SELF-FACE AND SELF-VOICE IN AUDIOVISUAL SPEECH INTEGRATION

CAN YOU MCGURK YOURSELF?

SELF-FACE AND SELF-VOICE IN AUDIOVISUAL SPEECH INTEGRATION

by

CHRISTOPHER ARUFFO, B.A., M.B.A., M.F.A.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2009)
(Psychology)

McMaster University
Hamilton, Ontario

TITLE: Can you McGurk yourself? Self-face and self-voice in audiovisual speech

AUTHOR: Christopher Aruffo, B.A. (Boston University), M.B.A. (University of California, Irvine), M.F.A. (University of Florida)

SUPERVISOR: Dr. David I. Shore

NUMBER OF PAGES: v, 55

ABSTRACT

This experiment used the McGurk effect to test the influence of identity on audiovisual speech. Participants recorded disyllables which were used to create McGurk stimuli. These stimuli were further manipulated so that the facial and vocal identities in each were either from the same speaker (matched) or from two different speakers (mismatched). When identities matched, self-produced speech was less susceptible to the McGurk effect. When identities were mismatched, participants were less susceptible to the McGurk effect if hearing their own voice, but were not affected by seeing their own face. These results suggest that vocal identity influences speech processing and that facial identity is processed independently of speech.

ACKNOWLEDGEMENTS

I thank my supervisor, Dr. David I. Shore, for his constant encouragement and enthusiasm as well as his efforts to shape my aesthetic curiosities into rational scientific inquiry. I wish to thank Bruce Milliken for essential perspective on the nature of graduate study, and Lee Brooks for helping me understand the critical difference between a prediction and an expectation. I thank my committee members, Karin Humphreys and David Feinberg, for their patience and consideration as I have grown into my role. I also thank all the members of the Multisensory Lab for their thoughtfulness and assistance. Finally I wish to thank my family, whose staunch and unwavering belief in my abilities underscores my ongoing dedication to success.

TABLE OF CONTENTS

INTRODUCTION

The experiment described here used the McGurk effect to measure the influence of identity recognition on audiovisual speech. The McGurk effect is an auditory illusion created by dubbing an auditory phoneme onto a discrepant visual phoneme, or *viseme* (McGurk & MacDonald 1976). A listener presented with the combined audiovisual stimulus often reports hearing an integrated sound, which may be one of two types (MacDonald & McGurk 1978): A *blend* effect occurs when a bilabial auditory phoneme is dubbed onto a non-labial viseme, causing a listener to hear an entirely different third phone. For example, auditory /ma/ dubbed onto visual /ka/ may be heard as /na/. A *combination* effect occurs when a non-labial auditory phoneme is dubbed onto a bilabial viseme, causing a listener to hear both sounds. For example, auditory /ga/ dubbed onto visual /ba/ may be heard as /bga/. Because these illusions only occur through successful integration of the auditory and visual modalities, the McGurk effect may be used to examine factors that influence the integration of audiovisual speech.

Audiovisual speech integration

Audiovisual speech integration, as indexed by the McGurk effect, is thought to be automatic, mandatory, and uninfluenced by attention. The McGurk effect's illusory percept remains impenetrable despite full knowledge of its nature (McGurk & MacDonald 1976) and irrespective of the sensory modality to which observers are instructed to attend (Massaro 1984). Moreover, the illusion can be supported by minimal information, such as a point-light representation of mouth and jaw motions with no facial features visibly present (Rosenblum & Saldaña 1996). Asynchrony between video and audio does not prevent integration, whether in time (Munhall et al 1996), space (Paré et al 2003), or both (Jones & Jarick 2006). Integration persists when observers can consciously detect that there are two disparate channels of information (Soto-Faraco & Alsius 2007); observers may clearly notice that visual and auditory channels are asynchronous (Soto-Faraco & Alsius 2009) and correctly identify which sensory modality leads (Vatakis & Spence 2007), recognize that the perceived phoneme does not match the apparent viseme (Summerfield & McGrath 1984), or notice that a face and voice are of different gender (Green et al 1991), without detriment to the illusory percept. The evidence combines to describe the McGurk illusion as "the result of an automatic process [i.e., audiovisual speech integration] that cannot be voluntarily prevented by the perceiver" (Navarra et al 2009).

Audiovisual speech integration can be prevented by reducing the quality of information presented in either sensory modality. This result is consistent with an optimal integration strategy (Ernst & Bülthoff 2002). When either the auditory or visual signal is of greater clarity, "the least ambiguous source of information has the most influence" on an integrated percept (Massaro 1998). Introducing ambiguity to either the visual or auditory channel will reduce that channel's contribution (Fixmer & Hawkins 1998; Colin et al 2002; Andersen et al 2009). Not all consonants contribute equally to the illusion (Cathiard, Schwartz, & Abry 2001), and some vowel contexts reduce the likelihood of consonant blending (Green, Kuhl, & Meltzoff 1988). When either modality is systematically degraded with noise, integration is biased proportionally to the other modality, and the subsequent likelihood of integration can be predicted mathematically (Massaro 2004).

Disunity of sources may prevent integration if sufficiently weighted. Integration is unlikely if either modality is recognized to be nonhuman or to be nonspeech, even when both sources are clearly perceived (Vatakis, Ghazanfar, & Spence 2008). The McGurk illusion is susceptible to such a disunity. Munhall et al (2009) created an animation of Rubin's face-vase profile illusion (Rubin 1915), manipulating the display so that either the vase or face was more prominent while the McGurk illusion sounded. When participants recognized a human silhouette producing the lip movement, the illusion was perceived; when the central figure appeared to be nonhuman (a vase), the same "lip" movement supported significantly fewer illusions. The lack of illusory percept from the face-vase presentation is therefore due to observers' conscious identification of the moving figure as a nonhuman object. A comparison of this result to Green et al (1991) suggests that the critical factor is an explicit identification. When an unknown face and voice are recognized to be of different gender, this cue may not be weighted strongly enough to induce disunity; perhaps an observer unconsciously accepts that an unknown speaker can speak with an unusual voice. By contrast, when visual and auditory sources are uniquely identified and thus definitively known to be separate and unrelated, this conscious knowledge provides adequate weight to prevent audiovisual integration of the two sources.

Identity disunity has been largely overlooked as a factor in audiovisual speech integration. To date, only one study has been published that purports to examine the influence of recognizable speakers on audiovisual speech (Walker, Bruce, & O'Malley 1995). In that study, four speakers were presented; an experimental group was familiar with all four while a control group was familiar with none. Each participant viewed the stimuli and reported what they heard. The number of illusions reported did not differ between familiar and unfamiliar groups— except when a speaker's voice was dubbed onto a different speaker's face. When faces and voices did not match, participants familiar with the speakers reported significantly fewer illusions. Familiar identity was thus demonstrated to have some effect on audiovisual speech, but left undetermined was whether either facial or vocal identity was primarily responsible for preventing the illusion. Because the experimental group was familiar with all four speakers, all faces and all voices were familiar, regardless of identity congruence, and the experimental design disallowed separate analysis of either.

The perceptual disunity observed for familiar speakers in the McGurk effect appears to be caused by facial or vocal weighting. Participants who recognized familiar faces could not specifically identify which voice they had heard, even though all voices were familiar. Failure to identify a familiar voice from a short utterance is not surprising (Bricker & Pruzansky 1966), but eliminates the explanation that participants were able to distinctly identify face and voice as belonging to different individuals. The illusion would therefore be prevented by optimal weighting. Participants reported the auditory phoneme, but Walker et al dismissed the possibility of auditory influence. Because faces could be recognized, Walker et al asserted that weighting occurs in the visual channel due to a previously-unobserved dependence between facial identity and facial speech, such that "when subjects were processing facial speech cues from familiar faces, they were able to use their knowledge concerning those particular faces" to prevent the illusion, but declined to elaborate on what that knowledge might be or how participants might be using it.

Facial identity recognition

Facial identity recognition and facial speech are drawn from different types of facial information (Bruce & Young 1986). Identity is derived from the invariant characteristics of a face and their relation to each other, or *structural information.* Speech and emotion are derived from the transitory appearances caused by muscular action, or *affective information.* Although judgments of structural or affective information may be made separately— it is possible to comprehend speech produced by an unknown face, or to identify an individual when they are not speaking— to what extent these processes may be integrated is the subject of debate.

