HHS Public Access

Author manuscript

Atten Percept Psychophys. Author manuscript; available in PMC 2017 February 01.

Published in final edited form as:

Atten Percept Psychophys. 2016 February ; 78(2): 583–601. doi:10.3758/s13414-015-1026-y.

Timing in Audiovisual Speech Perception: A Mini Review and New Psychophysical Data

Jonathan H. Venezia[1], **Steven M. Thurman**[2], **William Matchin**[3], **Sahara E. George**[4], and **Gregory Hickok**[1]

[1]Department of Cognitive Sciences, University of California, Irvine [2]Department of Psychology, University of California, Los Angeles [3]Department of Linguistics, University of Maryland [4]Department of Anatomy and Neurobiology, University of California, Irvine

Abstract

Recent influential models of audiovisual speech perception suggest that visual speech aids perception by generating predictions about the identity of upcoming speech sounds. These models place stock in the assumption that visual speech leads auditory speech in time. However, it is unclear whether and to what extent temporally-leading visual speech information contributes to perception. Previous studies exploring audiovisual-speech timing have relied upon psychophysical procedures that require artificial manipulation of cross-modal alignment or stimulus duration. We introduce a classification procedure that tracks perceptually-relevant visual speech information in time without requiring such manipulations. Participants were shown videos of a McGurk syllable (auditory /apa/ + visual /aka/ = perceptual /ata/) and asked to perform phoneme identification (/apa/ yes-no). The mouth region of the visual stimulus was overlaid with a dynamic transparency mask that obscured visual speech in some frames but not others randomly across trials. Variability in participants' responses (∼35% identification of /apa/ compared to ∼5% in the absence of the masker) served as the basis for classification analysis. The outcome was a high resolution spatiotemporal map of perceptually-relevant visual features. We produced these maps for McGurk stimuli at different audiovisual temporal offsets (natural timing, 50-ms visual lead, and 100-ms visual lead). Briefly, temporally-leading (∼130 ms) visual information did influence auditory perception. Moreover, several visual features influenced perception of a single speech sound, with the relative influence of each feature depending on both its temporal relation to the auditory signal and its informational content.

Keywords

audiovisual speech; multisensory integration; prediction; classification image; timing; McGurk; speech kinematics

The visual facial gestures that accompany auditory speech form an additional signal that reflects a common underlying source (i.e., the positions and dynamic patterning of vocal

Corresponding Author: Jonathan Venezia, University of California, Irvine, Irvine, CA 92697, Phone: (949) 824-1409, Fax: (949) 824-2307, jvenezia@uci.edu.

tract articulators). Perhaps, then, it is no surprise that certain dynamic visual speech features, such as opening and closing of the lips and natural movements of the head, are correlated in time with dynamic features of the acoustic signal including its envelope and fundamental frequency (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; K. G. Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; H. C. Yehia, Kuratate, & Vatikiotis-Bateson, 2002). Moreover, higher-level phonemic information is partially redundant across auditory and visual speech signals, as demonstrated by expert speechreaders who can achieve extremely high rates of accuracy on speech-(lip-) reading tasks even when effects of context are minimized (Andersson & Lidestam, 2005). When speech is perceived in noisy environments, auditory cues to place of articulation are compromised, whereas such cues tend to be robust in the visual signal (R. Campbell, 2008; Miller & Nicely, 1955; Q. Summerfield, 1987; Walden, Prosek, Montgomery, Scherr, & Jones, 1977). Together, these findings suggest that information in the acoustic speech signal is somewhat preserved in – or at least enhanced by – the visual speech signal. In fact, visual speech is quite informative as evidenced by large intelligibility gains in noise for audiovisual speech compared to auditory speech alone (Erber, 1969; MacLeod & Summerfield, 1987; Neely, 1956; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954). However, there remains the question of exactly *how* visual speech is informative. One possibility is that the combination of partially redundant auditory and visual speech signals leads to better perception via simple multisensory enhancement (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Calvert, Campbell, & Brammer, 2000; Stein & Stanford, 2008). A second possibility – one that has achieved considerable attention recently and will be explored further here – is that visual speech generates predictions regarding the timing or identity of upcoming auditory speech sounds (Golumbic, Poeppel, & Schroeder, 2012; Grant & Seitz, 2000; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008; Virginie van Wassenhove, Grant, & Poeppel, 2005).

Support for the latter position derives from experiments designed to explore perception of cross-modal (audiovisual) synchrony. Such experiments artificially alter the stimulus onset asynchrony (SOA) between auditory and visual signals. Participants are asked to judge the temporal order of the signals (i.e., visual-first or audio-first) or indicate whether or not they perceive the signals as synchronous. A highly-replicated finding from this line of research is that, for a variety of audiovisual stimuli, simultaneity is maximally perceived when the visual signal leads the auditory signal (see Vroomen & Keetels, 2010 for a review). This effect is particularly pronounced for speech (although see also Maier, Di Luca, & Noppeney, 2011). In a classic study, Dixon and Spitz (1980) asked participants to monitor audiovisual clips of either a continuous speech stream (man reading prose) or a hammer striking a nail. The clips began fully synchronized and were gradually desynchronized in steps of 51 ms. Participants were instructed to respond when they could just detect the asynchrony. Average detection thresholds were larger when the video preceded the sound, and this effect was greater for speech (258-ms vs. 131-ms) than the hammer scenario (188-ms vs. 75-ms). Subsequent research has confirmed that auditory and visual speech signals are judged to be synchronous over a long, asymmetric temporal window that favors visual-lead SOAs (∼50-ms audio-lead to ∼200-ms visual-lead), with replications across a range of stimuli including connected speech (Eg & Behne, 2015; Grant, Wassenhove, & Poeppel, 2004), words

(Conrey & Pisoni, 2006), and syllables (V. van Wassenhove, Grant, & Poeppel, 2007). Moreover, audiovisual asynchrony only begins to disrupt speech recognition when the limits of this window have been reached (Grant & Greenberg, 2001). In other words, results from simultaneity judgment tasks hold when participants are asked to simply identify speech. This has been confirmed by studies of the McGurk effect (McGurk & MacDonald, 1976), an illusion in which an auditory syllable (e.g., /pa/) dubbed onto video of an incongruent visual syllable (e.g., /ka/) yields a perceived syllable that matches neither the auditory nor visual component (e.g., /ta/). Indeed, the McGurk effect is robust to audiovisual asynchrony over a range of SOAs similar to those that yield synchronous perception (Jones & Jarick, 2006; K. G. Munhall, Gribble, Sacco, & Ward, 1996; V. van Wassenhove et al., 2007).

The significance of visual-lead SOAs

The above research led investigators to propose the existence of a so-called audiovisual-speech temporal integration window (Dominic W Massaro, Cohen, & Smeele, 1996; Navarra et al., 2005; Virginie van Wassenhove, 2009; V. van Wassenhove et al., 2007). A striking feature of this window is its marked asymmetry favoring visual-lead SOAs. Low-level explanations for this phenomenon invoke cross-modal differences in simple processing time (Elliott, 1968) or natural differences in the propagation times of the physical signals (King & Palmer, 1985). These explanations alone are unlikely to explain patterns of audiovisual integration in speech, though stimulus attributes such as energy rise times and temporal structure have been shown to influence the shape of the audiovisual integration window (Denison, Driver, & Ruff, 2012; Van der Burg, Cass, Olivers, Theeuwes, & Alais, 2009). Recently, a more complex explanation based on predictive processing has received considerable support and attention. This explanation draws upon the assumption that visible speech information becomes available (i.e., visible articulators begin to move) prior to the onset of the corresponding auditory speech event (Grant et al., 2004; V. van Wassenhove et al., 2007). This temporal relationship favors integration of visual speech over long intervals. Moreover, visual speech is relatively coarse with respect to both time and informational content – that is, the information conveyed by speechreading is limited primarily to place of articulation (Grant & Walden, 1996; D.W. Massaro, 1987; Q. Summerfield, 1987; Quentin Summerfield, 1992), which evolves over a syllabic interval of ∼200 ms (Greenberg, 1999). Conversely, auditory speech events (especially with respect to consonants) tend to occur over short timescales of ∼20-40 ms (D. Poeppel, 2003; but see, e.g., Quentin Summerfield, 1981). When relatively robust auditory information is processed before visual speech cues arrive (i.e., at short audio-lead SOAs), there is no need to “wait around” for the visual speech signal. The opposite is true for situations in which visual speech information is processed before auditory-phonemic cues have been realized (i.e., even at relatively long visual-lead SOAs) – it pays to wait for auditory information to disambiguate among candidate representations activated by visual speech.

