Published in final edited form as:

Annu Rev Linguist. 2019 January ; 5(1): 49–66. doi:10.1146/annurev-linguistics-011718-012353.

Cross-modal effects in speech perception

Megan Keough[1], **Donald Derrick**[2,3], **Bryan Gick**[§,1,4]

[1]Interdisciplinary Speech Research Lab, Department of Linguistics, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada [2]New Zealand Institute of Brain and Behaviour, University of Canterbury, Christchurch 8140, New Zealand [3]MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith, New South Wales 2751, Australia [4]Haskins Laboratories, Yale University, New Haven, CT 06511, USA

Abstract

Speech research during recent years has moved progressively away from its traditional focus on audition toward a more multisensory approach. In addition to audition and vision, many somatosenses including proprioception, pressure, vibration and aerotactile sensation are all highly relevant modalities for experiencing and/or conveying speech. In this article, we review both long-standing cross-modal effects stemming from decades of audiovisual speech research as well as new findings related to somatosensory effects. Cross-modal effects in speech perception to date are found to be constrained by temporal congruence and signal relevance, but appear to be unconstrained by spatial congruence. Far from taking place in a one-, two- or even three-dimensional space, the literature reveals that speech occupies a highly multidimensional sensory space. We argue that future research in cross-modal effects should expand to consider each of these modalities both separately and in combination with other modalities in speech.

Keywords

cross-modal effects; multisensory; speech perception; somatosensation

1. Introduction

While speech is often described in terms of sound alone, considerable research over the last half-century – and even earlier (e.g., Bell 1867) – supports the notion that speech perception is fundamentally multisensory. Indeed, some natural languages have been found to encode sequences of speech “sounds” without any sound at all (Gick et al. 2012), bringing into relief the need for more deeply multisensory approaches. Much of the evidence that has been used in favor of a multisensory view of speech has derived from laboratory studies of cross-modal effects. Although the field has made some progress in understanding how such effects work and how to interpret them, researchers continue to grapple with the most basic questions regarding multisensory perception, i.e.: What sensory inputs are relevant to speech

[§]Corresponding author: gick@mail.ubc.ca, Department of Linguistics, University of British Columbia, Vancouver, British Columbia V6T 1Z4.

perception? What information or criteria determines when we integrate? Where in processing does integration occur? The present paper starts by reviewing some of what we know about the most widely cited cross-modal effects, then explores an expanded range of effects relevant to multisensory speech perception, and finally endorses a broadening of the scope of research in cross-modal effects.

Until fairly recently, terms such as “cross-modal” and “multisensory” have been almost exclusively limited to two sensory modalities in the speech perception literature – audition and vision – with audition generally seen as the primary or dominant modality; that is, when studies of speech perception have discussed cross-modal effects, they have generally described influences of specific visual speech information on auditory processing of speech. For example, in their seminal work on audiovisual speech processing, Sumby & Pollack (1954) found that visual speech information can enhance accuracy of auditory speech perception – especially when audio signals are highly degraded. Their groundbreaking work led to a conclusion that now seems obvious: speech has visual characteristics in addition to auditory characteristics. As Bateson and colleagues framed it, “The motor planning and execution associated with producing speech necessarily generates visual information as a by-product.” (Vatikiotis-Bateson et al. 1996, p. 221). Perhaps more than a by-product, the visual information produced during speech may be integrated fundamentally into our representations of speech, as evidenced by the observation that blind and sighted speakers use different lip movements to produce the same speech sounds (Ménard et al. 2016).

Sumby & Pollack’s (1954) enhancement work, while groundbreaking, did not investigate what aspects of the visual signal were useful to perceivers and why, nor did they speculate about what the mechanism might be that allows perceivers to use this information. Nonetheless, their work opened the door to an important shift in thought: researchers could no longer consider audition to be the sole modality for the transmission and processing of speech. Twenty years later, McGurk & McDonald (1976) published their widely known finding that perceivers given incongruent AV stimuli will often choose neither the response consistent with the auditory stimulus nor that which is consistent with the visual stimulus. As we will discuss further below, the decades of scholarly energy, particularly among linguists and psychologists, following McGurk & McDonald (1976) that went into understanding the McGurk effect and AV speech more generally brought us great advances in understanding speech from at least a bimodal perspective; at the same time, however, this focus on AV speech diverted attention away from the development of theoretical models that cast speech in a more broadly multisensory space.

While researchers in linguistics and psychology were focused on AV speech, clinical researchers were also investigating cross-modal effects on perception in the form of using tactile speech information as a means of enhancing speech perception for impaired populations (e.g., Alcorn 1932; Sparks et al. 1978). The most well-known example is the Tadoma method (Alcorn 1932; Vivian 1966; Reed et al. 1978), a communication method most commonly used with deaf and blind individuals who lost their hearing and sight early in childhood (around 18 months of age). In Tadoma, the perceiver places her hand on the face of the interlocutor in such a way that he or she can feel much of the interlocutor’s articulatory movements and their sensory consequences (though see Reed et al. 1989). By

placing the thumb at the lips and fanning the fingers across the check and neck, the perceiver can feel the movements of the lips and jaw, vibrations at the neck, and airflow at the lips and nose. The success of deaf-blind individuals in using Tadoma to communicate certainly suggests that the tactile modality is a viable communicative sense. Since most of this work has come from a clinical perspective, the focus has naturally been on the use of tactile information as a means of communication for clinical populations.

Although there has been a great deal of research into how the resulting tactile information can teach impaired perceivers to communicate, there has been less of an attempt to consider how this tactile information – and the fact that it can be used at all – fits into conceptions of how speech perception works for a general population (i.e., outside of purely clinical applications). For example, there is little research in the clinical literature asking what the Tadoma method might tell us about how normally hearing and seeing populations make use of the tactile modality, or what this might tell us about cognition and communication more broadly. Instead, the assumption has persisted that tactile information can be recruited to support the auditory and visual streams with training if the other streams are unavailable or degraded. This in turn has the effect of reinforcing the assumption that tactile cues require a great deal of training in order to be effective (see, e.g., Bernstein et al. 1991). However, as we will discuss further in later sections, more recent psychological and linguistic research has shown that the tactile modality can have a modulatory effect on speech perception for even untrained perceivers, suggesting that speech perceivers use signal information from whatever sources they have available, whether audio, visual, or touch.

In the following sections, we first review some of the many important findings in audiovisual speech perception. We will discuss how these findings have shaped our understanding of the what the perceptual system detects and makes use of during AV speech and when this integration occurs. In Section 3, we will focus on identifying the other modalities relevant in speech perception as well as delineating the role of production. Then we will describe how broadening our research to include these additional modalities has enriched our understanding of multi-modal speech perception. Finally, we will offer some conclusions on how this additional knowledge bears out with respect to the traditional dichotomous view and where the field should go from here.

2. Audiovisual Speech

2.1 The McGurk Illusion: Fusion or Confusion?

Discussions of cross-modal effects in speech perception often begin with the McGurk effect (McGurk & MacDonald 1976), perhaps the most widely known and most studied cross-modal effect on speech perception. In the original study, the authors presented participants with incongruent auditory and visual speech stimuli (auditory *ba* dubbed over a visual *ga*). The participants were then asked to indicate what they heard. Their finding was that participants often responded with neither the syllable that matched the visual token nor with the syllable that matched the auditory token; rather, participants were significantly more likely to respond that the speaker had said *da*, a sequence not present in either modality. The authors' findings were interesting in part because they showed that visual information can influence auditory perception even when the acoustic signal is clear and not degraded. In

some ways, the McGurk illusion seems largely automatic, suggesting that perceivers cannot avoid integrating the visual information. For example, perceivers report hearing *da* even when explicitly told about the dubbing and instructed to attend to the auditory cue, and after receiving training at attending to the auditory cue (Massaro 1987). Moreover, perceivers appear to be surprisingly insensitive to such factors as gender congruence between face and voice (Green et al. 1991), degradation in the visual signal (Rosenblum & Saldaña 1996), and temporal asynchrony between audio and video signals (though the direction of the asynchrony appears to matter, as will be discussed below).