The prevailing theory of facial processing argues for separate coding to process each type of information. Separate coding has been demonstrated by prosopagnosic individuals, who show impaired ability to make structural judgment while showing no difficulty with affective judgments, or vice versa (Campbell, Landis, & Regard 1986). This dissociation suggests a separate functionality for each type. A separate functionality, however, would predict that structural and affective information have no influence upon each other, and experiments testing this prediction have been equivocal and conflicting. There is greater support for the assertion that the dissociation may not arise from functional independence, but from cognitive analysis of a single coding process (Calder & Young 2005).

The anatomical distribution of this coding process has been modeled by Haxby, Hoffman, & Gobbini (2000). In their model, initial facial perception is attributed to a single area— the inferior occipital gyri— and different visual areas perform subsequent analyses on affective or visual information. An extended system performs further refinements on these two types of information, determining identity, emotion, speech, spatially-directed attention, and other particulars. These areas are not necessarily compartmentalized, however, but appear to work in concert with each other.

The extent to which affective or structural information may be separated from each other remains undetermined. Each can be familiarized when the other is eliminated, but when both are present they are difficult to tease apart. For example, a nonstandard facial expression interferes with classification of a familiar but not an unfamiliar face (Kaufman & Schweinberger 2004). To explain this result, it is equally plausible to suggest either that a familiar structure automatically incorporates its affective aspect, causing it to differ from a remembered image, or that recognition is impaired by conflict between separate judgments of familiar-structure and unfamiliar-affective. A similar confound exists when observing that familiarity improves judgment of dynamic visual speech (Lander & Davies 2008); it is equally plausible to suggest either that familiar facial structure automatically guides comprehension of visual speech or that visual speech patterns have been familiarized independently of facial structure. In short, an apparent interdependence may be caused either by integration of affective information into a familiar structure or to dual familiarities which have arisen in parallel.

Structural information can prevent audiovisual integration without recourse to disunity. The McGurk effect may be reduced by manipulating facial structure, through inversion (Green 1994; Rosenblum, Yakel, & Green 2000), rotation (Jordan & Bevan 1997), translation into moving dots (Rosenblum & Saldana 1996), or rearrangement (Heitanen, Manninen, Sams, & Surakka 2001). However, movements outside of the oral area may be perceived as visual speech components (Thomas & Jordan 2004); therefore, any reduction of the McGurk effect caused by facial manipulations could be interpreted as a change in intelligibility rather than recognizability (Massaro & Cohen 1996), and the result is predicted by optimality rather than identity. This leaves an open question whether familiar facial recognition facilitates audiovisual speech.

Facial speech can cue disunity if sufficiently unambiguous. Unambiguous affective information either increases the likelihood of integration or overrides the auditory channel to produce a visual-only response (Massaro 2004). The opposite occurred with observers of familiar McGurk stimuli, who not only reported fewer illusions but consistently reported the auditory channel alone (Walker et al 1995), but this is not necessarily a contradiction. A familiar face might create such clarity in the visual modality that visual speech is made incompatible with a conflicting auditory signal, enabling participants to consciously select auditory information thus made separate. Although this scenario is speculative it may be comparable to the "dubbed movie effect," in which bilingual observers are able to attend to the auditory signal alone when conflicting audio and video are each presented with full clarity (Navarra et al 2009).

Self-produced speech may provide evidence for separate influence of structural and affective facial information on audiovisual speech. One's own face is familiar (Kircher et al 2001), but one's own visual speech is not likely to be familiar, as one does not habitually watch oneself speak. If structural familiarity does automatically facilitate visual speech processing, then a difference would be predicted for self-face McGurk stimuli. If visual speech is independently familiarized, then no difference would be observed between self-face and nonself-face McGurk stimuli.

Vocal identity recognition

If a familiar voice can affect audiovisual speech, traditional theories of voice perception do not adequately explain how. Auditory speech has, historically, been viewed mainly as a carrier of language, with vocal perception divided between *linguistic* and *extralinguistic* information (Schweda-Nicholson 1987). In this view, linguistic information is speaker-independent and abstract, comprehended by "normalizing" and ignoring a speaker's idiosyncratic vocal habits (Goldinger 1996; Pisoni 1992). Extralinguistic information is speaker-specific and exclusive of language-related production; judgments of identity or emotion occur when attention to language articulation can be minimized (Deffenbacher et al 1989). Examinations of vocal identity have historically been concerned with determining the conditions under which vocal identity judgment can be deemed reliable (R.D.G. 1923; McGeehee 1937) and the nonlinguistic variables which contribute to identity judgment (Yarmey 1995), explicitly disallowing any interaction between linguistic and extralinguistic information (Clifford 1980). Although judgments of vocal quality may indeed be performed separately from interpretation of linguistic content (Relander & Rämä 2009), only a minute fraction of the acoustic signal is necessary for linguistic comprehension (Remez 1981), which would render nearly the entire vocal stream "extralinguistic" and irrelevant to audiovisual speech.

An alternative view of vocal recognition mirrors the Bruce & Young (1980) model of facial recognition. Although the vocal apparatus is perceived only implicitly, its "appearance" being derived entirely from the sounds produced by its motion, it is nonetheless possessed of unchanging features and transitory actions whose production can be described, respectively, as structural and affective information (Belin, Fecteau, & Bédard 2004), processed in parallel with traditional "linguistic" information (Knösche et al 2002). While "[t]raditional accounts of talker recognition propose that an individual talker is identified by virtue of qualitative characteristics that are nondistinctive linguistically" (Sheffert et al 2002), a recognition model makes no such exclusion; and, once a role for linguistic information is acknowledged in identity judgment, it becomes increasingly compelling to imagine the roles of affective and structural information in vocal analysis to be the reverse of facial analysis.

Structural information alone is an unreliable determinant of vocal identity. An unknown voice is difficult to remember and recognize if it is not sufficiently distinct (Schmidt-Neilsen & Stern 1986). Structural information may seem to indicate distinct physical characteristics of a speaker, such as age (Ptacek & Sander 1966), body size (Lass et al 1978), or sex (Lass et al 1976), to the extent that voices may be matched to static images of their speakers with accuracy above chance levels (Krauss, Feyberg, & Morsella 2002), but these indications may be indirect, as unguided judgments of physical characteristics based on vocal qualities can be entirely inaccurate (McGehee 1944). That is, vocal structure may not directly and unambiguously represent a speaker's physical characteristics to a perceiver, but instead evoke an abstract categorization which can be matched to laboratory stimuli (Kramer 1963).

Affective information contributes directly to voice recognition. A salient change in emotional affect can entirely obscure structural information, rendering a speaker unrecognizable (Saslove & Yarmey 1980); accented speech makes a speaker more difficult to recognize without obscuring structure (Goldstein et al 1981). Aphasics with left-hemisphere lesions find it difficult to recognize a familiar voice (Paelecke-Habermann et al 2009). Bricker & Pruzansky (1966) tested the effect of linguistic content on identification accuracy from short speech samples and observed: firstly, that speaker identifiability was dependent on vowel type; secondly, that reversing a sample significantly impaired recognition, despite vocal structure being preserved; thirdly, that identification accuracy was better determined not by a sample's overall duration but by the amount of phonetic information contained within it. Despite their findings, these authors suggested that their results "pose[d] some difficulties for a model of talker-identification behavior based on attributes of voice quality" and did not propose an alternative.

Recent evidence supports a stronger role for affective information in identity judgment. Listeners perform perceptual analysis upon a language stream to represent its originating movement and identify its unique characteristics. This claim is reminiscent of the motor theory of speech perception (Liberman & Mattingly 1985) but is language-neutral and amodal; "[f]rom this perspective, the physical movements of a speech gesture can shape the acoustic and optic signals in a similar way, so that the signals take on the same overall form" (Rosenblum 2008). Training with silent visual speech improves intelligibility of novel auditory utterances from the same speaker (Rosenblum, Miller, & Sanchez 2007), and a recent matching paradigm has shown that listeners can identify the speakers of auditory samples when selecting between dynamic silent faces which have been controlled for distinct physical characteristics. In this paradigm, identification is better than chance, and is speaker-specific rather than utterance-specific (Kamachi et al 2003). Identification remains above chance when structural information is obscured, either vocally using white noise (Lachs & Pisoni 2004) or visually using a point-light display (Rosenblum et al 2006). Accuracy falls to chance levels when affective information is obscured, either vocally with reversed speech samples or visually with static faces (Kamachi et al 2003). Lander et al (2007) found that matching was disrupted by altering the manner of speech— articulating with different levels of precision, or changing a statement into a question— but was not impaired by altering content, even into an unfamiliar language (e.g., English listeners analyzing Japanese speakers).