These ideas have prompted a recent upsurge in neurophysiological research designed to assess the effects of visual speech on early auditory processing. The results demonstrate unambiguously that activity in the auditory pathway is modulated by the presence of concurrent visual speech. Specifically, audiovisual interactions for speech stimuli are observed in the auditory brainstem response at very short latencies (∼11 ms post-acoustic

onset), which, due to differential propagation times, could only be driven by leading (pre-acoustic onset) visual information (Musacchia, Sams, Nicol, & Kraus, 2006; Wallace, Meredith, & Stein, 1998). Moreover, audiovisual speech modifies the phase of entrained oscillatory activity in the auditory cortex (Luo, Liu, & Poeppel, 2010; Power, Mead, Barnes, & Goswami, 2012), suggesting that visual speech may reset the phase of ongoing oscillations to ensure that expected auditory information arrives during a high neuronal-excitability state (Kayser, Petkov, & Logothetis, 2008; Schroeder et al., 2008). Finally, the latencies of event-related potentials generated in the auditory cortex are reduced for audiovisual syllables relative to auditory syllables, and the size of this effect is proportional to the predictive power of a given visual syllable (L. H. Arnal, Morillon, Kell, & Giraud, 2009; Stekelenburg & Vroomen, 2007; Virginie van Wassenhove et al., 2005). These data are significant in that they appear to argue against prominent models of audiovisual speech perception in which auditory and visual speech are highly processed in separate unisensory streams prior to integration (Bernstein, Auer, & Moore, 2004; D.W. Massaro, 1987).

Controversy over visual-lead timing in audiovisual speech perception

Until recently, visual-lead dynamics were merely assumed to hold across speakers, tokens, and contexts. In other words, it was assumed that visual-lead SOAs were the norm in natural audiovisual speech (David Poeppel, Idsardi, & van Wassenhove, 2008). It was only in 2009 – after the emergence of prominent theories emphasizing an early predictive role for visual speech (David Poeppel et al., 2008; Schroeder et al., 2008; Virginie van Wassenhove et al., 2005; V. van Wassenhove et al., 2007) – that Chandrasekaran and colleagues (2009) published an influential study in which they systematically measured the temporal offset between corresponding auditory and visual speech events in a number of large audiovisual corpora in different languages. Audiovisual temporal offsets were calculated by measuring the so-called “time to voice,” which can be found for a consonant-vowel (CV) sequence by subtracting the onset of the first consonant-related visual event (this is the half-way point of mouth closure prior to the consonantal release) from the onset of the first consonant-related auditory event (the consonantal burst in the acoustic waveform). Using this method, Chandrasekaran et al. identified a large and reliable visual lead (~150 ms) in natural audiovisual speech. Once again, these data seemed to provide support for the idea that visual speech is capable of exerting an early influence on auditory processing.

However, Schwartz and Savariaux (2014) subsequently pointed out a glaring fault in the data reported by Chandrasekaran et al. – namely, time-to-voice calculations were restricted to isolated CV sequences at the onset of individual utterances. Such contexts include so-called *preparatory gestures*, which are visual movements that by definition precede the onset of the auditory speech signal (the mouth opens and closes before opening again to produce the utterance-initial sound). In other words, preparatory gestures are visible but produce no sound, thus ensuring a visual-lead dynamic. They argued that isolated CV sequences are the exception rather than the rule in natural speech. In fact, most consonants occur in vowel-consonant-vowel (VCV) sequences embedded within utterances. In a VCV sequence, the mouth-closing gesture preceding the acoustic onset of the consonant does not occur in silence and actually corresponds to a *different* auditory event – the *offset* of sound energy related to the preceding vowel. Therefore, they argued, audiovisual asynchrony for

consonants should be calculated as the difference between the onset of the consonant-related acoustic energy and the onset of the mouth-*opening* gesture that corresponds to the consonantal release.

Schwartz and Savariaux (2014) went on to calculate two audiovisual temporal offsets for each token in a set of VCV sequences (consonants were plosives) produced by a single French speaker: (A) the difference between the time at which a decrease in sound energy related to the sequence-initial vowel was just measurable and the time at which a corresponding decrease in the area of the mouth was just measureable, and (B) the difference between the time at which an increase in sound energy related to the consonant was just measureable and the time at which a corresponding increase in the area of the mouth was just measureable. Using this technique, Schwartz & Savariaux found that auditory and visual speech signals were actually rather precisely aligned (between 20-ms audio-lead and 70-ms visual-lead). They concluded that large visual-lead offsets are mostly limited to the relatively infrequent contexts in which preparatory gestures occur at the onset of an utterance. Crucially, all but one of the recent neurophysiological studies cited in the preceding subsection used isolated CV syllables as stimuli (Luo et al., 2010 is the exception).

While this controversy appears to be a recent development, earlier studies explored audiovisual-speech timing relations extensively, with results often favoring the conclusion that temporally-leading visual speech is capable of driving perception. In a classic study by Campbell and Dodd (1980), participants perceived audiovisual consonant-vowel-consonant (CVC) words more accurately than matched auditory-alone or visual-alone (i.e., lip-read) words even when the acoustic signal was made to drastically lag the visual signal (up to 1600 ms). A series of perceptual gating studies in the early 1990s seemed to converge on the idea that visual speech can be perceived prior to auditory speech in utterances with natural timing. Visual perception of anticipatory vowel rounding gestures was shown to lead auditory perception by up to 200 ms in V-to-V (/i/ to /y/) spans across silent pauses (M.-A. Cathiard, Tiberghien, Tseva, Lallouache, & Escudier, 1991; see also M. Cathiard, Lallouache, Mohamadi, & Abry, 1995; M.-A. Cathiard, Lallouache, & Abry, 1996). The same visible gesture was perceived 40-60 ms ahead of the acoustic change when vowels were separated by a consonant (i.e., in a CVCV sequence; Escudier, Benoît, & Lallouache, 1990), and, furthermore, visual perception could be linked to articulatory parameters of the lips (Abry, Lallouache, & Cathiard, 1996). Additionally, accurate visual perception of bilabial and labiodental consonants in CV segments was demonstrated up to 80 ms prior to the consonant release (Smeele, 1994). Subsequent gating studies using CVC words have confirmed that visual speech information is often available early in the stimulus while auditory information continues to accumulate over time (Jesse & Massaro, 2010), and this leads to faster identification of audiovisual words (relative to auditory alone) in both silence and noise (Moradi, Lidestam, & Rönnberg, 2013).