While the McGurk illusion has been observed in many studies, there is also considerable evidence that it is not nearly as robust as is often suggested and may even disappear in certain contexts. For example, although it is often noted that the McGurk effect persists when there are incongruities between aspects of the voice producing the auditory stimulus and the face producing the visual stimulus (Green et al. 1991), there are circumstances when an apparent mismatch in the source affects the McGurk effect. Walker et al. (1995) found that the illusion is greatly reduced if familiar faces and voices are used. In addition, the McGurk effect is sensitive to differences across vowel context (Green et al. 1988) such that while /Ci/ contexts elicit strong McGurk effects, /Ca/ is variable and /Cu/ is quite weak. Further, both coarticulatory cues to the following vowel (Green & Gerdeman 1995) and high attention loads can disrupt the McGurk effect (Alsius et al 2005; Alsius et al. 2007; Tiippana et al. 2014). Finally, the effect is highly subject-dependent (e.g., Basu Mallick et al. 2015), a fact that was confirmed by a reanalysis of a large corpus of McGurk data (Schwartz 2010).

Perhaps the larger issue, however, is that a key conclusion drawn from this effect is fallacious: namely the idea that true audiovisual integration *equals* a fused percept. The idea of fusion stems from the fact that the reported percept matches neither the auditory percept nor the visual one. Proponents argue that the novel percept emerges because the perceiver "fuses" place features from the auditory bilabial [ba] and the visual velar [ga] to arrive at the alveolar [da] (van Wassenhove 2013) a view that is problematic on several levels. First, it implies that [d] is somehow an intermediate segment between bilabial [b] and velar [g]. This further suggests that these stops exist on a continuum and the perceiver is simply selecting the midpoint between the articulatory features of the incongruent stimuli (e.g., de Gelder et al. 1996). While it is indeed true that the alveolar ridge is positioned between the lips and the soft palate, the muscles and structures involved in these articulations are so different that the idea of alveolar as a kind of bilabial-velar compromise is unlikely. This take on the fusion illusion also assumes that perceivers are correctly identifying the visual stimulus as /ga/, an assumption that is impossible to confirm. More to the point, as noted in Tippana (2014), the concept of fusion as the true or ideal example of integration ignores the other percepts reported in McGurk-style tasks.

Beyond the above issues with fusion, the idea that only a fused percept indicates true audiovisual integration has in some ways slowed movement toward understanding the more richly multisensory nature of speech. Audiovisual integration happens even when the McGurk effect does not (Brancazio et al. 2002). Once we move past the idea that fusion equals cross-modal effects on perception, the landscape of cross-modal speech effects starts to make a bit more sense. Instead, the usual integration response is that congruous stimuli

improves perception because one sense provides information missing from the signal given to the other sense. In contrast, incongruous stimuli generates confusion as the signals provide conflicting information that does not match what normally occurs in real-world speech. That is, processing incongruous cross-modal stimuli (as in studies of the McGurk effect) is, at least in some respects, a fundamentally different kind of task from processing congruous cross-modal stimuli; this is supported by neuroimaging evidence that the brain recruits additional cortical areas when processing incongruent AV speech (Erickson et al. 2014).

2.2 Beyond McGurk

Beyond the well-known McGurk effect, as described above, movements of the vocal tract involved in the production of speech sounds have well-known visual and auditory consequences (Munhall et al. 2004; Munhall & Vatikiotis-Bateson 2004; Vatikiotis-Bateson et al. 2000; Vatikiotis-Bateson et al. 1996). One of the most interesting discoveries regarding cross-modal AV effects is that these visual speech cues do not just come from observing the oral aperture; rather, perceivers can make use of phonetic information from kinematic movements of essentially all parts of the face: the lips, jaw, neck, cheeks, eyebrows - and even movements of the whole head (Vatikiotis-Bateson et al. 1996; Yehia et al. 1998); these movements have been shown to convey different types of speech information, and they participate in cross-modal audiovisual effects that operate under a variety of experimental conditions. For example, while we have considerable evidence that visual speech information can supplement a degraded or ambiguous acoustic signal, we also have evidence that visual speech information modulates auditory perception even when the auditory signal is clear and unmanipulated (e.g., Arnold & Hill 2001; Reisberg et al. 1987).

Some research in AV perception suggests further that auditory and visual speech cues do not combine in a simple, additive fashion, but that they are, rather, superadditive. Superadditivity refers to the observation that perceivers' responses to multisensory information are not merely the result of a summation of the unimodal parts; rather, perceivers gain a disproportionate benefit from the addition of cross-modal information. Behavioral studies have shown evidence of superadditivity in audiovisual speech (e.g., McGrath & Summerfield 1985) showing that when perceivers experience audiovisual speech, their responses are more accurate than expected based on the sum of their responses to visual cues and auditory cues in unimodal conditions. While this seems to be true of behavioral results, electrophysiological evidence suggests that audiovisual speech perception may be subadditive. Klucharev et al. (2003) reported that when ERP responses to audiovisual stimuli are compared to the sum of unimodal auditory and visual speech stimuli responses, the AV responses were found to be smaller than the sum of the unimodal responses. While these neural findings may at first appear paradoxical, they may be related to decreased ambiguity in the signal. Studies of ambiguous stimuli indicate that, when exposed to congruent cross-modal cues, the brain does not have to work as hard to process the signal as it would if the input were unimodal (Parker & Krug 2003). A recent meta-analysis drawing on the results of hundreds of studies of audiovisual speech (Baart 2016) reports that audiovisual speech reduces N1/P2 peaks compared to audio alone, confirming that visual speech production helps with the processing of auditory information. Similar decreases have been found where

lower ERP amplitudes were observed during the easier task of decoding normal speech compared to disordered speech (Theys & McAuliffe 2014a, 2014b).

Some of the most revealing cross-modal effects to date in AV speech have been seen in observing how brain activation responds to speech, both cross-modal and unimodal. Some of these findings challenge the view that cross-modal perception is dependent on higher order processing that integrates information processed in unimodal streams. A particularly interesting finding concerns the unimodal perception of visual speech cues during silent speech. Multiple studies have found that silent visual speech information generates activation in the areas of the brain primarily associated with auditory processing (Campbell et al. 2001; Calvert et al. 1997; MacSweeney et al. 2000). For example, Calvert et al. (1997) used fMRI to test the brain activation of normal hearing subjects during a variety of conditions: silent lip reading, heard speech, nonspeech lip/jaw movements, and speech-like lip and jaw movements. The authors reported activation in the primary auditory cortex not only in the condition where participants heard acoustic speech cues, but also during silent lipreading trials and trials with phonetically-plausible lip and jaw movements. Crucially, this activation did not occur during trials in which the participants saw non-speech lip and jaw movements. The authors interpret these findings as indicating that visual information influences the perception of auditory cues long before they are categorized into phonemic categories.

Similarly, incongruent visual-only speech information modulates responses in the auditory cortex. Sams et al. (1991) used MEG and a McGurk-style task to investigate in which part of the brain the visual information affects auditory processing. The authors note that classical theories assume that the visual information is processed in the occipital cortex before being sent to the angular gyrus for “reorganization into auditory form.” However, they did not find coherent activity in these two regions. Instead, they found responses in the primary auditory cortex. This is further evidence that visually presented articulatory movements modulate responses in a cortical area typically associated with auditory processing. It also complicates a model in which cross-modal stimuli are processed in unimodal streams and then integrated at later stages, a view further supported by evidence that auditory and visual speech information interact in lower-order structures such as the brainstem (Musacchia et al. 2006).