Identity judgments may be drawn from affective information by identifying a characteristic manner of articulation. Additional evidence has arisen from testing with *sine-wave speech*— a computerized transformation of a speech signal that "include[s] three or four time-varying sinusoids, each of which reproduces the center frequency and amplitude of a vocal resonance in a natural utterance" (Remez et al 1994). Although a transformed sample sounds highly artificial and nonhuman, observers who are instructed to interpret the sounds as speech find little difficulty in accurately doing so (Remez et al 1981). Despite the complete absence of structural information in sine-wave speech, familiar voices can be recognized without special training or practice (Remez, Fellowes, & Rubin 1997) and unfamiliar natural voices can be reliably matched to sine-wave transformations of those voices speaking different utterances (Fellowes, Remez, & Rubin 1997). Sheffert et al (2002) expanded these observations by training participants with samples of natural, reversed, or sine-wave speech; participants trained to recognize 10 different sine-wave speakers not only became able to do so but were also able to identify the natural speech of those same speakers. Participants trained with natural or reversed speech samples were able to make accurate identifications of other natural or reversed samples, but not of sine-wave speech; while the cause of their failure to recognize sine-wave speakers are speculative, their successes with reversed speech appears to demonstrate that structural vocal qualities may be recruited for identity judgment when affective information is ambiguous. Structural information may also be recruited once affective habits are well-learned; familiar voices are subjectively perceived as more distinctive, which suggests a sensitivity to structure (Schmidt-Neilsen & Stern 1984). Nonetheless, affective information seems to be the dominant determinant of vocal identity.

If vocal identity is drawn from affective information, it would be more likely to reduce the strength of the illusion. Audiovisual speech integration may be disrupted by disunity cues; if an observer recognizes their own voice paired with an unfamiliar face a dissociation may occur. However, if self-speech is processed with greater clarity than that of unknown speakers, then the illusion would become weaker in self-speech when identity is unified, and be made weaker still in mismatched-identity trials due to combined disunity and auditory weighting.

If vocal identity is drawn from structural information, self-voice would show little influence on the McGurk illusion. In stimuli featuring a face and voice from the same speaker, no disunity would exist, predicting no differences between self- and nonself-speech. Self-face and self-voice alone— mismatched to nonself-voice and nonself-face, respectively— would provide equivalent disunity, and disrupt the illusion equally, notwithstanding any additional effect of facial recognition.

Multimodal identity recognition

Vocal identity judgment may also incorporate affective visual information. Visual speech may either be an integral component of vocal identity or merely recruited when available to supplement ambiguous vocal information. Voices are learned more quickly when training is supplemented by visual speech (Sheffert & Olson 2004) and a familiar voice is more swiftly recognized when presented with a corresponding dynamic face (Schweinberger, Robertson, & Kaufmann 2007). An auditory sample of a familiar voice activates visual areas of the brain which an unfamiliar voice does not (von

Kriegstein & Giraud 2006; Rosa et al 2008) when recognizing identity and not language (von Kriegstein et al 2005); although it is not yet certain what a visual activation might represent, facial and vocal inputs do interact and integrate in identity judgment (Joassin et al 2004; Campanella & Belin 2007). If facial and vocal speech are integral, a greater bias toward a unified percept would be created, predicting a greater susceptibility to the McGurk illusion for self versus nonself speech in matched-identity stimuli.

An interaction of affective facial and vocal information calls into question the conclusion that a familiar voice does not influence the McGurk effect (Walker et al 1995). Such an interaction may indeed clarify the vague supposition that observers "use facial knowledge" to somehow pierce the illusion: if vocal identity can be recognized by characteristic articulation, and visual speech is integral to its characterization, then a conflict in manner between visual and auditory speech might be more easily detected when both are familiar, even if only one is identified by name. Indeed, explicit identification is not necessary to process a familiar voice advantageously (Hanley, Smith, & Hadfield 1998). Inability to identify a familiar voice may not be due to inaccurate perception, but inability to associate a voice with identifying personal traits (Hanley & Turner 2000); a recognized voice may be readily identified when facts associated with the speaker are offered as cues (Schweinberger & Herholz 1997). Even so, if there is an influence of familiar voice on the McGurk effect, it has not been tested.

It is difficult to test vocal familiarity exclusive of facial familiarity. Knowledge of a voice may increase auditory intelligibility (Craik & Kersner 1974), and a voice may be recognized as familiar as early as 200ms after onset (Beauchemin et al 2006), so a very short utterance such as a McGurk stimulus could become optimally weighted toward the auditory modality. For this to occur, however, a listener would have to be very familiar with the stimulus speaker. Nygaard & Pisoni (1998) used complete sentences to familiarize observers with unknown voices and discovered that listeners' knowledge did not transfer to novel single words. A listener would need significant exposure to a speaker to gain an advantage for short utterances, but in gaining such exposure a listener would gain equal familiarity with that speaker's visual speech.

Self-voice is an ideal stimulus for testing the influence of vocal familiarity on the McGurk effect. We are familiar with our own voices, but rarely watch ourselves speak. Although self-voice is heard differently due to its reaching the cochlea through bone conduction (Békésy 1949), recorded self-voice is recognized with near-perfect accuracy (Shuster & Durrant 2003; Kaplan et al 2008). If a familiar voice were to have an influence on the McGurk effect, it would likely be observed with self-voice as a reduction in the number of illusions reported. If vocal identity is multimodal, then the number of reported illusions for self stimuli would either be the same or more than for nonself.

Current experiment

The current experiment used self-produced speech to test whether familiar facial or vocal identity more strongly influences audiovisual integration. In addition to audiovisual *self* and *other*, trial blocks distinguished between self-face stimuli and self-voice stimuli that had been mismatched to nonself voices and faces, respectively. To maximize the influence of facial identity recognition, natural head motion was allowed (Knappmeyer, Thornton, & Bülthoff 2003). To maximize the influence of vocal identity recognition, disyllables such as /aga/ were used, as a speaker may be recognized from the initial vowel before a target phoneme (Beauchemin et al 2006).

Gender congruence was implemented as a between-subjects factor. Although gender incongruence is not expected to influence the McGurk illusion (Green et al 1991), even if speakers are familiar (Walker, Bruce, & O'Malley 1995), gender incongruence had not yet been tested with self-speech and was thus included in this experiment.

To avoid potential ceiling effects, an illusion of moderate strength was desirable. The vowel /a/ was selected as it demonstrates an effect stronger than /u/ but weaker than /i/ (Green 1988). Nonsense syllables were used to prevent lexical cuing. Seven non-discrepant syllables were included, for a total of nine different disyllables, because this set size diminishes the overall strength of the illusion (Amano & Sekiyama 1998).

EXPERIMENT 1

Method

Participants

A total of 11 McMaster University students, 5 male and 6 female, were recruited as participants, including 10 undergraduates and 1 graduate (age: 18-45 years). All participants reported normal hearing and normal or corrected-to-normal vision. Participants received course credit for their participation. Two of the participants had known each other as friends for four years, but no other participants had ever met or conversed with each other.

All participants gave informed consent to the procedures. All procedures were approved by the McMaster Board of Ethics.

Stimuli

Each participant recorded five repetitions of seven disyllables. The disyllables recorded were /aba/, /ada/, /aga/, /aka/, /ala/, /ama/, and /ana/. Participants were verbally instructed to articulate distinctly, to pronounce each vowel "a as in father, not uh as in about", to speak each repetition in a monotone or with falling pitch, and to wait at least five seconds between repetitions. The disyllables to be spoken were presented on 5" × 7" index cards held directly above the camera by the experimenter. To encourage distinct articulation, the index cards read "AH-BA", "AH-DA", "AH-GA", "AH-KA", "AH-LA", "AH-MA", and "AH-NA", all written in permanent black marker.

