "failed" McGurk effect is critically important. The finding suggests that simple identification responses to McGurk stimuli may be a limited—and even *misleading* means to evaluate how bimodal speech is combined. If this is true, then it may also be misguided to use the McGurk effect as a measure of subject differences in multisensory integration, as such (see above). It is possible that assuming extraction of the requisite information across streams, *all* observers *always* combine auditory and visual speech information. However, depending on a subject's linguistic and perceptual background, they may categorize that combined information as being more similar to the segment contained in auditory stream only, thereby failing to show the classic McGurk effect. It could be, then, that a more sensitive behavioral, motoric, and neurophysiological measure would show that the same subject does, in fact, combine the streams at a more fine-grained level. In this

sense, subject differences may reflect more how the fused, common information is categorized than any differences in how the streams are combined, as such.

This reevaluation of the McGurk effect also has theoretical import. Recall that the supramodal account proffers that streams that share informational commonalities, are—to some degree—always combined. This account argues that the combining of streams is a consequence of extraction of common information existent across the signals. As long as an observer is able to attend to that information in both modalities, then "integration" naturally occurs, and some evidence for combining should be observable. True subject differences, evident in more sensitive covert measures, would then be a consequence of the degree to which a subject can attend to the amodal information available across signals. With regard to the more overt McGurk effect measure, attention to common information would also bear on phonetic identification, as would the subject's categorization of fused segment. Regardless, the long-held assumption of the McGurk effect as the quintessential example of audiovisual speech integration should be reconsidered (see also Alsius et al., 2018).

A final result interpreted as demonstrating the incomplete combining of the audio and visual streams involves the phenomenon of semantic priming (Ostrand, Blumstein, Ferreira, & Morgan, 2016). They reported evidence that while identification of words is influenced by multisensory information, the semantic content of that word *is not* based on the integrated information. Their method involved presenting McGurk effect (McGurk & MacDonald 1976) words (e.g., Audio "bait" & Visual "date" which putatively produce the "heard" illusory perception of "date") as priming stimuli. These audiovisual words were presented as primes to test their facilitation of semantically related auditory-only target words. The question was whether the integrated words more strongly facilitate target words associated with the *integrated* primes ("date"➔time) or the *auditory* component of those primes ("bait"➔worm). Ostrand et al. (2016) found that only the auditory component of the McGurk stimulus ("bait") facilitated identification speed of the related audio-alone targets ("worm"). These findings have been interpreted as showing that auditory and visual speech integration is incomplete, at least up to the point that semantic processing begins (Ostrand et al., 2016).

However, a more recent test of this question has provided very different results (Dorsi, Rosenblum, & Ostrand, 2017). This new project used audiovisual segment combinations known to produce more compelling McGurk effects (e.g., audio "boat" + Visual "vote" = heard "vote"), than the prior project. Using these stimuli, results revealed that it was the *integrated word* that more strongly primed semantically related targets ("vote"➔election). Follow-up tests revealed that the degree to which an audiovisual stimulus primed based on the integrated versus auditory correlated with how strongly the McGurk effect worked with that particular stimulus. This result may suggest that the predominance of auditory-word priming reported by Ostrand et al. (2016) could reflect failures of their stimuli to induce strong McGurk effect perception. Regardless, this new

study suggests that the combining of streams is at least complete enough to induce semantic priming based on that combination.

In sum, there are a number of studies that, at first pass, seem to indicate that audiovisual speech integration does not completely use the common information available across the senses, and that influences of the individual streams remain. However, careful consideration and evaluation of the existing findings and methodologies indicates that this evidence is far from conclusive. Most critically, the much relied upon McGurk methodology appears to be misleading with regard to the issue. A similar story is emerging from the research on impenetrability of multisensory speech perception.

5. How Impenetrable Is Multisensory Speech Perception?

Early research on multisensory speech integration supported an automatic, impenetrable function. Much of this support came from findings that perceivers are influenced by the visual component of a McGurk stimulus even when told of the manipulation and instructed to report only the auditory component (Bertelson & de Gelder, 2004; Bertelson, Vroomen, Wiegeraad, & de Gelder, 1994; Colin et al., 2002; Green & Kuhl, 1991; Soto-Faraco & Alsius, 2009).