2.3 Spatiotemporal Congruence

While there has been some debate within the speech perception literature as to the degree to which speech is “special” and not governed by the same requirements as non-speech, work on AV speech has provided evidence that cross-modal effects are, at least in part, constrained by general properties of the natural world. Many researchers interested in multi-modal perception outside the domain of speech point to the importance that coincidence in space and time plays in governing perceptual integration (e.g., Holmes & Spence 2006; Macaluso & Driver 2005). Indeed, it appears that relative timing plays an important role in integrating cross-modal speech cues. For example, multiple studies have shown that there exists a temporal window within which integration occurs (Munhall et al. 1996; van Wassenhove et al. 2007). In other words, the onsets of the cross-modal cues must co-occur within a specific window of time in order to be integrated. This window is surprisingly large, suggesting that

perceivers are relatively forgiving when determining which incoming information relates to a perceptual event. In addition, the temporal window has been shown to be asymmetrical. Thus, in Munhall et al. (1996), participants showed a significant decline in integration when the visual speech cue (in the form of a video) preceded the auditory cue by more than 180ms, while in contrast, participants were much less forgiving when the auditory cues preceded the visual: the decline in integration occurred when the cues were offset by just 60ms. Munhall et al. (1996) suggest that this asymmetry can be explained by facts about how the cues behave in the natural world (i.e., that the speed of light is faster than the speed of sound). Perceivers may thus be sensitive to the natural relationship between the relative speeds of auditory and visual signals, and are more likely to integrate cross-modal cues that agree with basic principles of physics.

The story becomes murkier, however, when spatial congruence is considered. If perceivers are picking up a localized, distal source during speech perception (e.g., Fowler 1986), spatial incongruence of cross-modal cues may interfere with integration: stimuli coming from different directions are unlikely to originate at the same source. However, though the “spatial rule” appears to hold in multi-modal perception outside of speech (Soto-Faraco et al. 2003), it seems to have little to no effect on AV speech perception (Bertelson et al. 1994; Fisher & Pylyshyn 1994; Jones & Munhall 1997). Studies have shown little evidence that perceivers care about whether cross-modal cues appear to originate from the same spatial location. Indeed, it is this lack of constraint on spatial congruence that enables ventriloquism, surely the most popularly known of cross-modal speech illusions. Though some studies have tested very small degrees of dislocation (Bertelson et al. 1994; Fisher & Pylyshyn 1994), Jones & Munhall (1997) reported that participants showed a strong McGurk effect even when the auditory and visual stimulus were separated by as much as 90 degrees. Though these findings are somewhat expected given the robustness of the ventriloquism effect, it is surprising that perceivers might rely on synchrony alone. However, it remains possible that their findings are specific to audiovisual processing and not a fact about cross-modal speech perception more generally. Humans have a strong bias toward the visual stimulus when it comes to auditory localization (Bertelson & Aschersleben 1998) and that bias might override spatial information from other cues.

While studies of interactions in AV speech have provided many useful insights, they have left many questions unanswered. The broadest of these questions is whether observations about AV speech are specific to the audiovisual pairing, or whether these observations indicate more general properties that would obtain across any cross-modal (or multi-modal) pairing. Getting at this question requires moving beyond AV speech to compare additional modality pairings. The remainder of this paper focuses on studies of cross-modal effects that have attempted to extend the range of modalities.

3. Somatosensory Speech

3.1 The Many Somatosenses

Somatosensory space has been identified for decades as a rich yet largely unexplored frontier for speech research (e.g., Abbs & Gracco 1984; Gick et al. 2008; Ghosh et al. 2010; Kelso et al. 1984; Nasir & Ostry 2006; Perkell 2012; Tremblay et al. 2003). While the work

in this area has been vital in establishing somatosensory information as playing an important role in speech, the various effects referred to in these studies are often attributed to a single modality under the label “somatosensory”. Although the somatosenses have often been described as a single sense modality in the speech literature, the term “somatosense” applies to a broad range of different sensory modalities (Hsiao & Gomez-Ramirez 2011). As Hsiao & Gomez-Ramirez (2011) put it:

> “The somatosensory system is best conceptualized as a multi-modal, rather than a unimodal, processor, comprised of multiple parallel systems carrying information about numerous aspects of environmental stimuli. To the extent that a given object encountered in the real world may simultaneously generate multiple tactile impressions, what is remarkable is that the somatosensory system unites these disparate channels into a unified percept (141).”

The “disparate channels” described here – the various somatosenses – include senses such as pain, itch, temperature, pressure, vibration, and joint position, among others (Hsiao & Gomez-Ramirez 2011; Wilson et al. 2009), each with its own distinct organs, mechanisms and neural processes, and with some researchers further dividing the numerous somatosenses into subgroups (e.g., tactile, proprioceptive, and vestibular; Anderson and Fairgrieve, 1996). Thus, while perturbation studies have often examined the role of proprioception in speech (e.g., Kelso et al. 1984; Nasir & Ostry 2006), other studies have examined mainly pressure sense (e.g., Ghosh et al. 2010), others the vibrotactile sense (e.g., Bernstein et al., 1991), and still others the aerotactile sense (e.g., Gick & Derrick 2009, Derrick & Gick 2013). The various somatosenses should thus be viewed as being approximately as distinct from one another as they are from those modalities that have been more traditionally recognized in the speech literature (i.e., vision and audition) – and as such, we suggest that the same kinds of approaches traditionally used to investigate cross-modal processes in speech ought to be brought to bear on each of the different somatosenses. Minimally, in addition to the senses of audition and vision, this should apply to proprioception, pressure, vibration and aerotactile sensation, all of which should be considered highly relevant modalities for experiencing and conveying speech, and each of which merits study both separately and in combination with other modalities in speech. Far from taking place in a one-, two- or even three-dimensional space, speech should thus be viewed as occupying a sensory space that is indeed richly multidimensional.

Many of the above behavioral experiments relating to somatosensory input are focused on speech production rather than perception. Production has at times been left out of discussions of cross-modal effects on speech, or at times treated as an additional modality or sense. However, while there is no “production” sense, the movements of speech production do generate much of the somatosensory feedback shown to modulate speech perception and are thus important in any discussion of cross-modal effects. When we speak, we receive continuous sensory input through multiple modalities: the sense of air flowing across the articulators, the vibration of the vocal folds in the neck and other structures, pressure and proprioception from facial skin deformation, and of course, acoustic-auditory feedback, among others. Though we may not be consciously aware of this sensory feedback, it has been shown that manipulating the somatosensory feedback from our articulators during

speech production causes speakers to alter their production. For example, when the lips (Abbs & Gracco 1984) or jaw (Kelso et al. 1984) are mechanically perturbed, speakers adapt to compensate for the perturbation. This holds even when the perturbation does not alter the acoustic signal, suggesting that somatosensory targets themselves are relevant in the production of speech sounds independent of auditory goals (Tremblay et al. 2003; Nasir & Ostry, 2006).

