JSLHR

Research Article

Individual Differences in Susceptibility to the McGurk Effect: Links With Lipreading and Detecting Audiovisual Incongruity

Julia Strand,[a] Allison Cooperman,[a] Jonathon Rowe,[a] and Andrea Simenstad[a]

Purpose: Prior studies (e.g., Nath & Beauchamp, 2012) report large individual variability in the extent to which participants are susceptible to the McGurk effect, a prominent audiovisual (AV) speech illusion. The current study evaluated whether susceptibility to the McGurk effect (MGS) is related to lipreading skill and whether multiple measures of MGS that have been used previously are correlated. In addition, it evaluated the test–retest reliability of individual differences in MGS.

Method: Seventy-three college-age participants completed 2 tasks measuring MGS and 3 measures of lipreading skill. Fifty-eight participants returned for a 2nd session (approximately 2 months later) in which MGS was tested again.

Results: The current study demonstrated that MGS shows high test–retest reliability and is correlated with some measures of lipreading skill. In addition, susceptibility measures derived from identification tasks were moderately related to the ability to detect instances of AV incongruity.

Conclusions: Although MGS is often cited as a demonstration of AV integration, the results suggest that perceiving the illusion depends in part on individual differences in lipreading skill and detecting AV incongruity. Therefore, individual differences in susceptibility to the illusion are not solely attributable to individual differences in AV integration ability.

Speech perception is not an exclusively auditory process. Instead, visual information about a speaker, such as mouth shape and facial movements, supplement the auditory signal (Sumby & Pollack, 1954). As a result, listeners accurately perceive speech at more difficult signal-to-noise ratios when they can see and hear talkers relative to only hearing them (Middelweerd & Plomp, 1987). The benefit of the visual signal extends to many listening situations: Participants are better able to repeat passages of complex text when they can see the talker (Reisberg, McLean, & Goldfield, 1987), are better able to detect the presence of auditory stimuli at very low amplitudes when accompanied by a visual signal (Bernstein, Auer, & Takayanagi, 2004), and are less impaired by auditory distractors when visual information is present (Spence, Ranson, & Driver, 2000). A classic and compelling demonstration of the influence of the visual signal on auditory speech is the McGurk effect (McGurk & MacDonald, 1976). In this paradigm, a video of a face uttering a syllable is dubbed with an auditory stimulus of a different syllable, such as a visual "ga" with an auditory "ba." This often causes individuals to perceive a *fusion*, a third syllable that combines features of both the visual and auditory utterances (i.e., "da").

The McGurk effect is a robust illusion. For example, it can occur when the face and voice are different genders (Green, Kuhl, Meltzoff, & Stevens, 1991), when participants are explicitly warned that the auditory and visual signals may not match (Summerfield & McGrath, 1984), and even when the auditory and visual stimuli were recorded by the person perceiving the illusion (Aruffo & Shore, 2012). As a result, the McGurk effect is frequently cited as evidence for the powerful and automatic influence of the visual signal on auditory speech (Colin et al., 2002; Easton & Basala, 1982; Massaro, 1987; Rosenblum & Saldaña, 1996; Soto-Faraco, Navarra, & Alsius, 2004). However, there are large intersubject differences in susceptibility to the illusion: The frequency with which individuals perceive McGurk fusions varies greatly (0% to 100% of trials, Nath & Beauchamp, 2012; 1% to 91% of trials, Benoit, Raij, Lin, Jääskeläinen, & Stufflebeam, 2010; also see Cienkowski & Carney, 2002; Jin & Carney, 2000). These results suggest that some individuals are influenced by the visual signal nearly every time

[a]Carleton College, Northfield, MN

Correspondence to: jstrand@carleton.edu

Editor: Nancy Tye-Murray
Associate Editor: Kathleen Cienkowski

Received February 18, 2014
Revision received July 2, 2014
Accepted September 19, 2014
DOI: 10.1044/2014_JSLHR-H-14-0059

***Disclosure:** The authors have declared that no competing interests existed at the time of publication.*

they are presented with a McGurk stimulus whereas others rarely (or never) are.

The causes for the large individual differences in susceptibility to the McGurk effect (MGS) have not been determined. However, models of audiovisual (AV) integration specify at least two points during speech processing at which individual differences might emerge: unimodal identification and AV integration. Although implementations differ, models of AV integration agree that participants must first extract information from the unimodal (auditory and visual) signals and then combine the extracted information to arrive at an AV percept (Braida, 1991; Grant & Seitz, 1998; Massaro & Cohen, 2000; see Schwartz, Robert-Ribes, & Escudier, 1998, for a taxonomy of models of AV integration). MGS might vary across individuals due to individual differences in either the extraction stage or the integration stage (but see Massaro & Cohen, 2000, for arguments against individual differences in integration ability). Classically, research on the McGurk effect has focused on the mechanism of integration (see Green, 1998), and indeed, perceiving a McGurk fusion requires that integration occurs. Critically, however, failing to perceive an AV fusion does not mean that integration failed. Rather, the absence of a McGurk fusion may instead be caused by inaccurate or missing information from one of the unimodal signals. Therefore, it is not clear whether the observed variability in MGS represents individual differences in integration skill or in unimodal extraction ability (e.g., lipreading).

Models of AV integration would predict that poorer lipreading skill should lead to reduced MGS (Braida, 1991; Grant & Seitz, 1998; Massaro & Cohen, 2000), although this prediction has not yet been empirically supported. There are large individual differences in lipreading skill (Auer & Bernstein, 2007; Feld & Sommers, 2009; Lyxell & Holmberg, 2000), and poor lipreading will necessarily limit the amount that AV recognition scores differ from auditory-only scores. For instance, even with perfect AV integration, if a participant fails to extract any meaningful information from the visual signal, a McGurk fusion cannot occur. Therefore, individuals who are poor lipreaders may appear less susceptible to the McGurk effect simply because they are unable to correctly identify the linguistic information in the visual signal, and as a result, rely solely on the auditory signal. Indeed, when participants fail to report a McGurk fusion, they most commonly report the auditory portion alone (Cienkowski & Carney, 2002). This may suggest that individual differences in MGS are, in fact, partly attributable to individual differences in lipreading ability.