However, newer research has challenged the assumption of impenetrability and may suggest that both extraperceptual linguistic and non-linguistic factors can bear on integration. For example, visual influences on perceived segments are greater if that segment is part of a word than part of a non-word. If, for example, audio /ba/ is paired with visual /va/, it is perceived more often as ‘va’ when presented in the context of the word “valve” than in the non-word “vatch”: Brancazio, 2004; Barutchu, Crewther, Kiely, Murphy, & Crewther, 2008; but see Sams, Manninen, Surakka, Helin, & Kättö, 1998). In a similar way, the semantic context provided by a carrier sentence can affect how likely a visual influence is reported (e.g., Windmann, 2004, 2008). Because lexical and semantic processing are generally considered to occur later in the linguistic process, evidence for such influences may challenge the assumptions of impenetrable (and early) audiovisual integration.

However, other explanations have been provided for these findings (e.g., Brancazio 2004; Rosenblum, 2008). It might be that these downstream influences do not bear on *integration as such* but instead on the categorization of segments after the streams have been combined. As stated, perceived segments based on audiovisually incongruent stimuli tend to be less strong (e.g., Rosenblum & Saldana, 1992). This fact would open such segments to the downstream influences of lexical and semantic context—influences also known to affect ambiguous auditory-only segments (e.g., Connine & Clifton, 1987; Ganong, 1980). In fact, lexicality shows similar influences on ambiguous (e.g., noisy) auditory segments and incongruent multisensory segments, such that its influences are greater with longer response delays (Brancazio, 2004). This could mean that lexicality does not act on the in-

tegration process, as such, but instead on the ambiguous nature of a (already combined) segment composed of incongruent streams.

There also is evidence that attentional factors can influence perception of audiovisual speech. For example, only observers primed to hear sinewave simulations of auditory speech stimuli as speech (but not as nonspeech) will show a McGurk-type visual influence (Tuomainen, Andersen, Tiippana, & Sams, 2005). Also, asking subjects to attend to a visual distractor placed over the face reduces the McGurk effect significantly (Tiipanna, Andersen, & Sams, 2004; see also Munhall, Ten Hove, Brammer, & Paré, 2009). Other research shows that distractors presented in other modalities, including auditory and tactile, influence the consistency of the McGurk effect (e.g., Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Alsius, Alsius, Navarra, & Soto-Faraco, 2007; Mitterer & Reinisch, 2017). Importantly, in two separate studies, these attention tasks have been reported to influence *only* the McGurk effect: *not unimodal* (auditory or visual) speech identification. Following from this, it has thus been argued that attentional influences are not simply the result of depleted unimodal resources but instead have a direct influence on the integration of the modalities, as such (e.g., Alsius et al., 2005; Alsius et al., 2007; Mitterer & Reinisch, 2017; Navarra, Alsius, Soto-Faraco, & Spence, C. 2010; and see Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010). In this sense, these results clearly challenge the notion that the combining of streams is impenetrable.

However, alternative explanations for these findings exist. Perhaps attentional demands, regardless of their nature or modality, *do* actually suppress extraction of *unimodal* information, but in a way that is only reflected in McGurk effect responses. Consider the following possibility. For hearing observers, the visual modality is likely the more fragile for speech perception, and is more likely to be influenced by attentional demands, regardless of the source of distraction. In principle, it is true that these attentional demands should also influence unimodal visual-alone speech performance. As stated, however, studies have failed to find attentional influence on visual-alone responses, leading to the conclusion that attention acts directly on audiovisual integration (Alsius et al., 2005; Alsius et al., 2007).

However, upon re-evaluation of these studies, it seems possible that the visual-alone conditions used were not sensitive enough to show any effects. In both studies, visual-alone (lip-reading) identification was so poor (2–12% correct) that it is unlikely a distraction manipulation could have any meaningful influence on performance. Future research using easier visual-alone tasks could examine whether the distraction manipulations that influence the McGurk effect also influence visual-alone speech extraction. This research might also benefit from using use more covert measures (motoric, neurophysiological, VOT) to possibly reveal that perception of unimodal segments is affected by distraction. If distraction manipulations were found to influence both visual speech performance as well as McGurk responses, then attentional factors may not bear on the integration process itself. Instead, distraction may simply suppress visual information extraction, which in turn can change categorization of the combined information leading to a suppressed McGurk effect (see above). This would mean that distraction does not bear on the integration

function, as such. In fact, there is recent evidence using visual-world measures that in the context of distraction from visual speech, whatever visual information *is* extracted, is used in a seemingly automatic way (Mitterer & Reinisch, 2017).