This cross-modal interaction has been shown to feed into speakers' productions more generally. Somatosensory acuity has been shown to correlate with produced acoustic contrast distance between /s/ and /ʃ/ (Ghosh et al. 2010). This effect is independent of auditory acuity, which also suggests that speech sounds have independent somatosensory and auditory perceptual goals. It is also the case that inputs generated by production can influence and be influenced by auditory perception. In a novel "skin-stretching" paradigm, Ito et al. (2009) simulate the somatosensory consequences of lip spreading; they find that stretching the facial tissue in such a way that mimics skin deformation in vowel production influences the perception of the vowel sound heard by participants. Strikingly, the reverse also holds: hearing a vowel can shift the perceived direction of skin stretch (Ito & Ostry 2012). These cross-modal effects hold in spite of evidence that the lips do not contain muscle spindle proprioceptors to provide input regarding changes in muscle length (Frayne et al. 2016).

When our perception in one modality is affected by production, it is unclear whether we are responding to the sensations themselves or to the sensory consequences our internal models simulate, known as "efference copy" or "corollary discharge". When a motor command is sent to the motor system to initiate an action, a copy of this command signal, known as the efference copy, is sent to an internal forward model, which generates a prediction of the perceptual or sensory consequences of this act (Pickering & Garrod 2013). This prediction is a sensory signal known as corollary discharge (Scott 2012). Experimental investigations of corollary discharge in speech production have focused on the auditory consequences of speech production. For example, Tian & Poeppel (2010) showed that the activity of the auditory cortex during articulatory imagery tasks, where participants imagined producing certain sounds, is strikingly similar to its activity during the perception of actual auditory stimuli; this was argued to be a result of predicted auditory feedback (corollary discharge). It has further been argued that the corollary discharge involved in auditory imagery constitutes a sufficiently rich and detailed representation of the predicted sensory information that it can interfere with the perception of actual external auditory signals (Scott 2012). Our immediate responses to sensory feedback during everyday speech production are thus likely responses to corollary discharge rather than to external sources, opening possibilities for novel experimental approaches to understanding cross-modal effects involving speech production.

3.2 The Tadoma Method: A Multimodal Method

As mentioned above, we have long known from clinical research on the Tadoma method that somatosensory information can, with adequate training, provide a useful aid to communication. As the Tadoma method was initially developed for the those who suffer from loss of hearing and vision, it has often been used for populations who no longer have

access to speech information from hearing and sight, but who did have access to the visual and auditory cues at one time (Alcorn 1932; Vivian 1966). From a clinical point of view, research has thus focused more on using somatosensory cues as a substitution for missing cross-modal information rather than on the role of somatosensory input as a natural part of everyday speech perception. It has nevertheless been observed that individuals who lost their hearing and sight as young as 1.5 years old can become successful communicators with the Tadoma method, with the interpretation that the Tadoma input does not just provide access to the speech stream, but serves to “create a language base” for individuals whose language acquisition was interrupted at a fairly early stage (Reed 1996). Further to this point, experimental work shows that sensorimotor information from the articulators is both available and influential during speech perception for infants as young as six months old (Yeung & Werker 2009; Bruderer et al. 2016), suggesting that the somatosensory information about articulatory movement is already available to these individuals in some capacity by the time they lose their hearing and sight.

As noted above, the focus of the clinical research on the Tadoma method did not generally extend to discussions of speech perception in a non-clinical population. However, it did pave the way for later studies of cross-modal tactile effects in speech, starting with Fowler & Dekle (1991), an important early work in opening up cross-modal speech perception research to modality pairings beyond audition and vision. Perhaps just as importantly, the authors showed that somatosensory speech information is available to even untrained perceivers. In their study, the authors contrasted the influence on auditory perception of somatosensory speech information vs. orthographic cues in a McGurk-like task, asking whether the McGurk effect arises because of cue association in memory or because the cross-modal cues jointly specify the same event in the real world. The authors compared participant responses in two conditions with conflicting cross-modal cues. In one, auditory cues were simultaneously presented with either congruent or incongruent mouth movements. Participants placed a hand over the lips of a speaker and were asked to identify which syllable they had heard, as well as which they had felt. In the second condition, participants saw a congruent or incongruent printed syllable on a computer screen at the same time as they heard a syllable. As in the previous condition, participants responded with which syllable they had heard followed by which they had seen. The authors chose these two situations because in the first, the cross-modal cues jointly specify the real-world event and thus have a causal relationship in the natural word. However, most perceivers have little or no experience feeling the mouth movements of a speaker while listening. In contrast, the second condition offered a situation in which the cross-modal cues are associated only by social convention, yet the undergraduates tested in the study have experience with the visual and acoustic pairing of spelling and sounds. The authors found that the haptic information, and not the orthographic, influenced categorization, lending support to the idea that haptic speech information can be useful to perceivers without hours of training.

Gick et al. (2008) picked up this line of research, using the Tadoma method to further show that somatosensory information not only influences perception of incongruent syllables but that it can increase accuracy of congruent audio-tactile (AT) and visuo-tactile (VT) speech in a syllable identification task. The authors tested a group of perceivers with no previous training in the Tadoma method on their ability to identify syllables through bimodal pairings

of auditory, visual, and tactile speech cues. Accuracy improved by nearly 10% when tactile information was available to perceivers when paired with auditory-only or visual-only speech information. Their findings support the idea that perceivers do not need to have previous experience with information in those specific modality pairings for the cues enhance perception. This study also highlighted how using multiple modality pairings can help in understanding the role of superadditivity in speech perception. The degree to which the Tadoma information augmented auditory or visual speech perception varied considerably between individuals, suggesting that perceivers may use information from one modality more than another, in line with the AV findings above from Schwartz (2010) that some perceivers are more auditory and others are more visual. Gick et al. (2008) further found that participants whose accuracy increased with the addition of tactile cues in one modality pairing (e.g., AT) tended to benefit less from adding the same cues in the other modality pairing (e.g., VT).

While the Tadoma method has thus played an important role in the development of research into the relevance of somatosensation in speech, its history has been shaped by perhaps its most important characteristic, i.e., that Tadoma is not a unimodal method, but a (highly) multimodal one. That is, if we consider somatosensory space as comprising many distinct modalities, this gives a much more complex view of the Tadoma method than merely adding unimodal (“tactile”) information. On the contrary, the hand position for the Tadoma method is designed to allow the perceiver to pick up, at the very least, vibration from the larynx, air flow from the lips and pressure sensation from the moving articulators. Understanding the multisensory nature of the Tadoma method helps to explain both its power as a communicative tool for clinical populations as well as its challenges as an experimental tool for studying speech perception in a normative population. Naturally, interpreting the Tadoma method as if it conveyed a single sense modality rather than three or more would create confusion in any experimental paradigm, and attempting to create a Tadoma experiment with a single modality as an independent variable would seem certain to fail. Considering this likely confound, while some of the more descriptive aspects of a Tadoma study such as Gick et al. (2008) may remain useful, their results regarding superadditivity should be considered provisional, meriting future study using more easily controlled methods than Tadoma.

3.3 Aerotactile Speech Perception

Following a plosive sound such as [pa], air exits the mouth relatively slowly, traveling at a velocity an order of magnitude slower than the speed of sound, with the resulting pressure front dispersing and slowing loglinearly as it advances (Derrick et al. 2009). When this slow-moving pressure front strikes skin, it stimulates mechanoreceptors in the skin and in hair follicles that signal the presence of air flow across the skin, creating an aerotactile sensation (Gick & Derrick 2009); the absence of hair reduces the perceptibility of this sensation (Derrick & Gick 2013). The aerotactile sense is one that presents itself as an excellent candidate for experimental study of cross-modal speech effects. As with audible and visual speech information, the aerotactile signal (air flow) is transmitted externally, it is isolable (i.e., independent of other information such as vibration and pressure), and it is relatively easy to perturb and simulate precisely. Further, air flow can be applied to the skin from any direction and at any body location where hair is present, and is safe enough to be applied to

infants, opening doors for a range of novel kinds of experiments. For these reasons, the aerotactile sense has become the source of a number of novel cross-modal effects.