Only one prior study has collected data on MGS and lipreading ability from the same participants (Cienkowski & Carney, 2002) and found that lipreading ability was not related to the number of McGurk fusions participants reported. However, the lipreading materials in that study were full sentences, which may over- or underestimate the phonetic information that participants extract from the visual signal and, hence, affect their lipreading skill. For example, in a visual-only phoneme identification task, the phonemes /b/, /m/, and /p/ are very easily confusable because they use the same place of articulation (POA; Binnie, Montgomery, & Jackson, 1974). However, in a sentence such as “She read an interesting book,” a lipreader need not be able to distinguish between the /b/, /m/, and /p/ of the word “book” to understand the sentence because “pook” and “mook” are not real English words. An individual who cannot distinguish between /b/ and /p/ in isolation may therefore perform as well on sentence identification as an individual who can distinguish between the two phonemes. Alternatively, full-length sentence materials may underestimate the phonetic information that participants can extract. For example, a participant may be able to distinguish between two phonemes when they are presented in isolation but become unable to when they are presented more quickly and with greater coarticulation in a sentence context. Therefore, it remains unclear whether MGS, which is typically measured using syllable- rather than sentence-length materials, depends in part on individual differences in the ability to extract linguistic information from the visual signal. To establish whether it is appropriate to use MGS as a measure of AV integration, it is necessary to establish whether some of the variability in MGS can, in fact, be attributed to unimodal extraction ability.

Another factor that complicates the claim that MGS is a measure of individual differences in AV integration is that MGS has been measured in several ways, but these measures have not been compared. In prior studies, MGS has most commonly been quantified by showing participants multiple presentations of McGurk stimuli and asking them to identify the syllable that they perceived (Grant & Seitz, 1998; Nath & Beauchamp, 2012). An alternative method of measuring MGS is to ask participants to detect whether the auditory and visual signals are the same phoneme (congruent) or are different phonemes (incongruent; Benoit et al., 2010). Highly susceptible participants are those who report the fusion (in the case of the identification task) or fail to notice incongruity (in the detection task). Classic work on the illusion has stressed that, when participants perceive the illusion, they often do not detect that the auditory and visual stimuli were incongruent (Summerfield & McGrath, 1984). Given that, measures of MGS based on identification might be expected to correlate with measures of MGS based on incongruity detection with the assumption that, when participants notice incongruity, AV integration is prevented (but see Munhall, Gribble, Sacco, & Ward, 1996, and Soto-Faraco & Alsius, 2007, for evidence that the McGurk effect can still occur under noticeable temporal incongruity). This is in line with participants following the “unity assumption,” in which multisensory integration is more likely to occur if the two sensory sources of information seem highly consistent with one another (Vatakis & Spence, 2007; Welch & Warren, 1980). Indeed, manipulations in which the incongruity is more readily apparent, such as when the voice and face are of different genders, reduce the magnitude of, but do not eliminate, the illusion (Easton & Basala, 1982; Vatakis & Spence, 2007; but see also Green et al., 1991).

However, other work indicates that perception of the illusion does not depend on failing to detect the incongruity: Participants may sometimes be able to detect a temporal mismatch between the auditory and visual stimuli while still experiencing a perceptual fusion (Soto-Faraco & Alsius, 2007). In addition, evidence from neuroimaging studies demonstrates that there are distinct neural systems that mediate evaluating the relationship between cross-modal stimuli and those that underlie integration (Miller & D'Esposito, 2005). AV integration resulted in activation in the superior temporal sulcus and inferior frontal gyrus, whereas detecting temporal correspondence of auditory and visual stimuli tokens activated the superior colliculus, anterior intraparietal sulcus, and anterior insula, demonstrating cortical distinctions between AV integration and incongruity detection (Miller & D'Esposito, 2005). If detecting incongruity does not preclude AV integration, then individual differences in MGS based on incongruity tasks may not predict individual differences in MGS based on identification tasks. No studies to date have systematically compared individual differences in incongruity detection tasks and identification tasks to evaluate the relationship between the two.

Assessment of test–retest reliability of individual differences in MGS is also absent from the literature. Given that measures of MGS correlate with other measures of AV integration, such as susceptibility to the illusory flash effect (a nonspeech AV illusion; see Tremblay, Champoux, Bacon, & Theoret, 2007) and ability to benefit from the addition of visual information to auditory speech (Grant & Seitz, 1998), MGS is likely to be a stable trait on which individuals reliably differ. In addition, other related perceptual abilities, including lipreading (Macleod & Summerfield, 1990) and speech perception in noise (Bentler, 2000) show good test–retest reliability, so it would not be surprising to expect the same of MGS. However, no research has systematically tested this hypothesis by measuring MGS in the same participants in multiple sessions over time. Including a measure of test–retest reliability will also provide a comparison point for evaluating the correlation between measures of MGS derived from identification and incongruity detection tasks. That is, if there is a weak relationship between MGS based on identification and incongruity detection tasks, but MGS also shows poor stability over time, it could indicate that measures of MGS are unreliable. If, however, test–retest reliability is strong, and measures of MGS through identification and incongruity detection tasks are weakly correlated, it would suggest that the measures are reliable but that identification and incongruity detection tasks are measuring somewhat different processes.

Given the widespread use of the McGurk effect in research on AV integration, the current study seeks to evaluate whether MGS is in fact independent of lipreading skill even when measures are used that are highly sensitive to the amount of information extracted from the visual signal. In addition, the study will test whether individual differences in MGS derived from identification and incongruity detection tasks are comparable, following the assumption of earlier work that AV integration occurs when incongruity is not detected (Summerfield & McGrath, 1984). If measures of MGS based on syllable identification and measures based on incongruity detection are in fact correlated, it would indicate that individuals who are more likely to notice incongruity between the auditory and visual stimuli are also less likely to perceive McGurk fusions. We also evaluate test–retest reliability of individual differences in MGS.