A similar account can address findings showing how a preceding *bimodal* context can influence McGurk responses. A number of studies have shown that by priming subjects with audiovisually mismatched speech, subjects will show a smaller tendency to display the McGurk effect (Gau & Noppeney, 2016; Ganesh, Berthommier, Vilain, Sato, & Schwartz, 2014; Nahorna, Berthommier, & Schwartz, 2012, 2015). In these experiments, subjects are presented with a string of audiovisually incongruent sentences or syllables and are then asked to monitor for either a ‘ba’ or ‘da’ syllable. Critically, these target syllable are always comprised of an audio /ba/ - video /ga/ which is perceived as a ‘da’ when the McGurk effect occurs, and a ‘ba’ when it does not. Results show that when the target syllable is preceded by even a very short span of incongruent sentences or syllables, the frequency of McGurk effect ‘da’ responses is reduced (relative to when the preceding context is comprised of audiovisually *congruent* sentences or syllables).

Authors of these studies argue that the incoherence of the preceding context suppresses binding of the audio and visual components, thereby preventing integration and reducing the McGurk effect. Proponents also cite studies that ostensibly establish a neurophysiological basis for the interaction of coherence context and the McGurk effect (Gau & Noppeney, 2016; Ganesh et al., 2014). Regardless, because it seems the McGurk effect can be influenced by a preceding binding context, proponents argue that integration should not be considered an impenetrable and automatic process.

Importantly, considering multisensory perception as a two-stage process—involving first binding and then integration—has been discussed in a number of speech and nonspeech domains (e.g., Berthommier, 2004; Bregman, 1990; and for a review, see Chen & Spence, 2017). Often described as the “unity assumption,” it has been argued that in order for integration to occur, an initial evaluation of the streams is required to determine whether they indicate the same distal event. The outcome of this evaluation then provides a top-down influence on whether actual integration proceeds. Modern proponents of the unity assumption cite the incoherent context influences on the McGurk effect as support (Chen & Spence, 2017).

However, another interpretation of incoherent context influences can be offered that is more consistent with an impenetrable multisensory function. It could be that a preceding incoherent context simply serves as a distraction to deplete attentional resources from the extraction of unimodal speech information. As discussed above, a reduction in visual speech extraction would provide a smaller visual contribution to the combined information. This occurrence would then more likely induce a recovered segment that is phonetically categorized as the same as the auditory component (i.e., the McGurk effect would fail). Thus, rather than directly suppressing the binding and integration processes as such, an incoherent context may simply distract from using of all available visual informa-

tion. This interpretation would allow for the combining of information to be an impenetrable function.

If the incoherent preceding context does simply serve as a distractor from visual information extraction, then it should not just bear on performance with bimodal target stimuli but also on performance with unimodal stimuli (e.g., reducing accuracy or speed in categorizing visual-alone segments). Unfortunately, none of the studies demonstrating incoherent context effects (Gau & Noppeney, 2016; Ganesh et al., 2014; Nahorna et al., 2012, 2015) tested unimodal target stimuli. Thus, as with the aforementioned studies on distraction, it is difficult to reach any conclusion concerning whether attention bears on integration as such. This problem is exacerbated by the consistent reliance on the McGurk effect in these demonstrations. As stated, there are severe problems with using the McGurk effect as an index of integration. Recall the evidence that fusion can occur even when the McGurk effect does not (e.g., Brancazio & Miller, 2005), and the strength of the effect may reflect post-integration categorization instead of any characteristics of the integration processes itself. Thus, while incoherent preceding context disruption effects have been interpreted as showing direct influences on (penetrable) integration, much more research is needed before this interpretation is strongly supported.

In sum, while a number of findings have been interpreted as supportive of a highly penetrable integration function, it is clear that more research is needed before this conclusion can be legitimately accepted. Insufficient unimodal tests and reliance on the problematic McGurk effect methodology preclude any clear conclusions on this question. It could very well be that attention bears on unimodal information extraction and that the combining of the streams itself is impenetrable.