While it may seem difficult to imagine that the sensation of air flow on the skin could play a significant role in speech perception, it becomes less so when we remember that much spoken communication happens within the range of distances where air flow can be felt, particularly during language acquisition, and that, as we talk, we simultaneously perceive the sensations created by our own production, e.g., on our hands. Studies of aerotactile speech perception over the last decade have demonstrated that untrained perceivers incorporate aerotactile somatosensation as an informative part of the speech event. Gick & Derrick (2009) showed that aerotactile information can enhance or interfere with accurate auditory speech perception in a two-way forced choice task, even for participants who report being unaware of the airflow. When stop-initial syllables were accompanied by silent puffs of air applied to the participants' neck and hand, participants were significantly more likely to perceive the syllable as aspirated (i.e., /pa/ or /ta/), even in mismatch conditions (i.e., when the air puff occurred with an unaspirated token). In a control study, Gick & Derrick (2009, supplemental materials) compared the effect of applying a different somatosensory stimulus – a tap on the hand – in an otherwise identical experiment; not surprisingly, they observed no effect on auditory perception, highlighting the distinct effects of information conveyed through different somatosenses. The aerotactile effect on auditory perception has been replicated (Gick et al. 2010; Derrick & Gick 2013) and extended to enhancement of fricative identification (Derrick et al. 2014). Additionally, air flow applied to the skin has also been used to shift speech perception using tokens along a voicing continuum (Goldenberg et al. 2015), showing that the effect of aerotactile input on audible speech is stronger when auditory cues are more ambiguous.

While most aerotactile cross-modal studies have focused on effects on auditory perception, aerotactile cues have also been shown to apply even in the absence of audition, affecting the perception of visual-only speech. For example, Bicevskis et al. (2016) found that aerotactile information influences the perception of silently articulated bilabial stops. Participants were presented with silent videos of a speaker articulating /ba/ or /pa/. During some of the trials, they felt silent puffs of air on their skin. Just as in the above audio-aerotactile findings, the sensation of air-flow on the skin significantly affected the participants categorization such that participants were significantly more likely report that the speaker produced the voiceless aspirated token /pa/ when the silent video was accompanied by air-flow on the skin. Similarly, C. Chang, M. Keough, M. H. Schellenberg, & B. Gick (in prep) tested congenitally hard of hearing perceivers in a visual-aerotactile paradigm following the methods of Bicevskis et al. (2016), with results showing no effect of hearing loss, i.e., the hard of hearing perceivers were significantly more likely to respond /pa/ when they felt airflow, just like the control group with normal hearing. These results indicate that, rather than using the aerotactile information only in support of an ostensibly primary auditory signal, perceivers simply make use of whatever speech information is available, much as when the Tadoma method is used by deaf-blind perceivers to access speech information through multiple somatosenses. Based on this lack of an exclusive primary modality for speech, Bicevskis et al. (2016) describe speech as being "modality neutral".

An important issue raised by the study on congenitally hard of hearing perceivers (C. Chang, M. Keough, M. H. Schellenberg, & B. Gick, in prep) concerns the role of experience in cross-modal interactions. Understanding the role of experience in establishing links between different sensory streams could provide a key to determining whether these cross-modal links are learned or innate in humans. Although we can certainly learn from the experience of feeling our own air flow on our skin when we speak, it appears that the ability to integrate aerotactile cues during speech perception is not contingent on the perceiver having had prior experience with producing aspiration. The congenitally hard of hearing perceivers showed cross-modal effects of air flow on speech reading despite their not having contrastive aspiration in their own productions (C. Chang, M. Keough, M. H. Schellenberg, & B. Gick, in prep). Similarly, preverbal infants can be influenced by airflow cues in much the same way as normally hearing adults; when infants were presented with unaspirated /ba/ tokens accompanied by a puff of air on the skin, they perceived them to be more like an aspirated /pa/ (M. Keough, P. Kandhadai, H. H. Yeung, J. F. Werker, & B. Gick, in prep). These findings suggest that perceivers do not require experience feeling their own airflow during productions of aspirated and unaspirated tokens in order to integrate auditory and aerotactile information. This is in keeping with evidence that infants integrate other multi-sensory speech cues well before they begin speaking (e.g., Rosenblum et al. 1997).

3.4 Spatiotemporal (In-)Congruence and Ecological Validity

As discussed above, previous studies of audiovisual cross-modal effects have revealed that AV speech has some notable spatiotemporal properties. First, temporal congruence of signals is important for integration within an asymmetrical window in a direction consistent with the laws of physics (e.g., Munhall et al. 1996). Second, spatial congruence does not seem to be required for audiovisual integration in speech (e.g., Jones & Munhall 1997). While a certain degree of temporal congruence has also been shown to be important, perceivers do not integrate just any synchronous set of cross-modal cues. For example, Fowler & Dekle (1991) did not find evidence that simultaneously presented orthographic visual cues influenced auditory speech perception. Likewise, as described above, while Gick & Derrick (2009) found a significant effect of a light puff of air on the hand synchronous with auditorily presented plosives, they observed no effect of a light tap on the hand. Both of these results suggest that cues must be more than merely synchronous: rather, cross-modal stimuli must have a lawful, causal relationship in the real world in order to be perceived as integrated. These findings are corroborated by work showing that neural responses to auditory-tactile stimulation are modulated by congruence between the area of the body touched and the bodily origin of the acoustic signal (Shen et al. 2018). For two stimuli to exhibit a real-world causal relationship, however, suggests shared properties along multiple possible axes, e.g., temporal congruence, spatial congruence and signal relevance.

In order to test temporal congruence, Gick et al. (2010) tested perception of synchronous and asynchronous presentations of audible /pa/ and /ba/ plosives in combination with slight, inaudible, cutaneous air puffs to the skin. As with the previous AV findings, results of this audio-aerotactile study showed an asymmetrical window of integration for the enhancement effect, allowing up to 200 ms of asynchrony when the puff followed the audio signal, but only up to 50 ms when the puff preceded the audio signal. Interestingly, however, the same

asymmetrical window did not occur for the interference effect, consistent with the differential neural processing of congruent and incongruent cross-modal stimuli (Erickson et al. 2014). Bicevskis et al. (2016) observe a similar asymmetrical window in their visual-aerotactile study. It is notable that these asymmetries in both modality pairings occur in the direction that would be expected based on relative signal speed, supporting the view that the perceptual system is built (whether innately or through experience) to accommodate natural differences in physical transmission speed of multimodal signals.

Given perceivers' constraints on timing, one might expect perceivers to be similarly sensitive to spatial dislocation. The audio-aerotactile pairing is particularly useful here because both speech cues – the audio signal and the air puff – can be presented laterally, allowing for 180-degree dislocation of the crossmodal cues. Keough et al. (2016) describe a study in which participants took part in a two-alternative forced choice task in which they were presented with /ba/ and /pa/ syllables in noise. The syllables were presented one of three ways: in the left ear only, in the right ear only, or binaurally. Remarkably, participants showed the same levels of integration regardless of whether the auditory and tactile cues were presented from the same or opposite directions. These results add to the findings in AV speech perception discussed above as well as to audio-tactile studies with non-speech stimuli; for example, Sperdin et al. (2010) showed that participants are unable to pinpoint the origin of a unimodal auditory or tactile stimulus after being presented with a multimodal trial, even when informed that spatial location will be task-relevant. Another way it is possible to observe spatial congruence through tactile stimuli is through body location rather than signal direction. Indeed, one of the more intriguing outcomes of tactile work in speech is the degree to which speech perception is holistic in body space. In particular, perceivers integrate aerotactile speech information whether felt proximally, at the neck, or distally, at the hand (Gick & Derrick 2009) – or even at the ankle (Derrick & Gick 2013). It is difficult to imagine that perceivers have significant prior experience feeling another speaker's airflow on their ankles. Rather, these results suggest that the whole body participates in the perception of speech-related cues – another respect in which speech perception is unconstrained with regard to spatial congruence.