Although these gating studies are quite informative the results are also difficult to interpret. Specifically, the results tell us that visual speech information *can be* perceived earlier in time than auditory speech. However, since gating involves artificial manipulation (truncation) of the stimulus, it is unclear whether and how early visual information affects

perception of unaltered speech tokens. One possible interpretation of the gating results is that there is an *informational* offset in audiovisual speech that favors visual-lead. This offset may or may not map cleanly to physical asynchronies between auditory and visual speech signals, which may explain the (partial) disagreement between purely physical measures and psychophysical measures based on gating. Due to the coarticulatory nature of speech, the visual signal available during physically-aligned segments may nonetheless provide information about the position of vocal tract articulators that predicts the identity of upcoming auditory speech sounds. Such predictions may be reflected in reduced latencies of auditory cortical potentials during perception of audiovisual speech (L. H. Arnal et al., 2009; Stekelenburg & Vroomen, 2007; Virginie van Wassenhove et al., 2005). Conversely, a recent review of the neurophysiological literature suggests that these early effects are likely to be modulatory rather than predictive *per se*, given (a) the nature of the anatomical connections between early visual and auditory areas, and (b) the fact that high-level (e.g., phonetic) features of visual and auditory speech are represented downstream in the visual and auditory cortical pathways, suggesting that extensive modal processing is required prior high-level audiovisual interactions (Bernstein & Liebenthal, 2014).

The current study

To sum up the preceding literature review, predictive models of audiovisual speech perception that posit a strong role for temporally-leading visual speech are partially supported by physical and psychophysical measurements of audiovisual-speech timing. Indeed, it is clear that visual speech *can be* perceived prior to auditory speech. However, the time-course of perception may not map cleanly to physical measurements of the auditory and visual signals. Moreover, the level at which early visual information influences perception remains to be pinned down. Crucially, current results based on synchrony manipulations and gating are limited in that the natural timing (and/or duration) of audiovisual stimuli must be artificially altered in order to perform the experiments, and, therefore, these experiments make it impossible to track the perceptual influence of visual speech over time under perceptual conditions well-matched to those in which natural audiovisual perception occurs. Indeed, it may be the case that early visual speech information does not strongly influence perception when audiovisual signals are temporally aligned and when participants have access to the full-duration signal in each modality. Moreover, synchrony manipulations destroy the natural temporal relationship between physical features of the auditory and visual stimuli, which makes it difficult to precisely compare the time-course of perception to the timing of events in the physical signals.

Here, we present a novel experimental paradigm that allows precise measurement of the visual influence on auditory speech perception over time *without desynchronizing or truncating the stimuli*. Specifically, our paradigm uses a multiplicative visual noise masking procedure with to produce a frame-by-frame classification of the visual features that contribute to audiovisual speech perception, assessed here using a McGurk paradigm with VCV utterances. The McGurk effect was chosen due to its widely accepted use as a tool to assess audiovisual integration in speech. VCVs were chosen in order to examine audiovisual integration for phonemes (stop consonants in the case of the McGurk effect) embedded within an utterance, rather than at the onset of an isolated utterance.

In a psychophysical experiment, we overlaid a McGurk stimulus with a spatiotemporally correlated visual masker that randomly revealed different components of the visual speech signal on different trials, such that the McGurk effect was obtained on some trials but not on others based on the masking pattern. In particular, the masker was designed such that important visual features (lips, tongue, etc.) would be visible only in certain frames, adding a temporal component to the masking procedure. Visual information crucial to the fusion effect was identified by comparing the making patterns on fusion trials to the patterns on non-fusion trials (Ahumada & Lovell, 1971; Eckstein & Ahumada, 2002; Gosselin & Schyns, 2001; Thurman, Giese, & Grossman, 2010; Vinette, Gosselin, & Schyns, 2004). This produced a high resolution spatiotemporal map of the visual speech information that contributed to estimation of speech signal identity.

Although the masking/classification procedure was designed to work without altering the audiovisual timing of the test stimuli, we repeated the procedure using McGurk stimuli with altered timing. Specifically, we repeated the procedure with asynchronous McGurk stimuli at two visual-lead SOAs (50 ms, 100 ms). We purposefully chose SOAs that fell well within the audiovisual-speech temporal integration window so that the altered stimuli would be perceptually indistinguishable from the unaltered McGurk stimulus (Virginie van Wassenhove, 2009; V. van Wassenhove et al., 2007). This was done in order to examine whether different visual stimulus features contributed to the perceptual outcome at different SOAs, even though the perceptual outcome itself remained constant. This was, in fact, not a trivial question. One interpretation of the tolerance to large visual-lead SOAs (up to ∼200 ms) in audiovisual-speech perception is that visual speech information is integrated at roughly the syllabic rate (∼4-5 Hz; Arai & Greenberg, 1997; Greenberg, 2006; V. van Wassenhove et al., 2007). The notion of a “visual syllable” suggests a rather coarse mechanism for integration of visual speech. However, several pieces of evidence leave open the possibility that visual information is integrated on a finer grain. First, the audiovisual speech detection advantage (i.e., an advantage in detecting, rather than identifying, audiovisual vs. auditory-only speech) is disrupted at a visual-lead SOA of only 40 ms (Kim & Davis, 2004). Further, observers are able to correctly judge the temporal order of audiovisual speech signals at visual-lead SOAs that continue to yield a reliable McGurk effect (Soto-Faraco & Alsius, 2007, 2009). Finally, it has been demonstrated that multisensory neurons in animals are modulated by changes in SOA even when these changes occur within the window in which auditory and visual signals are perceptually bound (King & Palmer, 1985; Meredith, Nemitz, & Stein, 1987; Stein, Meredith, & Wallace, 1993), and the same effect is observed in humans (as measured in fMRI) using audiovisual speech (Stevenson, Altieri, Kim, Pisoni, & James, 2010).

In addition to producing spatiotemporal classification maps at three SOAs (synchronized, 50ms visual-lead, 100 ms visual-lead), we extracted the time-course of lip movements in the visual speech stimulus and compared this signal to the temporal dynamics of audiovisual speech *perception*, as estimated from the classification maps. The results allowed us to address several relevant questions. First, what precisely are the visual cues that contribute to fusion? Second, when do these cues unfold relative to the auditory signal (i.e., is there any preference for visual information that precedes onset of the auditory signal)? Third, are these

cues related to any features in the time-course of lip movements? And finally, do the particular cues that contribute to the McGurk effect vary depending on audiovisual synchrony (i.e., do individual features within “visual syllables” exert independent influence on the identity of the auditory signal)?

To look ahead briefly, our technique succeeded in producing high temporal resolution classifications of the visual speech information that contributed to audiovisual speech perception – i.e., certain frames contributed significantly to perception while others did not. It was clear from the results that visual speech events occurring prior to the onset of the acoustic signal contributed significantly to perception. Additionally, the particular frames that contributed significantly to perception, and the relative magnitude of these contributions, could be tied to the temporal dynamics of lip movements in the visual stimulus (velocity in particular). Crucially, the visual features that contributed to perception varied as a function of SOA, even though all of our stimuli fell within the audiovisual-speech temporal window integration window and produced similar rates of the McGurk effect. The implications of these findings are discussed below.

Methods

Participants

A total of 34 (6 male) participants were recruited to take part in two experiments. All participants were right-handed, native speakers of English with normal hearing and normal or corrected-normal vision (self-report). Of the 34 participants, 20 were recruited for the main experiment (mean age = 21.6 yrs, SD = 3.0 yrs) and 14 for a brief follow-up study (mean age = 20.9 yrs, SD = 1.6 yrs). Three participants (all female) did not complete the main experiment and were excluded from analysis. Potential participants were screened prior to enrollment in the main experiment to ensure they experienced the McGurk effect. One potential participant was not enrolled on the basis of a low McGurk response rate (< 25%, compared to a mean rate of 95% in the enrolled participants). Participants were students enrolled at UC Irvine and received course credit for their participation. These students were recruited via the UC Irvine Human Subjects Lab. Oral informed consent was obtained from each participant in accordance with the UC Irvine Institutional Review Board guidelines.