Method

Testing was completed in two sessions to enable test–retest comparisons.

Session One

Participants

Seventy-three participants (57 female, 16 male) between the ages of 18 and 22 (*M* = 20 years, *SD* = 1.3) were recruited from Carleton College. All participants were native English speakers and self-reported having normal hearing and normal or corrected-to-normal vision. Participants were compensated $10 for 1 hr of participation.

Stimuli

One male and one female with standard Midwestern accents served as speakers for all speech materials (materials were adopted from Strand & Sommers, 2011). Stimuli were recorded with a Cannon Elura 85 digital video camera connected to a Dell Precision PC at a 16-bit resolution and sampling rate of 48,000. Digital capture was done in Adobe Premiere Elements 1.0, auditory stimuli were equated for root–mean–square amplitude using Adobe Audition, and video editing was done with iMovie (Version 9.0.9). Visual stimuli measured 720 × 480 pixels, were presented at 30 frames per second and showed the speakers' head and shoulders directly facing the camera. To create the McGurk stimuli, we first overlaid the auditory track of one syllable on another AV track. The new audio track was then aligned with the original audio track to ensure that the onset of the consonant bursts matched (Munhall et al., 1996), and the original audio track was then deleted. Auditory stimuli were presented at approximately 68 dB sound pressure level via the computer's internal speakers.

Procedure

All participants were tested individually and completed tasks in a consistent order: McGurk identification, McGurk detection, lipreading consonants, and lipreading words. Given that the focus of the study was on individual differences, we used a consistent order for all participants rather than counterbalancing to ensure that no experimental differences influenced the performance of individual participants. Participants were seated in a sound-attenuating chamber at a comfortable distance from an iMac computer (OS X, 10.6) and were instructed to watch the face of the speaker for all tasks; between trials, they were asked to fixate on a cross positioned approximately where the mouth of the talker would appear. Stimulus presentation and participant responses to the speech tasks were controlled with PsyScope (Version X, B57). Instructions were given orally and in writing, and participants completed five practice trials for each task.

McGurk identification. Identification trials included AV consonant stimuli presented in an "aCa" context. This included congruent AV tokens (e.g., auditory "aBa" paired with visual "aBa") and stimuli expected to result in the McGurk effect:[1] $A_bV_f = AV_v$, $A_bV_g = AV_d$, $A_mV_g = AV_n$, $A_mV_t = AV_n$, $A_pV_g = AV_k$, $A_pV_k = AV_t$, $A_tV_b = AV_p$. The congruent audiovisual tokens were /b/, /d/, /f/, /g/, /k/, /m/, /n/, /p/, /t/, and /v/. Each participant completed 122 identifications: 80 were congruent AV tokens (10 tokens presented four times by two speakers) and 42 were McGurk tokens (seven tokens presented three times each by two speakers). After presentation of the stimulus item, participants were prompted to identify the syllable that they perceived from among 10 possibilities (b, d, f, g, k, m, n, p, t, v), and their responses were recorded via key press. The answer options included all the auditory and visual stimuli as well as all likely fusions, based on combined POA from the visual signal and voicing from the auditory signal (see Binnie et al., 1974). Congruent and McGurk stimuli were randomly intermixed but presented to all participants in the same order, blocked by speaker.

MGS in the identification task (MGS-ID) was quantified as the proportion of trials in which participants' responses were influenced by both the visual and auditory signal. This included reporting the expected fusion (e.g., $A_mV_t = AV_n$) or a response that was consistent with the POA of the visual signal and voicing of the auditory signal (e.g., $A_mV_t = AV_d$). This method of scoring MGS is somewhat more flexible than only counting optimal fusions as correct and ensures that all responses that showed simultaneous influence of the auditory and visual signals are counted as evidence of integration. However, the method also helps ensure that responses are only counted as McGurk fusions when they show the influence of both the auditory and visual signals. For example, $A_mV_t = AV_v$ would not be counted as a fusion because, although the participant reported a different phoneme than either the auditory or the visual signal, it does not match the POA of the visual signal, the most salient visual feature (G. A. Miller & Nicely, 1955). Following the procedures of Nath and Beauchamp (2012) and Cienkowski and Carney (2002), trials in which the visual stimulus alone was reported were also not counted as fusions.

McGurk detection. The McGurk detection task was similar to the identification task, but rather than identifying the syllable that was presented, participants were asked to report whether the auditory and visual signals matched by pressing keys labeled "yes" (indicating the auditory and visual stimuli were the same speech sound) or "no" (indicating they were different) following each stimulus. The same 122 trials (80 congruent stimuli, 42 McGurk stimuli) were presented along with 56 incongruent AV stimuli that were not expected to result in perceptual fusions. For example, A_gV_b often results in a combination response in which participants either report having heard "gba" or clearly detect the mismatch (McGurk & MacDonald, 1976). Following the procedures of Benoit et al. (2010), these trials were included out of concern that participants who are highly susceptible to MGS might never press the "no" key (indicating incongruity) to any of the congruent or McGurk trials. The incongruent trials included two repetitions of seven stimuli by each speaker, A_bV_t, A_fV_b, A_gV_b, A_gV_m, A_gV_p, and A_kV_p.[2] Although it might be preferable to present equal numbers of stimuli to which participants would respond "congruent" and "incongruent" to avoid response biases, the large individual variability in the ability to detect incongruity makes this difficult. Incongruent trials were randomly intermixed with the congruent and McGurk stimuli but presented to all participants in the same order, blocked by each speaker.