The foregoing studies on audio- and visual-aerotactile perception indicate that perceivers will integrate air-flow cues without the presence of a plausible real-world source, at least from a spatial point of view. However, based on the findings above, we know that signal relevance does seem to be important. Further to this, perceivers do seem to distinguish between possible though spatially dislocated sources and impossible ones; Keough et al. (2017) extended the methods of Bicevskis et al. (2016) using a video of an computer-generated face rather than a real face. They found that perceivers do not integrate visual and aerotactile information from a simulated face in the same manner as they do with a real human face. Instead of participant response being affected by the condition (i.e., air-flow or no air-flow), all perceivers' responses shifted categorically from /ba/ to /pa/ over the course of the experiment. These results suggest that the participants learned over the course of the experiment to associate the articulation in the video with /pa/ as they learned to associate more trials with air puffs. This response pattern was markedly different from that seen in Bicevskis et al. (2016) and aligns with other evidence indicating that perceivers are sensitive to the ecological validity of the source.

4. Conclusion

Historically, there has been an opposition in ways of looking at how and why we integrate cross-modal cues (e.g., Massaro 1987, 1998; Massaro & Chen 2008; Fowler 1986, 1996, 2010), a dichotomy has continued to influence assumptions about cross-modal effects and their interpretation. On one side is a Helmholtzian-based cognitivist approach in which perception is mediated by some higher-order cognitive process (e.g., Massaro 1987; Schwartz 2010), where the perceptual system detects and processes sensory cues from each modality separately; these cues are subsequently integrated into a percept through top-down processing. In one such model, Massaro's fuzzy logic model of speech perception (FLMP) theory (Massaro 1987, 1998), perceivers *evaluate* the auditory and visual speech signals, *integrate* them by consulting memory representations derived from commonly experienced information in the real world to *decide* upon the most appropriate percept. On the other side of the theoretical fence we see a "direct" or ecological approach to speech perception, such as those of Fowler (1986, 1991, 1995) and Bregman (1990), based on Gibson (1966, 1979/1986). In this view, perception does not involve the detection of sensory input at the receptors and top-down interpretation. Instead, proponents argue that because perception is the means through which perceivers evolved to interact with their environment, they directly experience the causal source of the sensations.

The present review includes work from both sides of this divide, as well as more recent work, our own included, which eschews the dichotomy. We see these two approaches as two different views of the same landscape, where one is perhaps more focused on the mechanics and the other on the end effect, but where, by and large, the two views have largely converged during recent decades. The individual in the real world receives different information about that world via the different senses regardless of what they expect. Most of these senses have been understood to be relevant only recently, and not all of them are relevant to speech at any given time. But as we have argued in this review, sensory information does not have to be frequently experienced in order to contribute, and it can derive from unexpected sources – as long as the cues relate to a real-world event. As laboratory experiments become more complex and sophisticated, and as researchers consider an ever richer array of relevant senses and their cross-modal effects, bottom-up cognitivist approaches appear increasingly ecological while top-down ecological approaches must become increasingly mechanistic. What emerges from the evidence of these additional modalities is that, while cross-modal effects in speech perception are constrained by temporal congruence and by signal relevance, they appear to be unconstrained by spatial congruence (either in terms of direction or location). In short, if two signals are roughly synchronous and are of an appropriate type to fit with a common source, we are likely to try to assign them to a single environmental cause.

Research into the tactile dimensions of speech perception is still in its infancy. However, the potential in this field of research is vast; hopefully, it will one day be as widespread as research into audiovisual speech perception is today. This research has already led to innovation in nasalance, turbulent speech airflow recording (Derrick et al. 2014; Derrick et al. 2015) and artificial air flow production techniques (Derrick & De Rybel 2015), and has helped uncover answers to questions about how the skin responds to speech air flow

compared to other types of air flow. But most importantly, it will usher in true multisensory speech perception research. In this way, continued research into cross-modal effects in speech perception will bring a deeper understanding of how humans sense and interact with the world.

Acknowledgements

This research was funded by NIH Grant DC-02717 to Haskins Laboratories.

Literature Cited

Abbs JH, Gracco VL, 1984. Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech. Journal of neurophysiology 51, 705–723. [PubMed: 6716120]

Alcorn S, 1932. The tadoma method. Volta Review 34, 195–198.

Alsius A, Navarra J, Campbell R, Soto-Faraco S, 2005. Audiovisual Integration of Speech Falters under High Attention Demands. Current Biology 15, 839–843. 10.1016/j.cub.2005.03.046 [PubMed: 15886102]

Alsius A, Navarra J, Soto-Faraco S, 2007. Attention to touch weakens audiovisual speech integration. Experimental Brain Research 183, 399–404. 10.1007/s00221-007-1110-1 [PubMed: 17899043]

Arnold P, Hill F, 2001. Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. British Journal of Psychology 92, 339–355. 10.1348/000712601162220

Baart M, 2016. Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays: Audiovisual speech integration at the N1 and P2. Psychophysiology 53, 1295–1306. 10.1111/psyp.12683 [PubMed: 27295181]

Basu Mallick D, F. Magnotti J, S. Beauchamp M, 2015. Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. Psychonomic Bulletin & Review 22, 1299–1307. 10.3758/s13423-015-0817-4 [PubMed: 25802068]

Bell AM, 1867. Visible Speech: The Science of Universal Alphabetics Or, Self-interpreting Physiological Letters, for the Writing of All Languages in One Alphabet. Simpkin, Marshall & Company.

Bernstein LE, Demorest ME, Coulter DC, O'Connell MP, 1991. Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. The Journal of the Acoustical Society of America 90, 2971–2984. [PubMed: 1838561]

Bertelson P, Aschersleben G, 1998. Automatic visual bias of perceived auditory location. Psychonomic bulletin & review 5, 482–489.

Bertelson P, Vroomen J, Wiegeraad G, Gelder B. de, 1994. Exploring the relation between McGurk interference and ventriloquism, in: Third International Conference on Spoken Language Processing.

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, 3531–3539. 10.1121/1.4965968 [PubMed: 27908052]

Bregman AS, 1990. Auditory scene analysis: the perceptual organization of sound. MIT Press, Cambridge, Mass.

Brancazio L, Miller JL, Mondini M, 2002. Audiovisual integration in the absence of a McGurk effect. The Journal of the Acoustical Society of America 111, 2433. 10.1121/1.4778348

Bruderer AG, Danielson DK, Kandhadai P, Werker JF, 2015. Sensorimotor influences on speech perception in infancy. Proceedings of the National Academy of Sciences 112, 13531–13536.

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS, 1997. Activation of auditory cortex during silent lipreading. science 276, 593–596. [PubMed: 9110978]

Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, Brammer MJ, David AS, 2001. Cortical substrates for the perception of face actions: an fMRI study of the specificity of

activation for seen speech and for meaningless lower-face acts (gurning). Cognitive Brain Research 12, 233–243. [PubMed: 11587893]

Derrick D, Anderson P, Gick B, Green S, 2009. Characteristics of air puffs produced in English “pa”: Experiments and simulations. The Journal of the Acoustical Society of America 125, 2272–2281. [PubMed: 19354402]

Derrick D, De Rybel T, 2015. System for audio analysis and perception enhancement. WO 2015/122785 A1.