6. Retiring the McGurk Effect

Arguably, work on multisensory speech has been one of the most active areas of perceptual psychology over the last 20 years. This research has contributed not only to the speech perception but also to our more general understanding of information integration, perceptual learning, and the overall architecture of the brain. Profoundly, audiovisual speech research has been critical in our new understanding of the perceptual brain as being built around multisensory input.

Certainly, the McGurk effect has been instrumental in the prevalence of multisensory speech research. Ironically, however, it may be just this phenomenon that has kept us from understanding the basic operations of multisensory speech perception. By mistakenly conflating the McGurk effect with speech integration itself, interpretations of the completeness and automaticity of multisensory may be incorrect. Future research should use more sensitive behavioral and neurophysiological measures of cross-modal influence. These measures can be used to test whether multisensory speech perception results from extraction of supramodal information across signals or the combining of modality-specific information through more standard cognitive processes.

Further Reading

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

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). **fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect**. *Journal of Neuroscience*, *30*(7), 2414–2417. doi:10.1523/JNEUROSCI.4865–09.2010 PMID: 20164324

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

Ghazanfar, A. A., & Schroeder, C. E. (2006). **Is neocortex essentially multisensory?**. *Trends in Cognitive Sciences*, *10*(6), 278–285. doi:10.1016/j.tics.2006.04.00

Mitterer, H., & Reinisch, E. (2017). Visual speech influences speech perception immediately but not automatically. *Attention, Perception, & Psychophysics*, *79*(2), 660–678.

Reich, L., Maidenbaum, S., & 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.

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

Rosenblum, L. D., Dorsi, J., & Dias, J. W. (2016). The impact and status of Carol Fowler's supramodal theory of multisensory speech perception. *Ecological Psychology*, *28*, 262–294.

Shams, L. (2011). Early integration and Bayesian causal inference in multisensory perception. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes* (pp. 217–232). Boca Raton, FL: CRC Press.

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 580–587.

References

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Science*, *4*, 267–278.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*(9), 839–843.

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, *183*(3), 399–404.

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

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., & Sams, M. (2009). **The role of visual spatial attention in audiovisual speech perception**. *Speech Communication*, *29*, 184–193. doi:10.1016/j.specom.2008.07.004

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

Barutchu, A., Crewther, S. G., Kiely, P., Murphy, M. J., & Crewther, D. P. (2008). When/b/ill with/g/ill becomes/d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology*, *20*(1), 1–11

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). **fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect**. *Journal of Neuroscience*, *30*(7), 2414–2417. doi:10.1523/JNEUROSCI.4865–09.2010 PMID: 20164324

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

Bertelson, P., & de Gelder, B. (2004). The psychology of multi-sensory perception. In C. Spence & J. Driver (Eds.), *Cross-modal space and cross-modal attention* (pp. 155–171). Oxford, U.K.: Oxford University Press.

Bertelson, P., Vroomen, J., Wiegeraad, G., de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. *Proceedings of the International Congress on Spoken Language Processing* (pp. 559–562). Yokohama, Japan: Acoustical Society of Japan.

Berthommier, F. (2004). A phonetically neutral model of the low-level audio-visual interaction. *Speech Communication*, *44*(1), 31–41.

Bicevskis, K., Derrick, D., & Gick, B. (2016). Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives. *The Journal of the Acoustical Society of America*, *140*(5), 3531–3539.

Borrie, S. A., McAuliffe, M. J., Liss, J. M., O'Beirne, G. A., & Anderson, T. J. (2013). The role of linguistic and indexical information in improved recognition of dysarthric speech. *The Journal of the Acoustical Society of America*, *133*, 474–482.

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

Brancazio, L., Best, C. T., & Fowler, C. A. (2006). Visual influences on perception of speech and nonspeech vocal-tract events. *Language and speech*, *49*(1), 21–53.

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

Bregman, A. S. (1990). *Auditory scene analysis* (Vol. 10). Cambridge, MA: MIT Press.

Brouwer, K., Gordon-Pershey, M., Hoffman, D., & Gunderson, E. (2015). Speech sound–production deficits in children with visual impairment: A preliminary investigation of the nature and prevalence of coexisting conditions. *Contemporary Issues in Communication Science and Disorders*, *42*, 33–46.

Burnham, D., Ciocca, V., Lauw, C., Lau, S., & Stokes, S. (2000). Perception of visual information for Cantonese tones. In *The 8th Australian International Conference on Speech Science & Technology, Canberra, Australia* (pp. 86–91).