To quantify MGS in the detection task, we first calculated the rates at which participants reported that the auditory and visual signals of the McGurk stimuli were congruent (MGS-detect), following the method of Benoit et al. (2010). Higher values of MGS-detect indicate that participants were less likely to notice the incongruity between the auditory and visual signals. Although rates of responding that McGurk stimuli are congruent may be interpreted as a measure of how often participants are integrating the mismatched auditory and visual stimuli, a limitation of this measure is that response biases can significantly distort the results. For example, participants who respond "incongruent" in every trial, regardless of stimulus type, would appear to have very low rates of MGS but would also be incorrect on every congruent trial.

In order to incorporate individual differences in responding to both the McGurk stimuli and the congruent stimuli, we also calculated d' values (Macmillan & Creelman, 2004; for other applications of d' in speech perception research, see Iverson et al., 2003; Kaplan-Neeman, Kishon-Rabin, Henkin, & Muchnik, 2006). These values (MGS-d') were calculated using both the rates at which participants incorrectly reported that the auditory and visual signals of the McGurk stimuli were congruent (i.e., the MGS-detect measure or false alarms in the language of signal detection theory) as well as the rates at which participants correctly reported that the congruent auditory and visual signals were congruent (i.e., hits). Given that d' represents the extent to which a participant is able to discriminate between the two trial types (congruent and McGurk stimuli), the highest

[1] Here, the auditory syllable is represented as the subscript of "A," the visual syllable is the subscript of "V," and the expected perception is the subscript of "AV."

[2] We do not report expected perceptions for the incongruent trials because we did not anticipate participants reporting fusions for these trials.

MGS-d' values are generated when a participant consistently reports that the congruent stimuli are congruent and the McGurk stimuli are incongruent. Following the recommendations of Brown and White (2005) and Hautus (1995), we applied the log-linear rule transformation and added 0.5 to each cell before calculating d' values. This enables d' values to be calculated from responses in which individual participants show 100% hit rates. Although MGS-detect and MGS-d' will necessarily be correlated because one is derived from the other, both measures are reported in the results because MGS-detect has been reported previously as a measure of integration (Benoit et al., 2010) and MGS-d' offers some additional computational benefits (see Results). We did not analyze the incongruent filler trials.

Lipreading consonants. Participants identified 120 trials of visual-only consonants in an "aCa" context: six presentations of the 10 consonants used as congruent stimuli in the McGurk tasks produced by the same two speakers used in the McGurk tasks. After each trial, participants were asked to identify the syllable using the same 10 options as in the McGurk identification task. All participants saw the stimuli in the same randomized order, blocked by speaker.

Accuracy in the consonant identification task (LR-C) was quantified as the proportion of trials in which participants selected the appropriate consonant. We also calculated the proportion of trials in which each participant correctly identified the POA of the phoneme, using the groupings {b, m, p} {f, v} and {d, n, t} {k, g}. POA is the feature most easily identified in lipread speech, and therefore, participants are far more likely to correctly identify the POA but misidentify information about voicing than the reverse pattern (Jackson, 1988). For example, for the LR-C score, a response of /p/ when presented with /b/ would be counted as incorrect, but for the POA score (LR-POA), it would count as correct. LR-POA scores are informative because they help assess the amount of information that was extracted from the visual signal: An incorrect response that shares the POA with the stimulus phoneme shows that the participant extracted more salient information out of the visual signal than an incorrect response that does not share the POA.

Lipreading words. Participants identified 60 visual-only consonant–vowel–consonant words (30 by each speaker). These stimuli were chosen from a larger set of words used in a prior study (Strand & Sommers, 2011) and were selected because they rendered a range of lipreading difficulty in that study (0%–83% accuracy, M = 36, SD = 22). All stimuli were presented at the end of a carrier phrase ("say the word . . ."). Participants lipread the sentence and were then prompted to type a free response identifying the final word. They were encouraged to guess when unsure. All participants saw the stimuli in the same randomized order, blocked by speaker. Accuracy in this task (LR-W) was quantified as the proportion of trials in which participants identified the word correctly. Prior to scoring, obvious typographical errors (e.g., "cat") and homophonous entries (e.g., "four" instead of "for") were corrected.

Session Two

Participants

All participants from session one were invited to participate in a follow-up session. Fifty-eight participants (49 female, nine male) between the ages of 18 and 22 (M = 19.88, SD = 1.27) completed this session.[3] The sessions were an average of 61 days apart. Participants were compensated $5 for 30 min of participation.

Procedure

Participants completed the McGurk identification task following the same procedures as in session one but with a re-randomized order of stimuli.

Results

Individual Variability

Replicating prior research (e.g., Benoit et al., 2010; Nath & Beauchamp, 2012), we found large individual variability in MGS when measured in both the identification task and the detection task. In the identification task, individual participants reported fused responses to McGurk stimuli in 0%–64% of trials (M = 24, SD = 16) in session one and in 2%–79% of trials (M = 24, SD = 18) in session two. The stimuli produced by the female speaker resulted in significantly more McGurk fusions (M = 27, SD = 19) than the male speaker (M = 22, SD = 16), $t(71) = 2.63$, $p = .01$, and a Levene's test revealed that the variability was somewhat larger for the female speaker than for the male, $t(71) = 2.1$, $p = .03$. The two talkers were not screened for equivalent visual intelligibility, and given established differences in how easy speakers are to lipread (Conrey & Gold, 2006), these differences in the rates of McGurk fusions may be attributable to speaker idiosyncrasies. When presented with McGurk stimuli, participants reported AV fusions in 24% of trials, with 23% reporting the expected fusion and 1% reporting a response consistent with POA of visual signal and voicing of the auditory signal. They reported the audio signal alone in 66% of trials, the visual signal alone in 6% of trials, and an unrelated response in 4% of trials.