Stimuli

Digital videos of a single male actor producing a sequence of vowel-consonant-vowel (VCV) non-words were recorded on a digital camera at a native resolution of 1080p at 60 frames per second. Videos captured the head and neck of the actor against a green screen. In post-processing, the videos were cropped to 500×500 pixels and the green screen was replaced with a uniform gray background. Individual clips of each VCV were extracted such that each contained 78 frames (duration 1.3 s). Audio was simultaneously recorded on separate device, digitized (44.1 kHz, 16-bit), and synced to the main video sequence in post-processing. VCVs were produced with a deliberate, clear speaking style. Each syllable was stressed and the utterance was elongated relative to a conversational speech. This was done to ensure that each event within the visual stimulus was sampled with the largest possible

number of frames, which was presumed to maximize the probability of detecting small temporal shifts using our classification method (see below). A consequence of using this speaking style was that the consonant in each VCV was strongly associated with the final vowel. An additional consequence was that our stimuli were somewhat artificial because the deliberate, clear style of speech employed here is relatively uncommon in natural speech.

In each VCV, the consonant was preceded and followed by the vowel /ɑ/ (as in 'father'). At least nine VCV clips were produced for each of the English voiceless stops – i.e, APA, AKA, ATA. Of these clips, five each of APA and ATA and one clip of AKA were selected for use in the study. To create a McGurk stimulus, audio from one APA clip was dubbed onto the video from the AKA clip. The APA audio waveform was manually aligned to the original AKA audio waveform by jointly minimizing the temporal disparity at the offset of the initial vowel and the onset of the consonant burst. This resulted in the onset of the consonant burst in the McGurk-aligned APA leading the onset of the consonant burst in the original AKA by 6 ms. This McGurk stimulus will henceforth be referred to as 'SYNC' to reflect the natural alignment of the auditory and visual speech signals. Two additional McGurk stimuli were created by altering the temporal alignment of the SYNC stimulus. Specifically, two clips with visual-lead SOAs within the audiovisual-speech temporal integration window (V. van Wassenhove et al., 2007) were created by lagging the auditory signal by 50 ms (V-Lead50) and 100 ms (V-Lead100), respectively. A silent period was added to the beginning of the V-Lead50 and V-Lead100 audio files to maintain duration at 1.3s.

Procedure

For all experimental sessions, stimulus presentation and response collection were implemented in Psychtoolbox-3 (Kleiner et al., 2007) on an IBM ThinkPad running Ubuntu Linux v12.04. Auditory stimuli were presented over Sennheiser HD 280 Pro headphones and responses were collected on a DirectIN keyboard (Empirisoft). Participants were seated ∼20 inches in front of the testing laptop inside a sound deadened chamber (IAC Acoustics). All auditory stimuli (including those in audiovisual clips) were presented at 68 dBA against a background of white noise at 62 dBA. This auditory signal-to-noise ratio (+6 dB) was chosen to increase the likelihood of the McGurk effect (Magnotti, Ma, & Beauchamp, 2013) without substantially disrupting identification of the auditory signal in isolation.

Each participant completed three days of testing spread over no more than a week. The task was phoneme identification on a six-point confidence scale. Participants were told they would be presented VCV non-words following the form /ɑ/-X-/ɑ/ (where "X" could be any consonant sound). On each trial of the experiment, participants were asked to indicate whether they perceived the non-word APA using the 1-6 keys on the response keyboard. The '1' key indicated highest confidence for APA and the '6' key indicated highest confidence for Not-APA, with the boundary between '3' and '4' corresponding to a categorical decision boundary. The response key was displayed to participants follows:

Sure apa 1 2 3 4 5 6 Sure Not apa

Phoneme identification was performed in three conditions: audio-only, unmasked audiovisual (Clear-AV), and masked audiovisual (Masked-AV). In the audio-only condition, stimuli were presented only over the headphones. Participants completed two blocks of 100 trials of auditory phoneme identification. The 100 trials comprised 50 trials in which the stimulus was APA, and 50 trials in which the stimulus was ATA. There were five separate APA tokens (including the APA audio used to create McGurk stimuli) and five separate ATA tokens, presented in random order (10 trials per token). On each trial, a black fixation cross was presented against a gray background over a jittered inter-trial interval (0.5-1.5s); at onset of the auditory signal, the fixation cross was replaced by the response key which remained on screen until participants made their response.

In the Clear-AV condition, stimuli were presented bimodally and the videos were unaltered (presented without a masker). Participants completed six blocks of 24 trials of audiovisual phoneme identification. In each block, 16 trials contained a McGurk stimulus, 4 trials contained one of four congruent APA videos, and 4 trials contained one of four congruent ATA videos. The congruent videos were included to provide occasional good exemplars of the most common percepts associated with the McGurk stimuli (though participants were not explicitly told that the McGurk stimuli were incongruent). The McGurk stimuli in each block were presented at a single SOA (SYNC, V-Lead50, or V-Lead100; 2 blocks each) with block order fully randomized. In each trial, a blank gray background appeared during a variable inter-trial interval (based on video loading times), followed by presentation of the video (1.3s); at the end of the video, the response key was flashed on screen and remained until participants made their response.

The crucial condition was Masked-AV, which differed from Clear-AV only in that a visual masker was added to the mouth region of the video on each trial. The masker disrupted the McGurk effect on some trials but not on others (see following section). Participants completed 24 blocks of 40 trials in the Masked-AV condition. In each block, there were 32 McGurk trials and 4 trials each of congruent APA or ATA. The McGurk stimuli in each block were presented at a single SOA (SYNC, V-Lead50, or V-Lead100; 8 blocks each) with block order fully randomized. The trial structure was identical to Clear-AV.

Visual masking technique

We implemented a visual masking technique to reveal the temporal dynamics of audiovisual perception in speech. This technique is similar to the multiplicative noise masking procedure known as “bubbles” (e.g. visual masking with randomly distributed Gaussian apertures; Gosselin & Schyns, 2001), which has been used successfully in several domains including face perception and in some of our previous work investigating biological motion perception (Thurman et al., 2010; Thurman & Grossman, 2011). Masking was applied to VCV video clips in the Masked-AV condition. For a given clip, we first down-sampled the clip to 120×120 pixels, and from this low-resolution clip we selected a 30×35 pixel region covering the mouth and part of the lower jaw of the speaker. The mean value of the pixels in this region was subtracted and a 30×35 mouth-region masker was applied as follows: (1) a random noise image was generated from a uniform distribution for each frame. (2) A Gaussian blur was applied to the random image sequence in the temporal domain (sigma =

2.1 frames) and in the spatial domain (sigma = 14 pixels) to create correlated spatiotemporal noise patterns. These were in fact low-pass filters with frequency cutoffs of 0.75 cycles/face and 4.5 Hz, respectively. Cutoff frequency was determined based on the sigma of the Gaussian filter in the frequency domain (or the point at which the filter gain was 0.6065 of maximum). The very low cutoff in the spatial domain produced a “shutter-like” effect when the noise masker was added to the mouth region of the stimulus – i.e., the masker tended to obscure large portions of the mouth region when it was opaque (Figure 1). (3) The blurred image sequence was scaled to a range of [0 1] and the resultant values were raised to the fourth power (i.e., a power transform) to produce essentially a map of alpha transparency values that were mostly opaque (e.g. close to 0), but with clusters of regions with high transparency (e.g. values close to 1). Specifically, “alpha transparency” refers to the degree to which the background image is allowed to show through the masker (1 = completely unmasked, 0 = completely masked, with a continuous scale between 1 and 0). (4) The alpha map was scaled to a maximum of 0.5 (a noise level found in pilot testing to work well with audiovisual speech stimuli). (5) The processed 30×35 image sequence was multiplied to the 30×35 mouth region of the original video separately in each RGB color frame. (6) The contrast variance and mean intensity of the masked mouth region was adjusted to match the original video sequence. (7) The fully processed sequence was up-sampled to 480×480 pixels for display. In the resultant video, a masker with spatiotemporally correlated alpha transparency values covered the mouth. Specifically, the mouth was (at least partially) visible in certain frames of the video, but not in other frames (Figure 1). Maskers were generated in real time and at random for each trial, such that no masker had the same pattern of transparent pixels. The crucial manipulation was masking of McGurk stimuli, where the logic of the masking process is as follows: when transparent components of the masker reveal critical visual features (i.e., of the mouth during articulation), the McGurk effect will be obtained; on the other hand, when critical visual features are blocked by the masker, the McGurk effect will be blocked. The set of visual features that contribute reliably to the effect can be estimated from the relation between behavioral response patterns and masker patterns over many trials. The specific masker created for each trial was saved for later analysis.