Callan, D. E., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2001). **Multimodal contribution to speech perception revealed by independent component analysis: A singlesweep EEG case study**. *Cognitive Brain Research*, *10*, 349–353. doi:10.1016/S0926-6410(00)00054-9

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, *16*(5), 805–816.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., & David, A. S. (1997). **Activation of auditory cortex during silent lipreading**. *Science*, *276*(5312), 593–596. doi:10.1126/science.276.5312.59

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*, 57–70.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). **The natural statistics of audiovisual speech**. *PLoS Computational Biology*, *5*(7), 1–18. doi:10.1371/journal.pcbi.1000436

Chen, Y. C., & Spence, C. (2017). Assessing the role of the “unity assumption” on multisensory integration: A review. *Frontiers in Psychology*, *8*, 445–455.

Colin, C., Radeau, M., Deltenre, P., Demolin, D., & Soquet, A. (2002). The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. *European Journal of Cognitive Psychology*, *14*, 475–491.

Connine, C. M., & Clifton Jr., C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(2), 291.

Dias, J. W., Cook, T. C., & Rosenblum, L. D. (2017). The McGurk effect and the primacy of multisensory perception. (pp. 791–796) In A. Shapiro & D. Todorovic (Eds.), *The Oxford compendium of visual illusions*. Oxford, U.K.: Oxford University Press.

Dias, J. W., & Rosenblum, L. D. (2011). Visual influences on interactive speech alignment. *Perception*, *40*, 1457–1466.

Dias, J. W., & Rosenblum, L.D. (2015). **Visibility of speech articulation enhances auditory phonetic convergence**. *Attention, Perception & Psychophysics*, *78*, 317–333. doi: 10.3758/s13414-015-0982-6

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121–144.

Dorsi, J., Rosenblum, L. D., & Ostrand, R. (2017, November 10). *What you see isn't always what you get, or is it? Reexamining semantic priming from Mcgurk stimuli*. Poster presented at the 58th meeting of the Psychonomics Society, Vancouver, Canada.

Fowler, C. A. (2004). Speech as a supramodal or amodal phenomenon. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 189–201). Cambridge, U.K.: Cambridge University Press.

Fuster-Duran, A. (1996). Perception of conflicting audio-visual speech: An examination across Spanish and German. In *Speechreading by humans and machines* (pp. 135–143). Berlin, Germany: Springer Berlin Heidelberg.

Ganesh, A. C., Berthommier, F., Vilain, C., Sato, M., & Schwartz, J. L. (2014). A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. *Frontiers in Psychology*, *5*, 1–13.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. *NeuroImage*, *124*, 876–886.

Gentilucci, M., & Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. *Experimental Brain Research*, *167*(1), 66–75.

Ghazanfar, A. A., & Schroeder, C. E. (2006). **Is neocortex essentially multisensory?**. *Trends in Cognitive Sciences*, *10*(6), 278–285. doi:10.1016/j.tics.2006.04.008

Green, K. P., & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 278–288.

Hardison, D. M (2005). Variability in bimodal spoken language processing by native and nonnative speakers of English: A closer look at effects of speech style. *Speech Communication*, *46*, 73–93

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry*, *51*, 59–67.

Hazan, V., Sennema, A., Iba, M., & 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. doi:10.1016/j.specom.2005.04.007

Hickok, G. (2009). **Eight problems for the mirror neuron theory of action understanding in monkeys and humans**. *Journal of Cognitive Neuroscience*, *21*(7), 1229–1243. doi:10.1162/jocn.2009.21189

Iacoboni, M. (2008). The role of premotor cortex in speech perception: Evidence from fmri and rtms. *Journal of Physiology-Paris*, *102*(1), 31–34.

Irwin, J. R., Whalen, D. H., & Fowler, C. A. (2006). A sex difference in visual influence on heard speech. *Perception & Psychophysics*, *68*(4), 582–592.

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

Jerger, S., Damian, M. F., Tye-Murray, N., & Abdi, H. (2017). Children perceive speech onsets by ear and eye. *Journal of child language*, *44*(1), 185–215.

von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A.-L., Kell, C. A., Grüter, T.,. . . Kiebel, S. J. (2008). **Simulation of talking faces in the human brain improves auditory speech recognition**. *Proceedings of the National Academy of Sciences*, *105*(18), 6747–6752. doi:10.1073/pnas.0710826105

von Kriegstein, K., & Giraud, A. L. (2006). **Implicit multisensory associations influence voice recognition**. *PLoS Biology*, *4*(10), 1809–1820. doi:10.1371/journal.pbio.0040326

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*(3), 367–376.