Participants also showed large variability in MGS in the detection task, with MGS-detect scores (proportion of trials in which participants reported McGurk trials were congruent) ranging from 11% to 75% (M = 33, SD = 19). Similarly, MGS-d' scores (which include information about MGS-detect rates as well as rates of correctly identifying congruent trials are congruent) ranged from 0.41 to 3.45 (M = 2.43, SD = 0.71). A d' score of zero would indicate that participants responded "congruent" equally often for congruent and McGurk trials, and a score of 4.65 would

[3] Although all participants were invited to return for the second session, testing took place during two different academic terms, and some participants were no longer on campus and were therefore unable to return for the second session.

indicate that participants responded "congruent" on 99% of congruent trials and on only 1% of McGurk trials.

Responses to congruent trials in the identification task showed very high accuracy, with scores ranging from 95% to 100% ($M = 99$, $SD = 1$). This indicates that the speech stimuli were intelligible, and participants were attending to the task. In the detection task, hit rates (responding that congruent trials were congruent) were also high, with scores ranging from 75% to 100% ($M = 97$, $SD = 4$). Because the hit rates showed less variability than the false alarm rates, it suggests that MGS-d' values are primarily influenced by rates at which participants responded that McGurk stimuli were congruent.

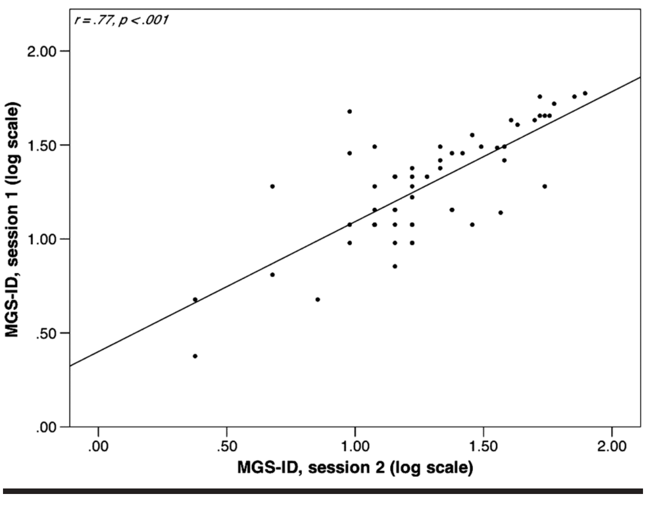
Test–Retest Reliability

MGS-ID scores were positively skewed and so were log transformed prior to analysis. Individual differences in MGS-ID was significantly correlated in sessions one and two, $r = .77$, $p < .001$ (see Figure 1), demonstrating good test–retest reliability. MGS-ID scores did not differ between session 1 and session 2, $t(57) = 0.20$, $p = .85$; Cohen's $d = 0.14$. Given that 20% of participants did not return for the second session, the measures of MGS-ID from the second session are not reported on further.

MGS-ID and MGS-d′

Like MGS-ID, MGS-detect was positively skewed and so was log transformed prior to analysis. Both MGS-detect and MGS-d' were significantly correlated with MGS-ID, although the magnitude of the correlations was relatively small (see Figure 2). The positive correlation between MGS-detect and MGS-ID indicates that participants who were more likely to perceive fusions in the identification task were also moderately more likely to report that the auditory and visual signals in McGurk trials were congruent. Because higher MGS-d' values indicate greater discrimination between congruent and McGurk stimuli, the negative correlation between MGS-ID and MGS-d' indicates that participants who were more likely to report fusions on McGurk identification trials were slightly worse at discriminating between congruent and McGurk trials in the detection task. That is, participants who showed more similar response patterns to McGurk and congruent trials in the detection task were also more likely to report fusions in the identification task.

Figure 1. Test–retest reliability of McGurk effect in the identification task (MGS-ID).