Derrick D, De Rybel T, Fiasson R, 2015. Recording and reproducing speech airflow outside the mouth. Canadian Acoustics 43.

Derrick D, Gick B, 2013. Aerotactile Integration from Distal Skin Stimuli. Multisensory Research 26, 405–416. 10.1163/22134808-00002427 [PubMed: 24649526]

Derrick D, O’Beirne GA, Rybel T. de, Hay J, 2014. Aero-tactile integration in fricatives: Converting audio to air flow information for speech perception enhancement, in: Fifteenth Annual Conference of the International Speech Communication Association.

Erickson LC, Zielinski BA, Zielinski JEV, Liu G, Turkeltaub PE, Leaver AM, Rauschecker JP, 2014. Distinct cortical locations for integration of audiovisual speech and the McGurk effect. Frontiers in Psychology 5. 10.3389/fpsyg.2014.00534

Fisher BD, Pylyshyn ZW, 1994. The cognitive architecture of bimodal event perception: a commentary and addendum to Radeau. Current Psychology of Cognition 92–96.

Fowler CA, 2010. Speech Production, in: Weiner IB, Craighead WE (Eds.), The Corsini Encyclopedia of Psychology. John Wiley & Sons, Inc., Hoboken, NJ, USA. 10.1002/9780470479216.corpsy0933

Fowler CA, 1996. Listeners do hear sounds, not tongues. J. Acoust. Soc. Am 99, 1730–1741. [PubMed: 8819862]

Fowler CA, 1986. An event approach to the study of speech perception from a direct-realist perspective. Status report on speech research 14, 139.

Fowler CA, Dekle DJ, 1991. Listening with eye and hand: Cross-modal contributions to speech perception. Journal of Experimental Psychology: Human Perception and Performance 17, 816–828. 10.1037/0096-1523.17.3.816 [PubMed: 1834793]

Frayne E, Coulson S, Adams R, Croxson G, Waddington G, 2016. Proprioceptive ability at the lips and jaw measured using the same psychophysical discrimination task. Experimental Brain Research 234, 1679–1687. 10.1007/s00221-016-4573-0 [PubMed: 26860522]

Ghosh SS, Matthies ML, Maas E, Hanson A, Tiede M, Ménard L, Guenther FH, Lane H, Perkell JS, 2010. An investigation of the relation between sibilant production and somatosensory and auditory acuity. The Journal of the Acoustical Society of America 128, 3079–3087. [PubMed: 21110603]

Gibson JJ, 1966. The senses considered as perceptual systems. Greenwood Press, Westport, Conn.

Gibson JJ, 1979. The ecological approach to visual perception. Houghton Mifflin, Boston.

Gick B, Bliss H, Michelson K, Radanov B, 2012. Articulation without acoustics: “Soundless” vowels in Oneida and Blackfoot. Journal of Phonetics 40, 46–53. 10.1016/j.wocn.2011.09.002

Gick B, Derrick D, 2009. Aero-tactile integration in speech perception. Nature 462, 502. [PubMed: 19940925]

Gick B, Ikegami Y, Derrick D, 2010. The temporal window of audio-tactile integration in speech perception. The Journal of the Acoustical Society of America 128, EL342–EL346. 10.1121/1.3505759 [PubMed: 21110549]

Gick B, Jóhannsdóttir KM, Gibraiel D, Mühlbauer J, 2008. Tactile enhancement of auditory and visual speech perception in untrained perceivers. The Journal of the Acoustical Society of America 123, EL72–EL76. 10.1121/1.2884349 [PubMed: 18396924]

Goldenberg D, Tiede MK, Whalen D, 2015. Aero-tactile influence on speech perception of voicing continua. In Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow: Int. Phon. Assoc. 5 pp.

Green KP, Gerdman A, 1995. Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. Journal of Experimental Psychology: Human Perception and Performance 21, 1409–1426. 10.1037/0096-1523.21.6.1409 [PubMed: 7490588]

Green KP, Kuhl PK, Meltzoff AN, 1988. Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. The Journal of the Acoustical Society of America 84, S155–S155. 10.1121/1.2025888

Green KP, Kuhl PK, Meltzoff AN, Stevens EB, 1991. Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. Perception & Psychophysics 50, 524–536. 10.3758/BF03207536 [PubMed: 1780200]

Holmes NP, Spence C, 2005. Multisensory Integration: Space, Time and Superadditivity. Current Biology 15, R762–R764. 10.1016/j.cub.2005.08.058 [PubMed: 16169476]

Hsiao S, Gomez-Ramirez M, 2011. Touch, in: Neurobiology of Sensation and Reward. p. 141.

Ito T, Ostry DJ, 2012. Speech sounds alter facial skin sensation. Journal of Neurophysiology 107, 442–447. 10.1152/jn.00029.2011 [PubMed: 22013241]

Ito T, Tiede M, Ostry DJ, 2009. Somatosensory function in speech perception. Proceedings of the National Academy of Sciences 106, 1245–1248. 10.1073/pnas.0810063106

Jones JA, Munhall KG, 1997. Effects of separating auditory and visual sources on audiovisual integration of speech. Canadian Acoustics 25, 13–19.

Kelso JS, Tuller B, Vatikiotis-Bateson E, Fowler CA, 1984. Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. Journal of Experimental Psychology: Human Perception and Performance 10, 812. [PubMed: 6239907]

Keough M, Schellenberg M, Gick B, 2016. Spatial congruence in multimodal speech perception. The Journal of the Acoustical Society of America 140, 3225–3225. 10.1121/1.4970186

Keough M, Taylor RC, Derrick D, Schellenberg M, Gick B, 2017. Sensory Integration from an Impossible Source: Perceiving Simulated Faces. Canadian Acoustics 45, 176–177.

Klucharev V, Möttönen R, Sams M, 2003. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Cognitive Brain Research 18, 65–75. 10.1016/j.cogbrainres.2003.09.004 [PubMed: 14659498]

Macaluso E, Driver J, 2005. Multisensory spatial interactions: a window onto functional integration in the human brain. Trends in Neurosciences 28, 264–271. 10.1016/j.tins.2005.03.008 [PubMed: 15866201]

Macaluso E, George N, Dolan R, Spence C, Driver J, 2004. Spatial and temporal factors during processing of audiovisual speech: a PET study. NeuroImage 21, 725–732. 10.1016/j.neuroimage.2003.09.049 [PubMed: 14980575]

MacSweeney M, Amaro E, Calvert GA, Campbell R, David AS, McGuire P, Williams SC, Woll B, Brammer MJ, 2000. Silent speechreading in the absence of scanner noise: an event-related fMRI study. Neuroreport 11, 1729–1733. [PubMed: 10852233]

Massaro DW, 1998. Perceiving talking faces: from speech perception to a behavioral principle, MIT Press/Bradford Books series in cognitive psychology. MIT Press, Cambridge, Mass.

Massaro DW, 1987. Speech perception by ear and eye: a paradigm for psychological inquiry. Erlbaum Associates, Hillsdale, N.J.