Lachs, L., & 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. doi:10.1121/1.1757454

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology*, *61*, 961–967.

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: How much spatial degradation can be tolerated? *Perception*, *29*(10), 1155–1168.

Magnotti, J. F., & Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Computational Biology*, *13*(2), e1005229.

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

Massaro, D. W., & Ferguson, E. L. (1993). Cognitive style and perception: The relationship between category width and speech perception, categorization, and discrimination. *The American Journal of Psychology*, *106*(1), 25–49.

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.

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

Ménard, L., Cathiard, M. A., Troille, E., & Giroux, M. (2015). Effects of congenital visual deprivation on the auditory perception of anticipatory labial coarticulation. *Folia Phoniatrica et Logopaedica*, *67*(2), 83–89.

Ménard L., Dupont S., Baum S. R., & Aubin J. (2009). Production and perception of French vowels by congenitally blind adults and sighted adults. *Journal of the Acoustical Society of America*, *126*, 1406–1414.

Ménard L., Leclerc A., & Tiede M. (2014). Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults. *Journal of Speech and Hearing Research*, *57*, 793–804.

Ménard L., Toupin C., Baum S., Drouin S., Au-bin J., & Tiede M. (2013). Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults. *Journal of the Acoustical Society of America*, *134*, 2975–2987.

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception, & Performance*, *72*, 1614–1625.

Mills, A. E. (1987). The development of phonology in the blind child. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 145–162). Hillsdale, NJ: Erlbaum.

Mitterer, H., & Reinisch, E. (2017). Visual speech influences speech perception immediately but not automatically. *Attention, Perception, & Psychophysics*, *79*(2), 660–678.

Mottonen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research*, *13*, 417–425.

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

Munhall, K. G., Ten Hove, M. W., Brammer, M., & Paré, M. (2009). Audiovisual integration of speech in a bistable illusion. *Current Biology*, *19*(9), 735–739.

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006) Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, *168*, 1–10.

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America*, *132*(2), 1061–1077.

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2015). Audio-visual speech scene analysis: Characterization of the dynamics of unbinding and rebinding the McGurk effect. *The Journal of the Acoustical Society of America*, *137*(1), 362–377.

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

Nath, A. R., & Beauchamp, M. S. (2012). **A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion**. *NeuroImage*, *59*(1), 781–787. doi:10.1016/j.neuroimage.2011.07.024 PMID: 21787869

Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion*, *11*(1), 4–11.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of L2 sounds. *Psychological Research*, *71*, 4–12.

Nuttall, H. E., Kennedy-Higgins, D., Hogan, J., Devlin, J. T., & Adank, P. (2016). The effect of speech distortion on the excitability of articulatory motor cortex. *NeuroImage*, *128*, 218–226.

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

Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition*, *151*, 96–107.

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

Pardo, J. S. (2013). **Measuring phonetic convergence in speech production**. *Frontiers in Psychology*, *4*, 1–5.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637–659.

Pascual-Leone, A., & Hamilton, R. (2001). The metamodal organization of the brain. *Progress in Brain Research, 134*, 1–19.

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

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation of humans viewing eye and mouth movements. *Journal of Neuroscience*, *18*, 2188–2199.

Reich, L., Maidenbaum, S., & 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.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–114). Hillsdale, NJ: Erlbaum.

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

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947–950.

Ricciardi, E., Bonino, D., Pellegrini, S., & Pietrini, P. (2014). **Mind the blind brain to understand the sighted one! Is there a supramodal cortical functional architecture?**. *Neuroscience & Biobehavioral Reviews*, *41*, 64–77. doi:10.1016/j.neubiorev.2013.10.006

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. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409.

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

Rosenblum, L. D., Dias, J.W., & Dorsi, J. (2017). **The supramodal brain: Implications for auditory perception**. *Journal of Cognitive Psychology*, *28*, 1–23. doi:10.1080/20445911.2016.1181691

Rosenblum, L. D., Dorsi, J., & Dias, J.W. (2016). The impact and status of Carol Fowler's supramodal theory of multisensory speech perception. *Ecological Psychology*, *28*, 262–294.