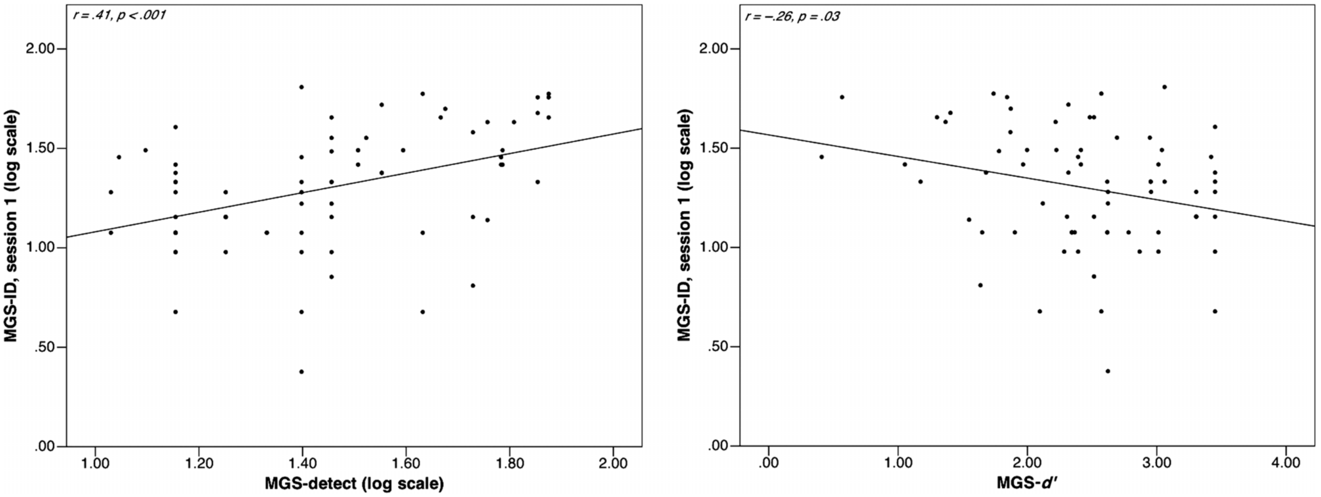
MGS and Lipreading

In line with previous research (Auer & Bernstein, 2007; Feld & Sommers, 2009; Tye-Murray, Sommers, & Spehar, 2007), participants showed large individual variability in lipreading scores. Word identification scores ranged from 2% to 55% correct ($M = 25$, $SD = 12$), consonant identification scores ranged from 28% to 49% correct ($M = 38$, $SD = 5$), and the consonant POA scores ranged from 72% to 99% correct ($M = 88$, $SD = 7$). Although lipreading scores differ across studies based on the number of response options (Watson, Qiu, Chamberlain, & Li, 1996), speaker intelligibility (Conrey & Gold, 2006; Kricos, 1985), and the population tested (Auer & Bernstein, 2007), these ranges are comparable to other published reports of individual variability in lipreading words (Sommers, Tye-Murray, & Spehar, 2005; Strand & Sommers, 2011), consonants (Demorest, Bernstein, & DeHaven, 1996; Feld & Sommers, 2011), and consonant POA (Jackson, 1988; Owens & Blazek, 1985). As expected from prior work (Bernstein, Demorest, & Tucker, 2000), the three measures of lipreading (LR-C, LR-POA, and LR-W) were all significantly intercorrelated (see Table 1).

Neither lipreading consonants nor lipreading words correlated with MGS-ID, but the more sensitive measure of accuracy at identifying POA was significantly correlated with MGS-ID, with better lipreaders showing greater susceptibility to the illusion. All three measures of lipreading were significantly correlated with MGS-d', with better lipreaders showing better discrimination between congruent and McGurk trials in the detection task. Measures of lipreading were also correlated with MGS-detect, with better lipreaders being more likely to notice incongruity when it occurred, although the correlation with LR-POA failed to reach significance ($p = .10$).

Given the multicollinearity between measures of MGS in the detection task and the lipreading measures, we conducted a hierarchical multiple regression to assess whether both accounted for unique variance in MGS-ID. MGS-d' added in the first step of the regression accounted for 7% ($\beta = -.26$, $p = .03$) of the variance in MGS-ID. In the second step, all three measures of lipreading were entered in stepwise fashion. The model selected LR-POA as the most powerful of the lipreading variables, which accounted for an additional 17% ($\beta = .44$, $p < .001$) of unique variance in MGS-ID, indicating that the relationship between LR-POA and MGS-ID was strengthened when MGS-d' was controlled for. When the other measures of lipreading were forced in as the second step (instead of LR-POA), they also accounted for significant variance in MGS-ID (8% of the

Figure 2. The relationship between individual differences in MGS-ID and MGS-detect (left) and between MGS-ID and MGS-*d′* (right).



variance for LR-C, β = .30, p = .02; 9% of the variance for the LR-W, β = .33, p = .01). If d' was entered after the POA measure instead of before it, it accounted for an additional 14% of the variance, indicating that the relationship between MGS-ID and MGS-d' becomes stronger when controlling for individual differences in lipreading.

A parallel analysis with the MGS-detect measure rendered a very similar pattern of results. MGS-detect accounted for 17% of variance in MGS-ID in the first step (β = .41, p < .001), and a stepwise model selected the POA measure as the most powerful predictor of the lipreading measures, accounting for an additional 16% of variance (β = .41, p < .001). When the consonant or word tasks were forced in the second step instead of the POA measure, both accounted for an additional 8% (β = .29, p = .01 for both) of unique variance. When MGS-detect was instead entered in the second step of the regression after LR-C/POA, it accounted for an additional 23% of the variance in MGS-ID (β = .49, p < .001).

Discussion

These results replicate and extend previous research demonstrating large intersubject variability in MGS (Benoit et al., 2010; Nath & Beauchamp, 2012). Although prior work has identified large individual variability in MGS, this study is the first to identify a relationship between MGS and lipreading skill. In addition, these results establish that individual differences in identification and detection are only moderately correlated, which suggests that individual differences in perceiving McGurk fusions does not entirely depend on failing to detect AV incongruity (see also Soto-Faraco & Alsius, 2007).

Table 1. Correlations among measures of susceptibility to the McGurk effect (MGS) and lipreading.

	MGS-*d′*	MGS-detect	LR-C	LR-POA	LR-W
MGS-ID	−.26*	.41**	.14	.32**	.15
MGS-*d′*		−.81**	.42**	.31**	.45**
MGS-detect			−.30**	−.19	−.29*
LR-C				.71**	.63**
LR-POA					.65**

*p < .05. **p < .001.

LR-C = lipreading consonants; LRC-POA = lipreading correct place of articulation in consonant task; LR-W = lipreading words

Although models of AV integration (Braida, 1991; Grant & Seitz, 1998; Massaro & Cohen, 2000) would predict that poorer lipreaders should perceive fewer AV fusions, this is the first empirical demonstration of that relationship. Good lipreading ability was associated with greater susceptibility in the identification task but poorer susceptibility in the detection task, shown by higher d' scores and lower false alarm rates. Therefore, good lipreaders are more likely to perceive fusions but also more likely to notice incongruity when it occurs. This finding is somewhat surprising as it means better lipreaders are both more and less susceptible to the McGurk effect, depending on how MGS is measured. This may suggest that task demands influence the strategies that good lipreaders are using. When the task is comprehension, better lipreaders use the information they extract from the visual signal to supplement the auditory signal. When the task is incongruity detection, extracting more information from the visual signal allows better evaluation of whether the visual and auditory signals are consistent. One reason that the relationship between MGS and lipreading may have emerged here and not in prior work (Cienkowski & Carney, 2002) is the measures of lipreading used. The POA measurement in the lipreading task allows for a much more fine-grained analysis of the information the participant was able to extract from the visual signal. Therefore, it may be more sensitive to individual differences in lipreading skill than sentence-length measures, which also include semantic and grammatical cues.