When recorded, participants were seated before a plain beige background in a sound-attenuated room. A digital video camera (JVC GZ-MG37U) and wireless lapel microphone (Shure PG185) were used. Participants affixed the microphone to their own clothing without assistance, and spoke the word "hello" to confirm a clear signal. To supplement the room's incandescent overhead light, two 60-watt lamps were placed at 45-degree angles to the participant's body. These lamps were situated approximately two feet away from the participant and on a vertical level with the participant's face, thus rendering the face clearly and fully visible. The video camera was mounted on a tripod approximately one meter away from the subject, also on a vertical level with the face. Video was shot in standard 4:3 aspect ratio and was framed to include the entire face. Video footage was recorded to the camera's internal hard drive in MPEG-2 format (720 × 480 pixels, 8.5 Mbps). The camera's audio footage was not used. Audio was digitally recorded from the microphone in lossless WAV format (16-bit depth, 44100 samples). Extraneous noises were muted, and the volume of each voice normalized, using Adobe Audition 3.

The WAV audio and MPEG-2 video of each session were combined and synchronized to within 6 ms accuracy. The resulting audiovisual streams were cut into stimulus segments of one disyllable each. Each stimulus was precisely four seconds long, displaying a dynamic image of a participant's face gazing silently at the camera for approximately three seconds before speaking one complete disyllable. Audiovisual segments were encoded in NTSC format, using the Windows Media Video 9 codec at 720 × 480 pixels with a variable average bitrate of 1.5 Mbps and the Windows Media

Audio 9.2 codec at 48kHz stereo sampling with variable average bitrate of 96 Kbps. All audiovisual editing was performed with Adobe Premiere CS4.

Experimental stimuli were created by combining video and audio of different disyllables. For each participant, auditory /aba/ was combined with visual /aga/, and auditory /ama/ combined with visual /aka/, making five unique repetitions of each discrepant disyllable. In each stimulus, the auditory consonant release was synchronized to the video within an accuracy of 12 ms, and the peak intensity of the initial vowel was synchronized within an accuracy of 16 ms.

Additional stimuli were created by mismatching facial and vocal identities. For each disyllable, each participant's face was combined with every other participant's voice. Audio and video were synchronized in the same manner and with the same level of accuracy.

A total of 1,584 unique stimuli were created. This included 495 matched-identity stimuli (11 participants × 9 disyllables × 5 repetitions of each disyllable) and 990 mismatched-identity stimuli (11 faces × 10 voices × 9 disyllables).

Design

Stimuli were organized into three blocks: *matched identity*, *mismatched self*, and *mismatched nonself*. All three blocks were presented to each participant twice, in random order. Each block consisted of 48 trials, with one stimulus per trial, for a grand total of 288 trials. A *matched identity* block comprised 24 self-speech and 24 nonself-speech stimuli. A *mismatched self* block comprised 24 self-face and 24 self-voice stimuli. A *mismatched nonself* block comprised 48 nonself-speech stimuli. Each set of 24 trials included 10 discrepant stimuli (5 trials × 2 disyllables /aba/-/aga/, /ama/-/aka/) and 14 non-discrepant stimuli (2 trials × 7 disyllables /aba/, /ada/, /aga/, /aka/, /ala/, /ama/, /ana/). Within each block, all stimuli were randomized, although each nonself face or voice was made to appear with equivalent frequency.

Gender congruence was implemented between-subjects. In any mismatched stimulus, the facial and vocal identities were either same-gendered (both male or both female) or cross-gendered (one male, one female). In mismatched blocks, 6 participants were presented with only same-gender stimuli and 5 participants were presented with only cross-gender stimuli.

For each of the two participants who knew each other, all stimuli featuring the other's face or voice were excluded from all blocks. Nonself faces and voices were randomly substituted as appropriate to each block, with a condition that the same facial or vocal identity would not be repeated.

Procedure

Participants were seated at a table in a sound-attenuated room. On this table was placed a laptop computer (Gateway W6501), running Windows Vista, featuring a 15" widescreen display at 1280 × 800 resolution. Display brightness was set to maximum. Participants were encouraged to adjust the screen to a comfortable viewing angle.

Stimuli were presented via a custom interface developed with Realbasic software. This interface did not completely obscure the computer desktop, but the desktop color was set to a solid dark blue, and the taskbar and desktop icons were hidden, leaving nothing to be seen on screen but the experimental interface. The interface consisted principally of a video area measuring 720 pixels in width and 420 pixels in height. Beneath the video area was a single button labeled "continue". Above the video area, a set of black dots indicated the number of blocks remaining, and a progress bar indicated the number of trials remaining in the current block. Participants were told that the dots and bar indicated how many trials remained in the experiment, but were not informed of the design.

Sounds were played through the laptop's built-in speakers. The laptop speakers were located directly below the display. Due to the relatively low power of the speakers, their volume was initially set to its maximum, and participants were played a single stimulus of their own face and voice speaking /ala/ to verify that the volume was neither inaudible nor uncomfortably loud.

To initiate the experiment and each subsequent trial, participants clicked "continue" using the laptop's built-in mouse button. All participants were informed that they could refrain from clicking "continue" at any time if they wished to take a break, although no participant did so in any session. During each stimulus presentation, the "continue" button was disabled and could not be clicked. When enabled, clicking "continue" caused the next stimulus to be presented without a delay.

Participants were instructed to repeat aloud whatever syllables they heard in each trial, and were informed that their looking away from the screen would produce invalid results. Participants were encouraged to respond to ambiguous stimuli "right away, with your first impression, rather than trying to second-guess yourself." The experimenter sat in the room with each participant, facing away from the video screen, and watched the participant's mouth as responses were made. If a spoken response was unclear or ambiguous, the experimenter prompted the participant to repeat their response. Each trial was recorded on paper as a single letter corresponding to the consonant spoken by the participant.

Results

Results were measured as the proportion of integrated responses to discrepant stimuli. A *non-integrated* response accurately reported only the auditory or visual channel (B, G, M, or K). Any other response was considered to be *integrated.* The proportions reported for both types of discrepant syllables were not significantly different (B-G, 65%; M-K, 56%) and these data were combined. Integrated responses produced by participants included D, F, L, N, T (/aða/), and A (/no consonant heard, reported as /a/-/a/). 98% of non-discrepant trials were reported correctly.

Gender congruence was not a significant factor. Fewer integrations were reported for cross-gender trials (57%) than for same-gender trials (63%), but this difference was not significant. A 2 × 2 repeated-measures ANOVA (self-face × gender congruence) was conducted on percentage of integrations, and this revealed no significant effect of gender nor a significant interaction.

Results for Experiment 1 are shown in Figure 1. In matched blocks, fewer integrations were reported for self trials (60%) than nonself trials (71%). In the mismatched condition, there were fewer integrations reported for self-face trials (60%) than nonself trials (64%), and fewer still reported for self-voice trials (50%).

Figure 1 about here

A series of paired-samples t-tests were conducted on these means. The difference between self-face and self-voice was significant ($t(10) = 2.74, p = .02$) but no other differences were found.

Discussion

The results appear to demand an increase in statistical power. Although there appear to be differences between self and nonself stimuli, only one significant effect was found: self-voice supported weaker audiovisual integration than self-face. However, neither self-face nor self-voice produced a level of integration significantly different from nonself stimuli in mismatched trials.

A second experiment was designed to clarify the comparison of self-face to self-voice. Stimuli were re-blocked to maximize the variety of face-voice combinations and increase the quantity of discrepant trials. Additional participants were recruited and, as the McGurk effect is thought to be unaffected by knowledge of the illusion (McGurk and MacDonald 1976), participants from Experiment 1 were invited to return.

The second experiment introduced a combination illusion. Discrepant stimuli in Experiment 1 were exclusively blend illusions. Combination responses have been offered as speculative evidence for the influence of familiar faces (Walker et al. 1995), so a discrepant combination was used in Experiment 2 to discover if a similar result would be seen. The combination composed of auditory G and visual B was selected because it could be created from the recordings already collected for Experiment 1.

To leave the set size unchanged, and reduce the overall quantity of stimuli to be created, the discrepant M-K blend was excluded from Experiment 2. Excluding /ama/ and /aka/ also necessitated the exclusion of /ana/. New participants were asked to record disyllables /aða/ and /abga/ instead.