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

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lipread me now, hear me better later: Crossmodal transfer of talker familiarity effects. *Psychological Science*, *18*, 392–396.

Rosenblum, L. D., & Saldaña, H. M. (1992). Discrimination tests of visually-influenced syllables. *Perception and Psychophysics*, *52*(4), 461–473.

Rosenblum, L. D., & 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.

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

Rosenblum, L. D., Smith, N. M. & Niehus, R. P. (2007). Look who's talking: Recognizing friends from visible articulation. *Perception*, *36*, 157–159.

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

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.

Sams, M., Manninen, P., Surakka, V., Helin, P., & 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., & Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception & Psychophysics*, *75*, 1359–1365.

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

Santi, A., Servos, P., Vatikiotis-Bateson, E., Kuratate, T., & Munhall, K. (2003). Perceiving biological motion: dissociating visible speech from walking. *Journal of Cognitive Neuroscience*, *15*(6), 800–809.

Sato, M., Buccino, G., Gentilucci, M., & Cattaneo, L. (2010). **On the tip of the tongue: Modulation of the primary motor cortex during audiovisual speech perception**. *Speech Communication*, *52*(6), 533–541. doi:10.1016/j.specom.2009.12.00

Schweinberger, S. R., & Soukup, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1748–1765.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in nonEnglish listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America*, *90*(4), 1797–1805.

Shams, L. (2011). Early integration and Bayesian causal inference in multisensory perception. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes* (pp. 217–232). Boca Raton, FL: CRC Press.

Simmons, D. C., Dias, J. W., Dorsi, J., & Rosenblum, L. D. (2015, May 20). *Crossmodal transfer of talker learning*. Poster presented at the 169th meeting of the Acoustical Society of America, Pittsburg, PA.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). **Listening to talking faces: Motor cortical activation during speech perception**. *NeuroImage*, *25*(1), 76–89. doi:10.1016/j.neuroimage.2004.11.006

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). **Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception**. *Cerebral Cortex*, *17*(10), 2387–2399. doi:10.1093/cercor/bhl147

Smalle, E. H., Rogers, J., & Möttönen, R. (2015). Dissociating contributions of the motor cortex to speech perception and response bias by using transcranial magnetic stimulation. *Cerebral Cortex*, *25*(10), 3690–3698

Soto-Faraco, S., Alsius, A. (2007). Conscious access to the unisensory components of a crossmodal illusion, *Neuroreport*, *18*, 347–350.

Soto-Faraco, S., Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 580–587.

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech Language and Hearing Research*, *57*, 2322–2331.

Strange, W., Jenkins, J., & Johnson, T. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, *74*(3), 694–705.

Striem-Amit, E., Dakwar, O., Hertz, U., Meijer, P., Stern, W., Pascual-Leone, A., & Amedi, A. (2011). The neural network of sensory-substitution object shape recognition. *Functional Neurology, Rehabilitation, and Ergonomics*, *1*(2), 271–278.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215.

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, U.K.: Erlbaum.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audiovisual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, *36A*, 51–74.

Sundara, M., Namasivayam, A. K., & Chen, R. (2001). **Observation-execution matching system for speech: A magnetic stimulation study**. *Neuroreport*, *12*(7), 1341–1344. doi:10.1097/00001756-200105250-00010

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410.

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

Thomas, S. M., & Jordan, T. R. (2002). Determining the influence of Gaussian blurring on inversion effects with talking faces. *Perception & Psychophysics*, *64*, 932–944.

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, *5*, 725–728.

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*(3), 457–472.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio–visual speech perception is special. *Cognition*, *96*(1), B13–B22.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language*, *50*(2), 212–230.

Windmann, S. (2008). Sentence context induces lexical bias in audiovisual speech perception. *Review of Psychology*, *14*(2), 77–91.

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

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

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). **Quantitative association of vocal-tract and facial behavior**. *Speech Communication*, *26*(1–2), 23–43. doi:10.1016/S0167-6393(98)00048-X

Zhu, L. L., & Beauchamp, M. S. (2017). Mouth and voice: A relationship between visual and auditory preference in the human superior temporal sulcus. *Journal of Neuroscience*, *37*(10), 2697–2708.

Lawrence D. Rosenblum

Department of Psychology, University of California, Riverside