Although this study demonstrates that MGS-ID and lipreading are not wholly independent, it is important to note that the relationship is rather weak. This may suggest that individual differences in MGS depend in part on unimodal extraction ability but also on other individual difference variables, which may include integration skill, although the current study cannot evaluate this claim. The links between MGS and lipreading also suggest that future studies seeking to quantify integration ability should measure lipreading skill to ensure that measures of MGS are separate from measures of visual-only identification.

The direct comparison of measures of MGS in identification and detection tasks revealed that individual variability in the two tasks is moderately correlated. This provides some support for the claim that fusions are less likely to occur when individuals notice incongruity. However, given the only modest relationship between MGS-ID and detection measures, it is clear that detecting incongruity does not preclude perceiving an AV fusion. The strong test–retest correlations of MGS-ID suggest good reliability of measuring MGS, so the relatively weaker correlations between MGS-ID and measures of detection suggest unresolved individual differences underlying completion of the two tasks. Because participants were not explicitly alerted to the fact that incongruity may occur in the identification task but were in the detection task, it is possible that task demands were influencing the differences. Until the explanations for task differences become clearer, future studies should use caution when making generalizations about MGS from studies that measure MGS with identification versus detection paradigms.

Although the detection task in the current study asked participants to determine whether the auditory and visual speech tokens matched, the results are reminiscent of studies that ask participants to determine whether the auditory and visual speech tokens are aligned in time. These studies have found that relatively large temporal asynchronies in the auditory and visual stimuli of congruent speech (Dixon & Spitz, 1980; Massaro, Cohen, & Smeele, 1996; McGrath & Summerfield, 1985) and McGurk stimuli (Kösem & van Wassenhove, 2012; Munhall et al., 1996) may still be perceived as synchronous, and McGurk fusions may occur even when the auditory and visual signals are significantly misaligned (Munhall et al., 1996). Furthermore, perceiving that auditory and visual stimuli are temporally congruent does not guarantee that a McGurk fusion will occur (van Wassenhove, Grant, & Poeppel, 2007). Taken together, the temporal alignment findings and the current research suggest that detecting incongruity in auditory and visual signals does not preclude McGurk fusions from occurring.

The correlation between MGS-d' and MGS-ID was smaller than the correlation between MGS-detect and MGS-ID. One explanation for this may be a computational limitation of the d' value. The maximum d' score obtained in the study was 3.45, which indicated that participants correctly identified every congruent stimulus as congruent and correctly identified all but four McGurk stimuli as incongruent. For all seven subjects who showed a d' score of 3.45, these four false alarms were to the stimuli $A_{ba}V_{fa} = AV_{va}$ (two produced by each speaker). Other participants did correctly note that those stimuli were incongruent, but they also had misses (indicating congruent stimuli were incongruent). Given that this created a ceiling effect at 3.45, the smaller relationship between MGS-d' and MGS-ID than between MGS-detect and MGS-ID might simply be attributable to a restriction of range. Measures of d' are most useful when participants show variability in responding to both congruent and incongruent tasks. In this case, however, there was relatively little variability in hit rates; that is, most participants correctly identified all congruent trials as congruent. Therefore, MGS-detect may provide a more accurate assessment of the relationship between detection and identification tasks than MGS-d'.

Of particular note was the finding that the relationship between MGS-ID and lipreading was strengthened when the detection measures were controlled for. This indicates that participants who showed the highest MGS-ID scores tended to be good lipreaders who were also less likely to detect incongruity in the detection task. The finding is particularly surprising because lipreading was correlated with the detection measures, with better lipreaders showing better incongruity detection. This finding may suggest that MGS-ID depends in part on the ability to extract relevant information from the visual signal and also on failing to notice the incongruity in auditory and visual signals.

Approximately a third of the variability in MGS-ID was accounted for by measures of MGS-detect and lipreading. Lipreading skill has been linked to cognitive abilities, such as working memory and processing speed (Feld & Sommers, 2009), but it is not clear whether these measures or other cognitive tasks also predict MGS. Indeed, the only behavioral tasks that have been demonstrated to correlate with MGS are other measures of AV integration (Tremblay et al., 2007) or speech perception (Grant & Seitz, 1998). Although the current study suggests that some of the variability in MGS is attributable to lipreading skill, explanations for the remaining variability in MGS remain unknown. Future work should explore the cognitive or perceptual mechanisms underlying the remaining individual variability in MGS.

Quantifying the ability to integrate auditory and visual information has both theoretical and clinical implications. Given that individual differences in MGS persist even after controlling for individual differences in lipreading ability, models of AV speech perception may be improved by including a mechanism that treats integration as an individual difference variable. Individual differences in integration ability also have clinical implications. Individuals who differ in the extent to which they integrate AV information may benefit from different methods of training and education following hearing loss. Future work should also address whether the extent to which individuals are able to integrate audiovisually may be improved with training or practice.

Acknowledgments

This work was supported in part by a grant from Howard Hughes Medical Institute to Carleton College Interdisciplinary Science and Math Initiative (Grant 52006286). We are grateful to Paul Mains and Harper Makowsky for assistance with data collection and to Sarah Meerts, Julie Neiworth, Joseph Slote, and two anonymous reviewers for helpful feedback on earlier versions.