Data Analysis

Performance in the audio-only and Clear-AV conditions was evaluated in terms of % APA responses and mean confidence rating. The same measures were tabulated for congruent stimuli in the Masked-AV condition. The analysis of primary interest involved constructing ‘classification movies’ (CMs) for the SYNC, V-Lead50, and V-Lead100 McGurk stimuli based on behavior collected in the Masked-AV condition. Briefly, a classification movie considers the transparency of frames in the Masked-AV stimuli in its ability to predict participant responses. Details of the classification analysis follow.

Data from McGurk trials were grouped by stimulus: SYNC, V-Lead50, V-Lead100 (256 trials each). Separately for each stimulus, reverse correlation of trial-by-trial behavior with trial-by-trial masker alpha values was employed to construct CMs. Alpha values ranged from 0 (most opaque) to 0.5 (most transparent), and a separate alpha movie (the masker) was stored for each trial (30×35 pixels, 78 frames). Behavior was converted from six-point

confidence ratings to binary weights: on trials for which participants responded APA (1-3 on the confidence scale), the response was assigned a value of -1; on trials for which participants responded Not-APA (4-6 on the confidence scale), the response was assigned a value of 1. APA responses were correct with respect to the auditory signal (no McGurk fusion occurred) while Not-APA responses were incorrect with respect to the auditory signal (McGurk fusion occurred)[1]. In other words, fusion responses were weighted positively and non-fusion responses were weighted negatively. In the reverse correlation, we performed a trial-by-trial sum of the masker alpha movies weighted by the behavioral response. The weights were scaled by the overall fusion rate (proportion Not-APA responses). Specifically, we performed a weighted sum of the alpha values (a) over all trials (*t*) for each pixel location (*x*,*y*) in each frame (*f*) of the masker alpha movies. For a given fusion rate, *FR*:

$$CM_{x,y,f}=\sum_{t=1}^{256}\begin{cases}(1-FR)*a_t \text{if response is fusion}\\ -FR*a_t \text{if response is}\sim\text{fusion}\end{cases} \quad \text{for } x=1,\ldots,30; y=1,\ldots,35; f=1,\ldots,78$$

This procedure is equivalent to a multiple linear regression with behavioral responses serving as the criterion variable and across-trial fluctuations in alpha transparency at each pixel serving as the predictor variables. The result was a CM in which large positive values appeared at pixels that were frequently transparent when fusion occurred, while large negative values appeared at pixels that were frequently transparent when fusion did not occur. CMs for SYNC, V-Lead50, and V-Lead100 were created separately for each participant.

In order to identify CM pixels that were reliably positive or negative across participants, we constructed statistical maps for each stimulus. Individual-participant CMs were first averaged across participants, pixel-by-pixel, to obtain a group CM. The group CM was then scaled, pixel-by-pixel, by the estimated standard error of the mean, creating a map of t-statistics with n-1 degrees of freedom. These t-statistic maps were used to identify significant pixels on the group CMs. To account for multiple comparisons, we controlled the False Discovery Rate (FDR; Benjamini & Hochberg, 1995) such that pixels were considered significant only when $q < 0.05$. Only the pixels in frames 10-65 were included in statistical testing and multiple comparison correction. These frames covered the full duration of the auditory signal in the SYNC condition[2]. Visual features that contributed significantly to fusion were identified by overlaying the thresholded group CMs on the McGurk video. The efficacy of this technique in identifying critical visual features for McGurk fusion is demonstrated in Supplementary Video 1, where group CMs were used as a mask to generate *diagnostic* and *anti-diagnostic* video clips showing strong and weak McGurk fusion percepts, respectively. In order to chart the temporal dynamics of fusion, we created group

[1]The term “fusion” refers to trials for which the visual signal provided sufficient information to override the auditory percept. Such responses may reflect true fusion or also so-called “visual capture.” Since either percept reflects a visual influence on auditory perception, we are comfortable using Not-APA responses as an index of audiovisual integration or “fusion.” See also “Design choices in the current study” in the Discussion.

[2]Frames occurring during the final 50 and 100 ms of the auditory signal in the V-Lead50 and V-Lead100 conditions, respectively, were excluded from statistical analysis; we were comfortable with this given that the final 100 ms of the V-Lead100 auditory signal included only the tail end of the final vowel

classification time-courses for each stimulus by first averaging across pixels in each frame of the individual-participant CMs, and then averaging across participants to obtain a one-dimensional group time-course. For each frame (i.e., time-point), a t-statistic with n-1 degrees of freedom was calculated as described above. Frames were considered significant when FDR $q < 0.05$ (again restricting the analysis to frames 10-65).

Temporal dynamics of lip movements in McGurk stimuli

In the current experiment, visual maskers were applied to the mouth region of the visual speech stimuli. Previous work suggests that, among the cues in this region, the lips are of particular importance for perception of visual speech (Chandrasekaran et al., 2009; Grant & Seitz, 2000; Lander & Capek, 2013; McGrath, 1985). Thus, for comparison with the group classification time-courses, we measured and plotted the temporal dynamics of lip movements in the McGurk video following the methods established by Chandrasekaran et al. (2009). The interlip distance (Figure 2, top), which tracks the time-varying amplitude of the mouth opening, was measured frame-by-frame manually an experimenter (JV). For plotting, the resulting time course was smoothed using a Savitzky-Golay filter (order 3, window = 9 frames). It should be noted that, during production of /aka/, the interlip distance likely measures the extent to which the lower lip rides passively on the jaw. We confirmed this by measuring the vertical displacement of the jaw (frame-by-frame position of the superior edge of the mental protuberance of the mandible), which was nearly identical in both pattern and scale to the interlip distance.

The “velocity” of the lip opening was calculated by approximating the derivative of the interlip distance (Matlab ‘diff’). The velocity time course (Figure 2, middle) was smoothed for plotting in the same way as interlip distance. Two features related to production of the stop consonant (/k/) were identified from the interlip distance and velocity curves. Stop consonants typically involve a rapid closing of the mouth before opening to produce the subsequent sound. To identify the temporal signature of this closing phase, we looked backward in time from the onset of the consonant burst to find the point at which the interlip distance just started to decrease. This was marked by a trough in the velocity curve, and corresponded to initiation of the closure movement. We then looked forward in time to find the next peak in the velocity curve, which marked the point at which the mouth was half-closed and beginning to decelerate. The time between this half-closure point and the onset of the consonant burst, known as ‘time-to-voice’ (Chandrasekaran et al., 2009), was 167 ms for our McGurk stimulus (Figure 2, yellow shading).