EXPERIMENT 2

Method

Participants

8 new participants (1 male, 7 female) were recruited for this experiment. Recordings of all 11 participants from Experiment 1 were used with consent, and 8 participants (3 male, 5 female) returned to provide new data. In total, 19 participants were presented in stimuli, and 16 provided testing data. All participants were McMaster University students, with 2 graduates and 15 undergraduates participating (age: 18-45 years). 12 participants were native English speakers, two of whom were trilingual (one speaking Cantonese and Mandarin Chinese, the other Russian and Hebrew). Two participants spoke English as a second language, with their first languages Italian and Chinese, respectively. All participants reported normal or corrected-to-normal vision. All but one participant reported normal hearing. One participant self-identified as having impaired hearing; this participant's data were not used, although their recorded image was used with consent. Participants received course credit for their participation.

Stimuli

The recording method was identical to that of Experiment 1. Participants were informed that "AH-THA" was to be pronounced with a /ð/ sound, "soft like *this* or *that*; not hard, like *think* or *thin*," and that "OB-GA" was to be pronounced with an /a/ sound, "not like *obey*, but like *object*." Some participants mispronounced these disyllables and were asked to repeat the sounds correctly. The result of the recording sessions for

Experiment 2 were 30 disyllables from each newly-recruited participant, including five repetitions each of /aba/, /ada/, /aga/, /ala/, /aða/, and /abga/. Only newly-recruited participants recorded new footage; participants from Experiment 1 did not record /aða/ and /abga/. The recorded footage was processed into four-second stimuli in the same manner as Experiment 1.

A total of 2,928 unique stimuli were created. Only new participants had recorded /aða/ and /abga/; from these were created 80 matched-identity stimuli (8 participants × 2 disyllables × 5 repetitions of each disyllable) and 112 mismatched-identity stimuli (8 faces × 7 voices × 2 disyllables). For all participants, 570 matched-identity stimuli (19 participants × 6 disyllables × 5 repetitions of each disyllable) and 2,052 mismatched-identity stimuli (19 faces × 18 voices × 6 disyllables) were created. Stimuli were edited and encoded as in Experiment 1. No stimuli from Experiment 1 were re-used.

Design

Stimuli were blocked as in Experiment 1. Within a block, each set of 24 trials was changed to comprise 18 discrepant trials (9 trials × 2 disyllables) and 6 non-discrepant trials (1 trial × 6 disyllables). Stimuli were again randomized within each block with the restriction that no two identical stimuli were presented consecutively. Gender congruence was not a factor.

Because /aða/ and /abga/ had not been recorded by returning participants, non-discrepant stimuli featuring these disyllables featured only new participants. One of each disyllable were presented in the *self* blocks for returning participants and the two trials were coded as *nonself*.

Procedure

The procedure was identical to Experiment 1 except for the following differences.

Participants were first screened for familiarity. Silent video clips of each of the 19 faces, speaking /ala/, were played at full size (720 × 480 pixels). Participants were instructed to speak aloud "yes" or "no" to indicate their familiarity with each image, where "familiar" meant having had verbal interaction with the person displayed. Each "yes" answer was noted on the participant's datasheet, and the interface was adjusted to disallow the faces and voices of all familiar participants from the session. Six participants indicated familiarity with 2 or 3 other participants; one participant indicated familiarity with 5 others.

Additional instructions were provided to returning participants. Returning participants were reminded that they were now aware this experiment involved an auditory illusion. They were informed that "decoys" had been included (i.e., the non-discrepant stimuli) as well as new syllables, and were told "you may find yourself saying things you know you never recorded." Returning participants were further instructed that, because of the illusion, it was especially important to respond quickly to ambiguous stimuli to avoid the possibility of second-guessing their perception based on their knowledge of the illusion. Each participant was told "The 'correct' answer is not what you figure out it might have been, but what you actually heard."

Responses were coded as before. *T* was used to report /aða/ and *O* to report /abga/. Because the stop consonant in /abga/ may be non-plosive, a participant was considered to have replied /abga/ if they closed their lips during their response, and /aga/ if their lips remained open.

Results

Results were measured as in Experiment 1. Integrated responses produced by participants included D, F, L, M, T, V, Y, and A. 92% of non-discrepant trials were reported correctly. Combination trials (auditory G, visual B) were not affected by experimental variables. The following analysis is based on blend illusions only (auditory B, visual G).

Differences were found between self and nonself trials. In matched blocks, fewer integrations were reported for self trials (56%) than nonself trials (76%). In mismatched blocks, fewer integrations were reported for self-voice trials (48%) than nonself trials (64%). 98% of non-integrated responses reported the auditory phoneme. These results are shown in Figure 2.

Figure 2 about here

A series of paired-samples t-tests were conducted on these means. In matched blocks, the difference between self and non-self was significant ($t(14) = 3.35, p < .01$). In mismatched blocks, self-voice was different from both nonself ($t(14) = 2.14, p = .05$) and self-face ($t(14) = 2.14, p = .05$), while nonself and self-face were not significantly different from each other.

Nonself trials showed a reduction in the McGurk effect when identities were mismatched. A paired-sample t-test was conducted to compare nonself matched (76%) and mismatched (64%). This test showed a significant difference ($t(14) = 3.21, p < .01$).

Knowledge of the experimental design did not influence the results. There was no effect of practice. Returning participants reported 60% integration in Experiment 1 and 61% in Experiment 2; these means were not different.

Integrated responses to combination stimuli did not vary between self and nonself stimuli. Responses between participants did vary, and the experimenter observed possible causes by interviewing each participant.

Discussion

Self-produced auditory speech disrupts the McGurk illusion. Self stimuli supported significantly fewer illusions than nonself stimuli, and the illusion was disrupted by self-voice rather than self-face. Self and nonself faces supported an equal proportion of illusory percepts.

Mismatching nonself identities disrupted the McGurk illusion. This result is surprising, as the illusion has previously been shown not to be affected by the identity of unknown speakers. Participants' preference to report the auditory channel suggests that the illusion was not heard because visual information could be ignored.

Discrepant combination stimuli were unaffected by experimental factors. The critical factors in combination integration appeared to be expectation, i.e., whether a listener was linguistically predisposed to interpret a silent lip closure as a stop consonant and whether the participant knew "obga" was a valid response. These reasons are not directly supported by the data, but are inferred from participants' subjective reports.

GENERAL DISCUSSION

The current experiment examined the influence of self-speech on audiovisual speech, using the McGurk effect as an index. Self-speech was compared with nonself-speech, and the influences of self-face and self-voice were separately tested by mismatching vocal and facial identities. Self-speech was observed to weaken the McGurk effect, which weakening was caused by vocal rather than facial information.

Vocal processing

The results provides evidence for an influence of vocal identity on audiovisual speech. Self-speech supported a weaker illusion, indicating that greater weighting occurs in the auditory modality, which weighting is further observed in the weaker illusion produced by mismatched self-voice compared to self-face.

It may be argued that the proportion of incongruent stimuli (75% of all trials) could have created a response bias. Recalibration and adaptation have been shown to exist in audiovisual speech perception, such that responses following persistent exposure to intermodal conflict may become biased toward the less ambiguous component (Bertelson, Vroomen, & de Gelder 2003). However, while these perceptual influences do persist for a longer time than the intertrial interval occurring in the current experiment (Vroomen et al 2004), if such an influence were present here it would have been observed across all trials and all blocks— self and nonself, matched and mismatched.

Facial processing

The data argue for separate processing of facial recognition and facial speech. While familiarity does facilitate visual speech (Yakel et al 2000; Lander & Davies 2008), this may be attributed to familiarity with characteristic linguistic actions of a face, rather than structural knowledge, as no effect of facial self-recognition— neither facilitation nor interference— was automatically conferred to self-speech in this experiment.

Until recently, behavioral research explored the question of functional independence in facial recognition by using static photographs as stimuli (Roark et al 2003). Participants were asked either to categorize a single image or to match images to each other. Criteria for these judgments were either structural, analyzing identity (e.g., *familiar* or *unfamiliar*), or affective, analyzing either expression (e.g., *happy* or *angry*) or facial speech (e.g., *ee* or *oo*). Separability was tested by varying one of the types of information and measuring its influence upon the other. For example, faces may represent different emotional expressions while reaction times for identity judgments are measured. A change in reaction time signals an influence of one feature upon the other.