References

Aruffo, C., & Shore, D. I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. *Psychonomic Bulletin & Review, 19,* 66–72. doi:10.3758/s13423-011-0176-8

Auer, E. T., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research, 50,* 1157–1165. doi:10.1044/1092-4388(2007/080)

Benoit, M. M., Raij, T., Lin, F.-H., Jääskeläinen, I. P., & Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping, 31,* 526–538. doi:10.1002/hbm.20884

Bentler, R. (2000). List equivalency and test-retest reliability of the Speech in Noise Test. *American Journal of Audiology, 9,* 84–100. doi:10.1044/1059-0889(2000/010)

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44,* 5–18. doi:10.1016/j.specom.2004.10.011

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62,* 233–252. doi:10.3758/bf03205546

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research, 17,* 619–630. doi:10.3758/bf03211678

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 43*(A), 647–677. doi:10.1080/14640749108400991

Brown, G. S., & White, K. G. (2005). The optimal correction for estimating extreme discriminability. *Behavior Research Methods, 37,* 436–449. doi:10.3758/bf03192712

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-visual speech perception and aging. *Ear and Hearing, 23,* 439–449. doi:10.1097/00003446-200210000-00006

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology, 113,* 495–506. doi:10.1016/S1388-2457(02)00024-X

Conrey, B., & Gold, J. M. (2006). An ideal observer analysis of variability in visual-only speech. *Vision Research, 46,* 3243–3258. doi:10.1016/j.visres.2006.03.020

Demorest, M. E., Bernstein, L. E., & DeHaven, G. P. (1996). Generalizability of speechreading performance on nonsense syllables, words, and sentences: Subjects with normal hearing. *Journal of Speech and Hearing Research, 39,* 697–713. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8844551

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9,* 719–721. doi:10.1068/p090719

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics, 32,* 562–570. doi:10.3758/BF03204211

Feld, J., & Sommers, M. S. (2009). Lipreading, processing speed, and working memory in younger and older adults. *Journal of Speech, Language, and Hearing Research, 52,* 1555–1565. doi:10.1044/1092-4388(2009/08-0137)

Feld, J., & Sommers, M. S. (2011). There goes the neighborhood: Lipreading and the structure of the mental lexicon. *Speech Communication, 53,* 220–228. doi:10.1016/j.specom.2010.09.003

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104,* 2438–2450. doi:10.1121/1.423751

Green, K. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 3–26). East Sussex, UK: Psychology Press.

Green, K., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50,* 524–536. doi:10.3758/BF03207536

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers, 27,* 46–51. doi:10.3758/BF03203619

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Ketterman, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87,* 47–57. doi:10.1016/S0

Jackson, P. L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90,* 99–115.

Jin, S., & Carney, A. E. (2000). Sources of variability for talker and listener effects in auditory-visual speech perception. *The Journal of the Acoustical Society of America, 107,* 2888. doi:10.1121/1.428736

Kaplan-Neeman, R., Kishon-Rabin, L., Henkin, Y., & Muchnik, C. (2006). Identification of syllables in noise: Electrophysiological and behavioral correlates. *The Journal of the Acoustical Society of America, 120,* 926. doi:10.1121/1.2217567

Kösem, A., & van Wassenhove, V. (2012). Temporal structure in audiovisual sensory selection. *PloS One, 7,* e40936. doi:10.1371/journal.pone.0040936

Kricos, P. (1985). Differences in visual intelligibility across talkers. *The Volta Review, 8,* 5–14. doi:10.1007/978-3-662-13015-5_4

Lyxell, B., & Holmberg, I. (2000). Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11–14 years). *British Journal of Educational Psychology, 70,* 505–518. doi:10.1348/000709900158272

Macleod, A., & Summerfield, Q. A. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24,* 29–43. doi:10.3109/03005369009077840

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. East Sussex, UK: Psychology Press.

Massaro, D. W. (1987). *Speech perception by ear and eye*. Hillsdale, NJ: Erlbaum.

Massaro, D. W., & Cohen, M. M. (2000). Tests of auditory–visual integration efficiency within the framework of the fuzzy logical model of perception. *The Journal of the Acoustical Society of America, 108,* 784. doi:10.1121/1.429611

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America, 100,* 1777–1786. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8817903

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. doi:10.1121/1.392336

McGurk, H., & MacDonald, J. (1976, December 23). Hearing lips and seeing voices. *Nature, 264,* 746–748. doi:10.1038/264746a0

Middelweerd, M. J., & Plomp, R. (1987). The effect of speech-reading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America, 82,* 2145–2147. doi:10.1121/1.395659

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27,* 338. doi:10.1121/1.1907526

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 25,* 5884–5893. doi:10.1523/JNEUROSCI.0896-05.2005

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58,* 351–362. doi:10.3758/BF03206811

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

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research, 28,* 381–393. Retrieved from http://jslhr.highwire.org/cgi/content/abstract/28/3/381

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 lipreading* (pp. 97–113). Hillsdale, NJ: Erlbaum.

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,* 318–331. doi:10.1037//0096-1523.22.2.318

Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In B. Burnham, D. Campbell, & R. Dodd (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 85–108). East Sussex, UK: Psychology Press.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26,* 263–275. doi:10.1097/00003446-200506000-00003

Soto-Faraco, S., & Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *NeuroReport, 18,* 347–350. doi:10.1097/WNR.0b013e32801776f9

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: Evidence from the speeded classification task. *Cognition, 92,* B13–B23. doi:10.1016/j.cognition.2003.10.005

Spence, C., Ranson, J., & Driver, J. (2000). Cross-modal selective attention: On the difficulty of ignoring sounds at the locus of visual attention. *Perception & Psychophysics, 62,* 410–424. doi:10.3758/BF03205560

Strand, J. F., & Sommers, M. S. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *The Journal of the Acoustical Society of America, 130,* 1663. doi:10.1121/1.3613930

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26,* 212. doi:10.1121/1.1907309

Summerfield, Q. A., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 36*(A), 51–74. doi:10.1080/14640748408401503

Tremblay, C., Champoux, F., Bacon, B. A., & Theoret, H. (2007). Evidence for a generic process underlying multisensory integration. *The Open Behavioral Science Journal, 1*(1), 1–4. doi:10.2174/1874230007010011000

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). The effects of age and gender on lipreading abilities. *Journal of the American Academy of Audiology, 18,* 883–892. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18496997

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45,* 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69,* 744–756. doi:10.3758/BF03193776

Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *The Journal of the Acoustical Society of America, 100,* 1153–1162. doi:10.1121/1.416300

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667. doi:10.1037//0033-2909.88.3.638