We also calculated audiovisual asynchrony for the SYNC McGurk stimulus as in Schwarz and Savariaux (2014). An acoustic intensity contour was measured by extracting the speech envelope (Hilbert transform) and low-pass filtering (FIR filter with 4-Hz cutoff). This slow envelope was then converted to a dB scale (arbitrary units). The interlip distance curve was upsampled using cubic spline interpolation to match the sampling rate of the envelope. The onset of mouth closure was defined as the point at which the interlip distance was reduced by 0.15cm relative to its peak during production of the initial vowel (Figure 3, blue trace, 0.15cm↑), and the corresponding auditory event was defined as the point at which the envelope was reduced by 3dB from its initial peak (Figure 3, green trace, 3dB↓). The onset

of mouth opening was defined as the point at which the interlip distance increased by 0.15cm following the trough at vocal tract closure (Figure 3, blue trace, 0.15cm↑), and the corresponding auditory event was defined as the point at which the envelope increased 3dB from its own trough (Figure 3, green trace, 3dB↑). We repeated this analysis using the congruent AKA clip from which the McGurk video was derived (i.e., using the original AKA audio rather than the "dubbed" APA audio as in McGurk). For the SYNC McGurk stimulus, the audiovisual asynchrony at mouth closure was 163-ms visual-lead and the audiovisual asynchrony at mouth opening was 33-ms audio-lead (Figure 3, top). For the congruent AKA stimulus, the audiovisual asynchrony at mouth closure was 140-ms visual-lead and the audiovisual asynchrony at mouth opening was 32-ms audio-lead. These measurements indicate that our "dubbed" McGurk stimulus retained the audiovisual temporal characteristics of the congruent AKA utterance from which the McGurk video was drawn. More importantly, these measurements suggest a very precise audiovisual temporal relationship (within ∼30 ms) at the consonant in the VCV utterance, while measurements based on time-to-voice (Chandrasekaran et al., 2009) suggest a significant visual-lead (167 ms). A major advantage of the current experiment is the ability to determine unambiguously whether temporally-leading visual speech information occurring during the time-to-voice influences estimation of auditory signal identity in a VCV context.

It should be noted that various articulators including the upper and lower lips, jaw, tongue, and velum vary in terms of the timing of their movement onsets and offsets and the timing of their velocity maxima and minima (V. L. Gracco & Lofqvist, 1994; Kollia, Gracco, & Harris, 1995; Löfqvist & Gracco, 1999; McClean, 2000). Moreover, intra-articulator kinematic patterns and inter-articulator timing relations are sensitive to a number of factors including vowel context, specific movement goals, number of planned sounds in a vocal sequence, and speaking rate (Adams, Weismer, & Kent, 1993; V. Gracco, 1988; V. L. Gracco & Lofqvist, 1994; Löfqvist & Gracco, 1999, 2002; Parush, Ostry, & Munhall, 1983; Saltzman & Munhall, 1989). Therefore, the temporal relationship between articulator kinematics and the acoustic signal is not captured completely in Figures 3 and 4, which only track interlip distance and velocity. However, the choice to track interlip distance was motivated by the fact that changes in the oral aperture were among the most salient *visual* cues in the masker region of our /aka/ stimulus (see 'Visual masking technique' subsection above). Other articulators were visible only intermittently (e.g., the tongue) or their visible signals occurred mostly outside the classification region (e.g., the cheeks and jaw).

Results

Audio-only and Clear-AV

Auditory APA stimuli were perceived as APA 90% (1% SEM) of the time on average, and the mean confidence rating was 1.78 (0.07 SEM). Auditory ATA stimuli were perceived as APA 9% (2% SEM) of the time on average, and the mean confidence rating was 5.22 (0.14 SEM). The APA audio used to create the McGurk stimuli was perceived as APA 89% (2% SEM) of the time on average, and the mean confidence rating was 1.82 (0.11 SEM). Overall, this indicates that some perceptual uncertainty was introduced for auditory stimuli at the

+6dB SNR chosen for auditory presentation, but overall auditory-only perception was quite accurate.

For reporting the results of the Clear- AV condition, we will focus on the McGurk stimuli (performance for congruent AV stimuli was at ceiling). Recall that in McGurk stimuli, an auditory APA was dubbed on a visual AKA. Responses that did not conform to the identity of the auditory signal were considered fusion responses. The SYNC stimulus was perceived as APA 5% (3% SEM) of the time on average, with a mean confidence rating of 5.34 (0.16 SEM). The V-Lead50 stimulus was perceived as APA 6% (3% SEM) of the time on average, with a mean confidence rating of 5.33 (0.15 SEM). The V-Lead100 stimulus was perceived as APA 6% (3% SEM) of the time on average, with a mean confidence rating of 5.34 (0.17 SEM). Three conclusions are clear from these data. First, a very large proportion of responses (>90%) deviated from the identity of the auditory signal, indicating a high rate of fusion. Second, this rate of fusion did not differ significantly across the McGurk stimuli ($F(2,15) = 0.32$, $p = .732$, $\eta_p^2 = .041$), nor did confidence ratings ($F(2,15) = 0.01$, $p = .986$, $\eta_p^2 = .002$), suggesting that the McGurk stimuli were all perceptually bound despite the asynchrony manipulation. Third, McGurk stimuli were judged as Not-APA with roughly the same frequency and confidence as for auditory ATA stimuli, suggesting a very strong influence of the visual stimulus on auditory signal identity. This was the intended effect of adding low-intensity white noise to the auditory signal.

In a follow-up experiment with a separate group of participants, unmasked (Clear-AV) clips of the SYNC McGurk stimulus along with congruent APA, AKA, and ATA stimuli were presented in a 4-AFC design (12 trials per stimulus, random order). Participants were asked to indicate the identity of the stimulus using the numerical keypad (1=APA, 2=AKA, 3=ATA, 4=OTHER). This experiment followed the same trial structure as the main experiment, other than the 4-AFC response prompt. Stimulus presentation equipment and auditory levels were identical to the main experiment (including the addition of auditory white noise). The SYNC McGurk stimulus was judged as ATA 92% (± 3% SEM) of the time on average, indicating a high degree of fusion. All congruent stimuli were perceived accurately >90% of the time.

Masked-AV

Congruent APA videos were perceived as APA 95% of the time on average, while congruent ATA videos were perceived as APA 4% of the time on average, indicating that perception of congruent videos was largely unaffected by the masker. The SYNC McGurk stimulus was perceived as APA 40% (4% SEM) on average, with a mean confidence rating of 3.87 (0.80 SEM). The V-Lead50 McGurk stimulus was perceived as APA 37% (4% SEM) on average, with a mean confidence rating of 3.97 (0.71 SEM). The V-Lead100 McGurk stimulus was perceived as APA 33% (4% SEM) on average, with a mean confidence rating of 4.13 (0.65 SEM). Thus, we observed a net increase (relative to Clear-AV) of APA responses equal to 35% for SYNC, 31% for V-Lead50, and 27% for V-Lead100, indicating a significant reduction of fusion responses due to the masker. This reduction was significant for all three conditions (SYNC: $t(16) = 10.6$, $p < .001$, $d = 2.57$; V-Lead50: $t(16) = 11.3$, $p < .001$, $d = 2.75$; V-Lead100: $t(16) = 9.2$, $p < .001$, $d = 2.24$). In fact

this reduction, and the variation it induced across trials, provided the basis for classification of the visual features that contribute to fusion.

Example frames from the FDR-corrected classification movie (CM) for the SYNC stimulus are presented in Figure 4 (see Supplementary Figs. 1-2 for V-Lead50 and V-Lead100 CMs). Some comments are warranted. First, there are several frames in which significant negative-valued pixels can be identified (i.e., pixels that were reliably transparent on non-fusion trials). Since we were primarily interested in the pattern of positive-valued pixels (i.e., those that drove fusion), we will restrict further discussion almost entirely to positive pixels/frames. Second, since the masker region was rather small (i.e., confined to the mouth), and because a high spatial correlation was induced in the maskers, it is difficult to make meaningful conclusions about the specific spatial patterns revealed in individual frames of the CMs. We were primarily interested in the temporal dynamics of fusion, so from this point forward we will focus on the classification time-courses.