Evidence from static images showed an apparent dependence between structural and affective information which asymmetrically favored structural processing. Structural information interfered with affective judgments, but not vice versa (Kaufmann & Schweinberger 2005; Campbell, Brooks, & de Haan 1996); structural processing was unconsciously activated during affective tasks, but not vice versa (Campbell & de Haan 1998). The asymmetry may be explained by noting that "structural encoding… provide[s] information for the analysis of facial speech, and for the analysis of expression," (Bruce & Young 1986), which information may be obscured by violating structural coherence (Hietanen et al 2001) or facilitated by increasing familiarity (Schweinberger & Soukup 1998).

Structural familiarity does not provide a general facial processing advantage. Although familiar and unfamiliar faces are processed differently (Dubois et al 1999), identity judgment of a single familiar face can be facilitated or impaired by affective information (Kaufman & Schweinberger 2004), and familiarity confers no advantage to affective comparisons between faces (Young et al 1986). It may be that introducing familiarity causes interdependence between structural and affective aspects of a face, such that the two types of information may influence each other where previously they could not (Ganel & Goshen-Gottstein 2004; Levy & Bentin 2008).

If familiarity increases mutual influence between structural and affective information, this influence should be intensified by dynamic stimuli. Dynamic affective stimuli activate more regions of the brain than do static photographs, either from changing expression (LaBar et al 2003) or continuous speech. Dynamic speech, particularly, activates language-specific regions not activated by static images (Calvert & Campbell 2003). Whatever effects of familiarity might be observed in static photographs would, presumably, be more pronounced when observing dynamic stimuli.

Familiarity with dynamic faces produces some evidence for interdependence similar to that from static faces. Familiarity facilitates facial speech processing from a single speaker (Lander & Davies 2008); facial speech facilitates identification of a single familiar face (Lander & Chuang 2005); switching between familiar faces incurs a cost in affective processing (Yakel, Rosenblum, & Fortier 2000; Kaufmann & Schweinberger 2005).

However, dynamic speech also exhibits an independence not observable in static photographs. When dynamic, familiar faces can be recognized from affective information alone. Knappmeyer, Thornton, & Bülthoff (2003) obscured structural information by animating a single computer-generated face with the tracked speech movements of different actors; participants were familiarized with the actors via this single identical face, and participants subsequently became able to identify each of the actors from their speech movements. Rosenblum, Niehus, & Smith (2007) removed structural information entirely with a point-light display. The experimenters placed reflective dots at the points of articulation, but then spread other dots randomly across the rest of the face; participants viewing dynamic speech in these displays were able to identify their friends. The success of these judgments could be explained by the *supplemental information hypothesis* (Roark et al 2003): that judgments of familiar identity are supplemented by "identity-specific facial motion [which is] encoded in addition to the invariant structure of a face," but the presence of a known invariant structure is not necessary. A prosopagnosic individual may identify faces from their idiosyncratic motions and yet be unable to recognize the structure of those same faces (Steede, Tree, & Hole 2007). Normal participants can accurately match unknown speaker identities from dynamic speech in point-light images, even when the rest of the face is made invisible by placing reflective dots on the articulators alone (Rosenblum et al 2002).

In short, facial speech may be familiarized independently of facial identity. The equivalence of integration for both self-face and nonself-face support this conclusion. Facial self-speech shows no influence of familiarity on audiovisual integration.

Nonself stimuli

Nonself stimuli supported a weaker illusion when mismatched. This result is unexpected and unprecedented (Green et al 1991). Because all nonself speakers were unfamiliar to the participants, and all stimuli were synchronized to within 12ms, neither identity disunity, temporal disunity, nor vocal familiarity explains the result. The McGurk effect is also not influenced by varying speakers from trial to trial (Rosenblum & Yakel 2001), although if this were an influence it would be equally present in both matched and mismatched blocks. Returning participants demonstrated knowledge of the illusion to have no effect.

A possible explanation is a conflict of vocal prosody and visual movement. Participants were not encouraged to hold their heads rigidly during recording sessions and were not discouraged from speaking naturally. Head movement is correlated with vocal prosody, and naive participants are able to match voices to heads according to head movements (Munhall et al 2004). The presence of head movements, and natural differences in vocal style, may have caused a disunity of manner between auditory and visual inputs which prevented integration.

Alternatively, an expectation of disunity may have influenced the result. Fusion illusions can be modulated by expectation, being rendered unlikely when a stimulus transforms a word into a nonword (Brancazio 2004; Barutchu et al 2008) or more likely if an observer hearing the same stimulus expects a nonword (Windmann 2004), regardless of whether their expectation violates semantic context (Sams et al 1998). The variability of the evidence suggests that context and expectation are not external influences which allow or prevent the integration of audiovisual channels, but are themselves weighted components of speech information which integrate with sensory input into a final perceptual decision (Massaro 1989). Because these participants were fully aware that their own faces and voices were being mismatched to other identities, perhaps the expectation that other identities were also so mismatched was a strong enough influence to allow visual speech to be ignored, thus preventing integration.

CONCLUSION

The main finding of this study is that self-voice can prevent integration of audiovisual speech. Secondarily, it is observed that recognizing one's own face neither facilitates nor interferes with audiovisual speech. Finally, the data show an influence of mismatching facial and vocal identities which has not previously been seen in studies of the McGurk effect.

FIGURE CAPTIONS

Figure 1. Results of Experiment 1 for both identity conditions. Error bars represent within-subject standard error of mean.

Figure 2. Results of Experiment 2 for both identity conditions. Error bars represent standard error of mean.

Figure 1

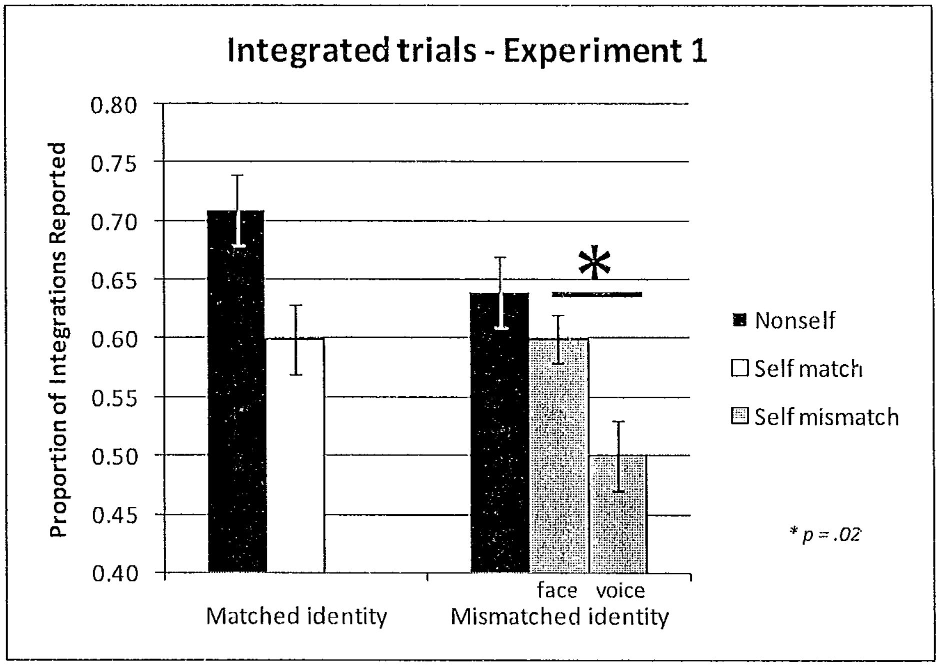
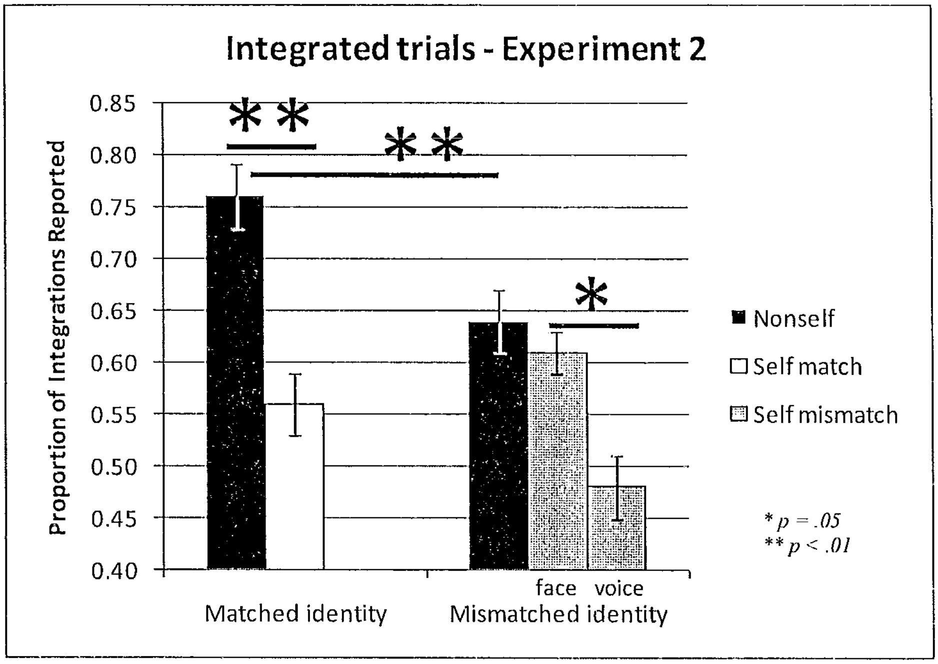