Massaro DW, Chen TH, 2008. The motor theory of speech perception revisited. Psychonomic Bulletin & Review 15, 453–457. 10.3758/PBR.15.2.453

McGrath M, Summerfield Q, 1985. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. The Journal of the Acoustical Society of America 77, 678–685. 10.1121/1.392336 [PubMed: 3973239]

Mcgurk H, Macdonald J, 1976. Hearing lips and seeing voices. Nature 264, 746–748. 10.1038/264746a0 [PubMed: 1012311]

Meister IG, Wilson SM, Deblieck C, Wu AD, Iacoboni M, 2007. The Essential Role of Premotor Cortex in Speech Perception. Current Biology 17, 1692–1696. 10.1016/j.cub.2007.08.064 [PubMed: 17900904]

Ménard L, Turgeon C, Trudeau-Fisette P, Bellavance-Courtemanche M, 2016. Effects of blindness on production–perception relationships: Compensation strategies for a lip-tube perturbation of the French [u]. Clinical Linguistics & Phonetics 30, 227–248. 10.3109/02699206.2015.1079247 [PubMed: 26403592]

Munhall KG, Gribble P, Sacco L, Ward M, 1996. Temporal constraints on the McGurk effect. Perception & Psychophysics 58, 351–362. 10.3758/BF03206811 [PubMed: 8935896]

Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E, 2004. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychological science 15, 133–137. [PubMed: 14738521]

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. 10.1007/s00221-005-0071-5 [PubMed: 16217645]

Nasir SM, Ostry DJ, 2006. Somatosensory Precision in Speech Production. Current Biology 16, 1918–1923. 10.1016/j.cub.2006.07.069 [PubMed: 17027488]

Parker AJ, Krug K, 2003. Neuronal mechanisms for the perception of ambiguous stimuli. Current Opinion in Neurobiology 13, 433–439. 10.1016/S0959-4388(03)00099-0 [PubMed: 12965290]

Perkell JS, 2012. Movement goals and feedback and feedforward control mechanisms in speech production. Journal of Neurolinguistics 25, 382–407. [PubMed: 22661828]

Pickering MJ, Garrod S, 2013. An integrated theory of language production and comprehension. Behavioral and Brain Sciences 36, 329–347. 10.1017/S0140525X12001495

Reed CM, 1996. The implications of the Tadoma method of speechreading for spoken language processing, in: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference On. IEEE, pp. 1489–1492.

Reed CM, Durlach NI, Braida LD, Schultz MC, 1989. Analytic Study of the Tadoma Method: Effects of Hand Position on Segmental Speech Perception. Journal of Speech Language and Hearing Research 32, 921. 10.1044/jshr.3204.921

Reed CM, Rubin SI, Braida LD, Durlach NI, 1978. Analytic Study of the Tadoma Method: Discrimination Ability of Untrained Observers. Journal of Speech Language and Hearing Research 21, 625. 10.1044/jshr.2104.625

Reisberg D, Mclean J, Goldfield A, 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli.

Rosenblum LD, Saldaña HM, 1996. An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology: Human Perception and Performance 22, 318. [PubMed: 8934846]

Rosenblum LD, Schmuckler MA, Johnson JA, 1997. The McGurk effect in infants. Perception & Psychophysics 59, 347–357. [PubMed: 9136265]

Sams M, Aulanko R, Hämäläinen M, Hari R, Lounasmaa OV, 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. 10.1016/0304-3940(91)90914-F [PubMed: 1881611]

Schwartz J-L, 2010. A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. The Journal of the Acoustical Society of America 127, 1584–1594. 10.1121/1.3293001 [PubMed: 20329858]

Scott M, 2012. Speech imagery as corollary discharge (PhD Thesis). University of British Columbia.

Shen G, Meltzoff AN, Marshall PJ, 2018. Touching lips and hearing fingers: effector-specific congruency between tactile and auditory stimulation modulates N1 amplitude and alpha desynchronization. Experimental Brain Research 236, 13–29. 10.1007/s00221-017-5104-3 [PubMed: 29038847]

Soto-Faraco S, Kingstone A, Spence C, 2003. Multisensory contributions to the perception of motion. Neuropsychologia 41, 1847–1862. [PubMed: 14527547]

Sparks DW, Kuhl PK, Edmonds AE, Gray GP, 1978. Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of segmental features of speech. The Journal of the Acoustical Society of America 63, 246–257. [PubMed: 632417]

Sperdin HF, Cappe C, Murray MM, 2010. Auditory–somatosensory multisensory interactions in humans: Dissociating detection and spatial discrimination. Neuropsychologia 48, 3696–3705. 10.1016/j.neuropsychologia.2010.09.001 [PubMed: 20833194]

Studdert-Kennedy M, Mattingly IG, 2014. Modularity and the Motor theory of Speech Perception: Proceedings of A Conference To Honor Alvin M. Liberman Taylor and Francis, Hoboken.

Sumby WH, Pollack I, 1954. Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America 26, 212–215. 10.1121/1.1907309

Theys C & McAuliffe M (2014a) Neurophysiological correlates associated with the perception of dysarthric speech. American Speech-Language-Hearing Association Annual Convention. Florida, USA.

Theys C & McAuliffe M (2014b) Auditory processing of dysarthric speech: an EEG study. Australasian Winter Conference on Brain Research. Queenstown, New Zealand.

Tian X, Poeppel D, 2010. Mental imagery of speech and movement implicates the dynamics of internal forward models. Frontiers in Psychology 1, 166. [PubMed: 21897822]

Tiippana K, 2014. What is the McGurk effect? Frontiers in Psychology 5. 10.3389/fpsyg.2014.00725

Tiippana K, Andersen TS, Sams M, 2004. Visual attention modulates audiovisual speech perception. European Journal of Cognitive Psychology 16, 457–472. 10.1080/09541440340000268

Tremblay S, Shiller DM, Ostry DJ, 2003. Somatosensory basis of speech production. Nature 423, 866. [PubMed: 12815431]

van Wassenhove V, 2013. Speech through ears and eyes: interfacing the senses with the supramodal brain. Frontiers in Psychology 4. 10.3389/fpsyg.2013.00388

van Wassenhove V, Grant KW, Poeppel D, 2007. Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45, 598–607. 10.1016/j.neuropsychologia.2006.01.001 [PubMed: 16530232]

Vatikiotis-Bateson E, Kuratate T, Munhall K, Yehia H, 2000. The production and perception of a realistic talking face. Proceedings of LP 98, 439–460.

Vatikiotis-bateson E, Munhall KG, Hirayama M, Lee Y, Terzopoulos D, 1996. The dynamics of audiovisual behavior in speech, in: Speechreading by Humans and Machines: Models, Systems, and Applications, Volume 150 of NATO ASI Series. Series F: Computer and Systems Sciences. Springer-Verlag, pp. 221–232.

Vatikiotis-Bateson E, Munhall KG, Hirayama M, Lee YV, Terzopoulos D, 1996. The Dynamics of Audiovisual Behavior in Speech, in: Stork DG, Hennecke ME (Eds.), Speechreading by Humans and Machines. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 221–232. 10.1007/978-3-662-13015-5_16

Vatikiotis-Bateson E, Munhall KG, Kasahara Y, Garcia F, Yehia H, 1996. Characterizing audiovisual information during speech. IEEE, pp. 1485–1488. 10.1109/ICSLP.1996.607897

Vivian RM, 1966. TADOMA METHOD-TACTUAL APPROACH TO SPEECH AND SPEECHREADING. Volta Review 68, 733–737.

Walker S, Bruce V, O'Malley C, 1995. Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. Perception & Psychophysics 57, 1124–1133. 10.3758/BF03208369 [PubMed: 8539088]

Wilson EC, Reed CM, Braida LD, 2009. Integration of auditory and vibrotactile stimuli: Effects of phase and stimulus-onset asynchrony. The Journal of the Acoustical Society of America 126, 1960–1974. [PubMed: 19813808]

Yehia H, Rubin P, Vatikiotis-Bateson E, 1998. Quantitative Association Of Orofacial And Vocal-Tract Shapes, in: In Proceedings of the Workshop on Audio-Visual Speech Processing. pp. 41–44.

Yeung HH, Werker JF, 2009. Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. Cognition 113, 234–243. [PubMed: 19765698]