Classification time-courses for the SYNC, V-Lead50, and V-Lead100 stimuli are plotted in Figure 5 along with a trace of the auditory waveform from each stimulus. Significant frames (FDR-corrected) are labeled with red circles. Positive values occur at frames that tended to be transparent on fusion trials and opaque on non-fusion trials, so we conclude that significant positive frames contributed reliably to fusion (for demonstration Supplementary Video 1). Recall that the V-Lead50 and V-Lead100 stimuli were created by shifting the auditory signal relative to the video, so in Figure 5 the frame number corresponds to identical visual information across all three SOAs.

In Figure 5 several results are immediately apparent: (1) each of the classification time-courses reaches its peak at the same point in time; (2) the morphology of the SYNC time-course differs from the V-Lead50 and V-Lead100 time-courses; (3) there are more significant frames in the SYNC time-course than the V-Lead50 or V-Lead100 time-courses. Regarding (1), the exact location of the peak in each time-course was frame 42, and this pattern was rather stable across participants. For the SYNC stimulus, 11 of 17 participants had their classification peak within ± 2 frames of the group peak and 14 of 17 participants had a local maximum within ± 2 frames of the group peak. For the V-Lead50 stimulus, these proportions were 12/17 and 15/17, respectively; and for the V-Lead100 stimulus, 13/17 and 16/17, respectively. Regarding (2), the most obvious difference in morphology concerns the width of the time-courses where they significantly exceed zero. The SYNC time-course is clearly wider than the V-Lead50 or V-Lead100 time-courses, owing primarily to an increased contribution of early frames (tested directly below). Regarding (3), the SYNC stimulus contained the most significant positive frames and the only significant negative frames. The significant positive region of the SYNC time-course ranged from frame 30 through 46 (283.33 ms), while this range was 38 through 45 (133.33 ms) and 38 through 46 (150 ms) for the V-Lead50 and V-Lead100 time-courses, respectively. Several significant negative frames bracketed the significant positive portion of the SYNC time-course. Briefly, we speculate that participants learned to attend to a wider range of visual information in the SYNC condition (evidenced by the increased number of significant positive frames), which allowed some neighboring uninformative frames to occasionally drive perception away from fusion.

In Figure 6 we zoom in on the classification time-courses where they contain significant positive frames. We plot the time-courses aligned to the lip velocity curve over the same time period. Stages of oral closure are labeled on the velocity curve. The shaded regions from Figure 2 are reproduced, accounting for shifts in the audio for the V-Lead50 and V-Lead100 stimuli. Two features of Figure 6 are significant. First, the peak region on each classification time-course clearly corresponds to the region of the lip velocity curve describing acceleration of the lips toward peak velocity during the release of airflow in production of the consonant /k/. Second, eight significant frames in the SYNC time-course fall in the time period prior to the onset of the consonant-related auditory burst (shaded yellow in Fig. 6), while the V-Lead50 and V-Lead100 time-courses contain zero significant frames in this period. This suggests that the SYNC time-course is significantly different from the V-Lead50 and V-Lead100 time-courses this region. To test this directly, we averaged individual-participant time-courses across the eight-frame window in which SYNC contained significant ‘pre-burst’ frames (fr. 30-37) and computed paired t-tests comparing SYNC to V-Lead50 and V-Lead100, respectively. In fact, SYNC was marginally greater than V-Lead50 ($t(16) = 2.05$, $p = .057$) and significantly greater than V-Lead100 ($t(16) = 2.79$, $p = .013$).

Discussion

We have developed a novel experimental paradigm for mapping the temporal dynamics of audiovisual integration in speech. Specifically, we employed a phoneme identification task in which we overlaid McGurk stimuli with a spatiotemporally correlated visual masker that revealed critical visual cues on some trials but not on others. As a result, McGurk fusion was observed only on trials for which critical visual cues were available. Behavioral patterns in phoneme identification (fusion or no fusion) were reverse correlated with masker patterns over many trials, yielding a classification time-course of the visual cues that contributed significantly to fusion. This technique provides several advantages over techniques used previously to study the temporal dynamics of audiovisual integration in speech. First, as opposed to temporal gating (M.-A. Cathiard et al., 1996; Jesse & Massaro, 2010; K. G. Munhall & Tohkura, 1998; Smeele, 1994) in which only the first part of the visual or auditory stimulus is presented to the participant (up to some pre-determined “gate” location), masking allows presentation of the entire stimulus on each trial. Second, as opposed to manipulations of audiovisual synchrony (Conrey & Pisoni, 2006; Grant & Greenberg, 2001; K. G. Munhall et al., 1996; V. van Wassenhove et al., 2007), masking does not require the natural timing of the stimulus to be altered. As in the current study, one can choose to manipulate stimulus timing to examine changes in audiovisual temporal dynamics relative to the unaltered stimulus. Finally, while techniques have been developed to estimate natural audiovisual timing based on physical measurements of speech stimuli (Chandrasekaran et al., 2009; Schwartz & Savariaux, 2014), our paradigm provides behavioral verification of such measures based on actual human perception. To the best of our knowledge, this is the first application of a “bubbles-like” masking procedure (Fiset et al., 2009; Thurman et al., 2010; Thurman & Grossman, 2011; Vinette et al., 2004) to a problem of multisensory integration.

In the present experiment, we performed classification analysis with three McGurk stimuli presented at different audiovisual SOAs – natural timing (SYNC), 50-ms visual-lead (V-Lead50), and 100-ms visual-lead (V-Lead100). Three significant findings summarize the results. First, the SYNC, V-Lead50, and V-Lead100 McGurk stimuli were rated nearly identically in a phoneme identification task with no visual masker. Specifically, each stimulus elicited a high degree of fusion suggesting that all of the stimuli were perceived similarly. Second, the *primary* visual cue contributing to fusion (peak of the classification time-courses, Figs. 5-6) was identical across the McGurk stimuli (i.e., the position of the peak was not affected by the temporal offset between the auditory and visual signals). Third, despite this fact, there were significant differences in the contribution of a *secondary* visual cue across the McGurk stimuli. Namely, an early visual cue – that is, one related to lip movements that preceded the onset of the consonant-related auditory signal – contributed significantly to fusion for the SYNC stimulus, but not for the V-Lead50 or V-Lead100 stimuli. The latter finding is noteworthy because it reveals that (a) temporally-leading visual speech information can significantly influence estimates of auditory signal identity, and (b) individual visual speech features exert independent influence on estimates of auditory signal identity.

Temporally-leading visual speech information influences auditory signal identity

In the Introduction, we reviewed a recent controversy surrounding the role of temporally-leading visual information in audiovisual speech perception. In particular, several prominent models of audiovisual speech perception (Luc H Arnal, Wyart, & Giraud, 2011; Bever, 2010; Golumbic et al., 2012; Power et al., 2012; Schroeder et al., 2008; Virginie van Wassenhove et al., 2005; V. van Wassenhove et al., 2007) have postulated a critical role for temporally-leading visual speech information in generating predictions of the timing or identity of the upcoming auditory signal. A recent study (Chandrasekaran et al., 2009) appeared to provide empirical support for the prevailing notion that visual-lead SOAs are the norm in natural audiovisual speech. This study showed that visual speech leads auditory speech by ∼150 ms for isolated CV syllables. A later study (Schwartz & Savariaux, 2014) used a different measurement technique and found that VCV utterances contained a range of audiovisual asynchronies that did not strongly favor visual-lead SOAs (20-ms audio-lead to 70-ms visual-lead). We measured the natural audiovisual asynchrony (Figs. 2-3) in our SYNC McGurk stimulus (which, crucially, was a VCV utterance) following both Chandrasekaran et al. (2009) and Schwartz & Savariaux (2014). Measurements based on Chandrasekaran et al. suggested a 167-ms visual-lead, while measurements based on Schwartz & Savariaux suggested a 33-ms audio-lead. When we measured the time-course of the *actual* visual influence on auditory signal identity (Figs. 5-6, SYNC), we found that a large number of frames within the 167-ms visual-lead period exerted such influence. Therefore, our study demonstrates unambiguously that temporally-leading visual information can influence subsequent auditory processing, which concurs with previous behavioral work (M. Cathiard et al., 1995; Jesse & Massaro, 2010; K. G. Munhall et al., 1996; Sánchez-García, Alsius, Enns, & Soto-Faraco, 2011; Smeele, 1994).