Integrated trials - Experiment 1
Proportion of Integrations Reported
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
*
face
voice
Matched identity
Mismatched identity
■ Nonself
□ Self match
▨ Self mismatch
* p = .02

Figure 2


Integrated trials - Experiment 2
Proportion of Integrations Reported
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
**
**
*
face
voice
Matched identity
Mismatched identity
■ Nonself
□ Self match
▨ Self mismatch
*p = .05
**p < .01

REFERENCES

Amano, J. and Sekiyama, K. (1998). The McGurk effect is influenced by the stimulus set size. In D. Burnham, J. Robert-Ribès, and E. Vatikiotis-Bateson (Eds.), *Proceedings of the Auditory-Visual Speech Processing Conference* (pp. 43–48). Terrigal, Australia.

Andersen, T., Tiippana, K., Laarni, J., Kojo, I., and Sams, M. (2009). The role of visual spatial attention in audiovisual speech perception. *Speech Communication, 51*, 184–193.

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

Beauchemin, M., DeBeaumont, L., Vannasing, P., Turcotte, A., Arcand, C., Belin, P., and Lassonde, M. (2006). Electrophysiological markers of voice familiarity. *European Journal of Neuroscience, 23*, 3081–3086.

Békésy, G. (1949). The structure of the middle ear and the hearing of one's own voice by bone conduction. *Journal of the Acoustical Society of America, 21(3)*, 217–232.

Belin, P., Fecteau, S., and Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences, 8(3)*, 129–135.

Bertelson, P., Vroomen, J., and de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science, 14(6)*, 592–597.

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

Bricker, P. and Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40(6)*, 1441–1449.

Bruce, V. and Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.

Calder, A. and Young, A. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6*, 641–651.

Calvert, G., and Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15(1)*, 57–70.

Campanella, S. and Belin, P. (2008). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11(12)*, 535–543.

Campbell, R., and de Haan, E. (1998). Repetition priming for face speech images: speech-reading primes face identification. *British Journal of Psychology, 89*, 309–323.

Campbell, R., Brooks, B., and de Haan, E. (1996). Dissociating face processing skills: decisions about lip-read speech, expression, and identity. *The Quarterly Journal of Experimental Psychology, 49A(2)*, 295–314.

Campbell, R., Landis, T., and Regard, M. (1986). Face recognition and lipreading: a neurological dissociation. *Brain. 109*, 509–521.

Cathiard, M., Schwartz, J., and Abry, C. (2001). Asking a Naive Question About the McGurk Effect: Why Does Audio [b] Give More [d] Percepts With Visual [g] Than With Visual [d]? *In In Auditory-Visual Speech Processing 2001, 138–142.*

Clifford, B. (1980). Voice identification by human listeners: on earwitness reliability. *Law and Human Behavior, 4(4)*, 373–394.

Colin, C., Radeau, M., Deltenre, P., Demolin, D., and Soquet, A. (2002). The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. *European Journal of Cognitive Psychology, 14(4)*, 475–491.

Craik, F. and Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology, 26(2)*, 274–284.

Deffenbacher, K., Cross, J., Handkins, R., Chance, J., Goldstein, A., Hammersley, R., and Read, D. (1989). Relevance of voice identification research to criteria for evaluating reliability of an identification. *Journal of Psychology, 123(2)*, 109–119.

Dubois, S., Rossion, B., Schiltz, C., Bodart, J., Michel, C., Bruyer, R., and Crommelinck, M. (1998). Effect of familiarity on the processing of human faces. *Neuroimage, 9*, 278–289.

Ernst, M. and Bülthoff, H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8(4),* 162–169.

Fellowes, J., Remez, R., and Rubin, P. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics, 59(6)*, 839–849.

Fixmer, E. and Hawkins, S. (1998). The influence of quality of information on the McGurk effect. *In Auditory-Visual Speech Processing '98*, 27–32.

Ganel, T. and Goshen-Gottstein, Y. (2004). Effects of familiarity on the perceptual integrality of the identity and expression of faces: the parallel-route hypothesis revisited. *Journal of Experimental Psychology: Human Perception and Performance, 30(3)*, 583–597.

Goldinger, S. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(5)*, 1166–1183.

Goldstein, A., Knight, P., Bailis, K., and Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society, 17(5)*, 217–220.

Green, K. (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.

Green, K. (1994). The influence of an inverted face on the McGurk effect. *Journal of the Acoustical Society of America, 95*, 3014.

Green, K., Kuhl, P., Meltzoff, A., and Stevens, E. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception and Psychophysics, 50(6)*, 524–536.

Hanley, J. and Turner, J. (2000). Why are familiar-only experiences more frequent for voices than for faces? *Quarterly Journal of Experimental Psychology, 53A(4)*, 1105–1116.

Hanley, J., Smith, S., and Hadfield, J. (1998). I recognize you but I can't place you: an investigation of familiar-only experiences during tests of face and voice recognition. *Quarterly Journal of Experimental Psychology, 51A(1)*, 179–195.

Haxby, J., Hoffman, E., and Gobbini, M. (2000). The distributed human neural system for face perception. *Trends In Cognitive Sciences, 4(6)*, 223–233.

Hietanen, J., Manninen, P., Sams, M., and Surakka, V. (2001). Does audiovisual speech perception use information about facial configuration? *European Journal of Cognitive Psychology, 13(3)*, 395–407.

Hill, H. and Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology, 11,* 880–885.

Joassin, F., Maurage, P., Bruyer, R., Crommelinck, M., and Campanella, S. (2004). When audition alters vision: an event-related potential study of the cross-modal interactions between faces and voices. *Neuroscience Letters, 369,* 132–137.

Jones, J. and Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time. *Experimental Brain Research, 174,* 588–594.

Jordan, T. and Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23(2)*, 388–403.

Kamachi, M., Hill, H., Lander, K., and Vatikiotis-Bateson, E. (2003). Putting the face to the voice: matching identity across modality. *Current Biology, 13*, 1709–1714.

Kaplan, J., Aziz-Zadeh, L., Uddin, L., and Iacoboni, M. (2008). The self across the senses: an fMRI study of self-face and self-voice recognition. *Social Cognitive and Affective Neuroscience, 3*, 218–223.

Kaufmann, J., and Schweinberger, S. (2004). Expression influences the recognition of familiar faces. *Perception, 33*, 399–408.

Kaufmann, J., and Schweinberger, S. (2005). Speaker variations influence speechreading speed for dynamic faces. *Perception, 34*, 595–610.

Kircher, T., Senior, C., Phillips, M., Rabe-Hesketh, S., Benson, P., Bullmore, E., Brammer, M., Simmons, A., Bartels, M., and David, A. (2001). Recognizing one's own face. *Cognition, 78*, B1–B15.

Knappmeyer, B., Thornton, I., and Bülthoff, H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43,* 1921–1936.

Knösche, T., Lattner, S., Maess, B., Schauer, M., and Friederici, A. (2002). Early parallel processing of auditory word and voice information. *Neuroimage, 17*, 1493–1503.

Kramer, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin, 60(4)*, 408–420.

Krauss, R., Freyberg, R., and Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618–625.

LaBar, K., Crupain, M., Voyvodic, J., and McCarthy, G. (2003). Dynamic perception of facial affect and identity in the human brain. *Cerebral Cortex, 13*, 1023–1033.

Lachs, L. and Pisoni, D. (2004). Crossmodal source identification in speech perception. *Ecological Psychology, 16(3)*, 159–187.

Lander, K. and Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition, 12(3)*, 429–442.

Lander, K. and Davies, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology, 61(7)*, 961–967.

Lander, K., Hill, H., Kamachi, M., and Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance, 33(4),* 905–914.

Lass, N., Beverly, A., Nicosia, D., and Simpson, L. (1978). An investigation by means of direct estimation of speaker height and weight identification. *Journal of Phonetics, 6*, 69–76.

Lass, N., Hughes, K., Bowyer, M., Waters, L., and Bourne, V. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America, 59(3)*, 675–678.

Levy, Y. and Bentin, S. (2008). Interactive processes in matching identity and expressions of unfamiliar faces: evidence for mutual facilitation effects. *Perception, 37*, 915–930.

Liberman, A. and Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

MacDonald, J. and McGurk, H. (1978). Visual influences on speech production processes. *Perception and Psychophysics, 24*, 253–257.

Massaro, D. (1984). Children's perception of visual and auditory speech. *Child Development, 55*, 1777–1788.

Massaro, D. (1989). Testing between the TRACE model and the fuzzy logical model of perception. *Cognitive Psychology, 21*, 398–421.

Massaro, D. (1998). Illusions and issues in bimodal speech perception. *Proceedings of Auditory Visual Speech Perception '98.* (pp. 21-26). Terrigal-Sydney Australia, December, 1998.

Massaro, D. (2004). From multisensory information to talking heads. In G. Calvert, C. Spence, and B. Stein (eds.), *The Handbook of Multisensory Processes* (pp. 153–176). Cambridge, MA: MIT Press.

Massaro, D. and Cohen, M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics, 58(7)*, 1047–1065.

McGehee, F. (1944). An experimental study of voice recognition. *Journal of General Psychology, 31,* 53–65.

McGrath, M. and Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America, 77(2)*, 678–685.

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

Munhall, K., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58(3)*, 351–362.

Munhall, K., Jones, J., Callan, D., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science, 15(2)*, 133–137.

Munhall, K., ten Hove, M., Brammer, M., and Paré, M. (2009). Audiovisual integration of speech in a bistable illusion. *Current Biology, 19*, 735–739.

Navarra, J., Alsius, A., Soto-Faraco, S., and Spence, C. (in press). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion (2009).*

Newman, R. and Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics, 35*, 85–103.

Nygaard, L. and Pisoni, D. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60(3)*, 355–376.

Nygaard, L., Sommers, M., and Pisoni, D. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.

Paelecke-Habermann, Y., Somborski, K., Paelecke, M., Knörgen, M., Kneidel, O., and Gaul, C. (2009). Recognizing people by their voices: an fMRI-study of healthy people and patients after stroke. *Clinical Neuropsychology, 120(1)*, e69.

Paré, M., Richler, R., ten Hove, M., and Munhall, K. (2003). Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. *Perception & Psychophysics, 65(4)*, 553–567.

Pisoni, D. (1992). Talker normalization in speech perception. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.) Speech Perception, Speech Production, and Linguistic Structure. Tokyo: OHM.

Ptacek, P. and Sander, E. (1966). Age recognition from voice. *Journal of Speech and Hearing Research, 9*, 273–277.

R. D. G., Jr. (1923). Admissability of telephone conversations in evidence. *Virginia Law Review, 9(6),* 446–451.

Relander, K. and Rämä, P. (2009). Separate neural processes for retrieval of voice identity and word content in working memory. *Brain Research, 1252,* 143–151.

Remez, R., Fellowes, J., and Rubin, P. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance. 23(3)*, 651–666.

Remez, R., Rubin, P., Berns, S., Pardo, J., and Lang, J. (1994). On the perceptual organization of speech. *Psychological Review, 101(1)*, 129–156.

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (1981). Speech perception without traditional speech cues. *Science, 212(4497)*, 947–950.

Roark, D., Barrett, S., Spence, M., Abdi, H., and O'Toole, A. (2003). Memory for moving faces: psychological and neural perspectives on the role of motion in face recognition. *Behavioral and Cognitive Neuroscience Reviews, 2(1)*, 15–46.

Rosa, C., Lassonde, M., Pinard, C., Keenan, J., and Belin, P. (2008). Investigations of hemispheric specialization of self-voice recognition. *Brain and Cognition, 68(2),* 204–214.

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

Rosenblum, L. and Saldaña, H. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22(2)*, 318–331.

Rosenblum, L., Miller, R., and Sanchez, K. (2007). Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. *Psychological Science, 18(5)*, 392–396.

Rosenblum, L., Niehus, R., and Smith, N. (2007). Look who's talking: recognizing friends from visible articulation. *Perception, 36*, 157–159.

Rosenblum, L., Smith, N., Nichols, S., Hale, S., and Lee, J. (2006). Hearing a face: cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics, 68(1)*, 84–93.

Rosenblum, L. and Yakel, D. (2001). The McGurk effect from single and mixed speaker stimuli. *Acoustic Research Letters Online, 2(2)*, 67–72.

Rosenblum, L., Yakel, D., and Green, K. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 26(3)*, 806–819.

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

Rubin, E. (1915). Synoplevde Figurer (Kopenhagen: Gyldendalske).

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

Saslove, H. and Yarmey, A. (1980). Long-term auditory memory: speaker identification. *Journal of Applied Psychology, 65(1)*, 111–116.

Schmidt-Neilsen, A. and Stern, K. (1985). Identification of known voices as a function of familiarity and narrow-band coding. *Journal of the Acoustical Society of America, 77(2)*, 658–663.

Schmidt-Neilsen, A. and Stern, K. (1986). Recognition of previously unfamiliar speakers as a function of narrow-band processing and speaker selection. *Journal of the Acoustical Society of America, 79(4)*, 1174–1177.

Schweda-Nicholson, N. (1987). Linguistic and extralinguistic aspects of simultaneous interpretation. *Applied Linguistics, 8(2)*, 194–205.

Schweinberger, S. and Herholz, A. (1997). Recognizing famous voices: influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research, 40(2)*, 453–463.

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

Schweinberger, S., Robertson, D., and Kaufmann, J. (2007). Hearing facial identities. *Quarterly Journal of Experimental Psychology, 60(10)*, 1446–1456.

Sheffert, S. and Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics, 66(2)*, 352–362.

Sheffert, S., Pisoni, D., Fellowes, J., and Remez, R. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance, 28(6)*, 1447–1469.

Shuster, L. and Durrant, J. (2003). Toward a better understanding of the perception of self-produced speech. *Journal of Communication Disorders, 36,* 1–11.

Soto-Faraco, S. and Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport, 18(4)*, 347–350.

Soto-Faraco, S. and Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance, 35(2)*, 580–587.

Steede, L., Tree, J., and Hole, G. (2007). I can't recognize your face but I can recognize its movement. *Cognitive Neuropsychology, 24(4)*, 451–466.

Thomas, S. and Jordan, T. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 30(5)*, 873–888.

Vatakis, A. and Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69(5)*, 744–756.

Vatakis, A., Ghazanfar, A., and Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8(9),* 1–11.

von Kriegstein, K. and Giraud, A. (2006). Implicit multisensory associations influence voice recognition. *Public Library of Science Biology, 4(10)*, e326.

von Kriegstein, K., Eger, E., Kleinschmidt, A., and Giraud, A. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*, 48–55.

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

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., and Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication, 44*, 55–61.

Walker, S., Bruce, V., and O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception and Psychophysics, 57(8),* 1124–1133.

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

Yakel, D., Rosenblum, L., and Fortier, M. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics, 62(7)*, 1405–1412.

Yarmey, A. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law, 1(4)*, 792–816.

Young, A., McWeeny, K., Hay, D., and Ellis, A. (1986). Matching familiar and unfamiliar faces on identity and expression. *Psychological Research, 48*, 63–68.