However, our data also suggest that the temporal position of visual speech cues relative to the auditory signal may be less important than the informational content of those cues. As

mentioned above, classification time-courses for all three of our McGurk stimuli reached their peak at the same frame (Figs. 5-6). This peak region coincided with an acceleration of the lips corresponding to the release of airflow during consonant production. Examination of the SYNC stimulus (natural audiovisual timing) indicates that this visual-articulatory gesture unfolded over the same time period as the consonant-related portion of the auditory signal. Therefore, the *most influential* visual information in the stimulus temporally overlapped the auditory signal. This information remained influential in the V-Lead50 and V-Lead100 stimuli when it preceded the onset of the auditory signal. This is interesting in light of the theoretical importance placed on visual speech cues that lead the onset of the auditory signal. In our study, the most informative visual information was related to the actual release of airflow during articulation, rather than closure of the vocal tract during the stop, and this was true whether this information preceded or overlapped the auditory signal in time. As such, while visual information about consonant identity was indeed available prior to onset of the auditory signal, the relative contribution of different visual cues depended as much (or more) on the information content of the visual signal as it did on the temporal relationship between the visual and auditory signals.

The relatively weak contribution of temporally-leading visual information in the current study may be attributable to the particular stimulus employed to produce McGurk effects (visual AKA, auditory APA). In particular, the visual velar /k/ in AKA is less distinct than other stops during vocal tract closure and makes a relatively weak prediction of the consonant identity relative to, e.g., a bilabial /p/ (L. H. Arnal et al., 2009; Q. Summerfield, 1987; Quentin Summerfield, 1992; Virginie van Wassenhove et al., 2005). Moreover, the particular AKA stimulus employed in our study was produced using a clear speech style with stress placed on each vowel. The amplitude of the mouth movements was quite large, and the mouth nearly closed during production of the stop. Such a large closure is atypical for velar stops and, in fact, made our stimulus similar to typical bilabial stops. If anything, this reduced the strength of early visual cues – namely, had the lips remained farther apart during vocal tract closure, this would have provided strong perceptual evidence against APA, and so would have favored not-APA (i.e., fusion). Whatever the case, the present study provides clear evidence that both temporally-leading and temporally-overlapping visual speech information can be quite informative.

Individual visual speech features exert independent influence on auditory signal identity

Previous work on audiovisual integration in speech suggests that visual speech information is integrated on a rather coarse, syllabic time-scale (see, e.g., V. van Wassenhove et al., 2007). In the Introduction we reviewed work suggesting that it is possible for visual speech to be integrated on a finer grain (Kim & Davis, 2004; King & Palmer, 1985; Meredith et al., 1987; Soto-Faraco & Alsius, 2007, 2009; Stein et al., 1993; Stevenson et al., 2010). We provide evidence that, in fact, individual features within “visual syllables” are integrated non-uniformly.

In our study, a baseline measurement of the visual cues that contribute to audiovisual fusion is given by the classification time-course for the SYNC McGurk stimulus (natural audiovisual timing). Inspection of this time course reveals that 17 video frames (30-46)

contributed significantly to fusion (i.e., there were 17 positive-valued significant frames). If these 17 frames compose a uniform “visual syllable,” this pattern should be largely unchanged for the V-Lead50 and V-Lead100 time-courses. Specifically, the V-Lead50 and V-Lead100 stimuli were constructed with relatively short visual-lead SOAs (50 ms and 100 ms, respectively) that produced no behavioral differences in terms of McGurk fusion rate. In other words, each stimulus was equally well bound within the audiovisual-speech temporal integration window. However, the set of visual cues that contributed to fusion for V-Lead50 and V-Lead100 was different than the set for SYNC. In particular, all of the early significant frames (30-37) dropped out from the classification time-course when the auditory signal was delayed – there were only 8 video frames (38-45) that contributed to fusion for V-Lead50, and only 9 video frames (38-46) contributed to fusion for V-Lead100. Overall, early frames had progressively less influence on fusion as the auditory signal was lagged further in time, evidenced by follow-up t-tests indicating that frames 30-37 were marginally different for SYNC vs. V-Lead50 ($p = .057$) and significantly different for SYNC vs. V-Lead100 ($p = .013$). Of crucial importance, the temporal shift from SYNC to V-Lead50 had a nonlinear effect on the classification results – i.e., a 50 ms shift in the auditory signal, which corresponds to a three-frame shift with respect to the visual signal, reduced or eliminated the contribution of *eight* early frames (Figs. 5-6; also compare Fig. 4 to Supplementary Fig. 1 for a more fine-grained depiction of this effect). This suggests that the observed effects cannot be explained merely by postulating a fixed temporal integration window that slides and “grabs” any informative visual frame within its boundaries. Rather, discrete visual events contributed to speech-sound “hypotheses” of varying strength, such that a relatively low-strength hypothesis related to an early visual event (frames labeled ‘pre-burst’ in Fig. 6) was no longer significantly influential when the auditory signal was lagged by 50 ms.

Thus, we suggest in accordance with previous work (Green, 1998; Green & Norrix, 2001; Jordan & Sergeant, 2000; K. Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004; Rosenblum & Saldaña, 1996) that dynamic (perhaps kinematic) visual features are integrated with the auditory signal. These features likely reveal some key timing information related to articulatory kinematics but need not have any particular level of phonological specificity (Chandrasekaran et al., 2009; K. G. Munhall & Vatikiotis-Bateson, 2004; Q. Summerfield, 1987; H. Yehia, Rubin, & Vatikiotis-Bateson, 1998; H. C. Yehia et al., 2002). Several findings in the current study support the existence of such features. Immediately above, we described a nonlinear dropout with respect to the contribution of early visual frames in the V-Lead50 classification relative to SYNC. This suggests that a discrete visual feature (likely related to vocal tract closure during production of the stop) no longer contributed significantly to fusion when the auditory signal was lagged by 50 ms. Further, the peak of the classification time-courses was identical across all McGurk stimuli, regardless of the temporal offset between the auditory and visual speech signals. We believe this peak corresponds to a visual feature related to the release of air in consonant production (Figure 6). We suggest that visual features are weighted in the integration process according to three factors: (1) visual salience (Vatakis, Maragos, Rodomagoulakis, & Spence, 2012), (2) information content, and (3) temporal proximity to the auditory signal (closer = greater weight). To be precise, representations of visual features are activated with strength proportional to visual salience and information content (both high for the ‘release’ feature

here), and this activation decays over time such that visual features occurring farther in time from the auditory signal are weighted less heavily ('pre-release' feature here). This allows the auditory system to "look back" in time for informative visual information. The 'release' feature in our McGurk stimuli remained influential even when it was temporally distanced from the auditory signal (e.g., V-Lead100) because of its high salience and because it was the only informative feature that remained activated upon arrival and processing of the auditory signal. Qualitative neurophysiological evidence (dynamic source reconstructions form MEG recordings) suggests that cortical activity loops between auditory cortex, visual motion cortex, and heteromodal superior temporal cortex when audiovisual convergence has not been reached, e.g. during lipreading (L. H. Arnal et al., 2009). This may reflect maintenance of visual features in memory over time for repeated comparison to the incoming auditory signal.

Design choices in the current study

Several of the specific design choices in the current study warrant further discussion. First, in the application of our visual masking technique, we chose to mask only the part of the visual stimulus containing the mouth and part of the lower jaw. This choice obviously limits our conclusions to mouth-related visual features. This is a potential shortcoming since it is well known that other aspects of face and head movement are correlated with the acoustic speech signal (Jiang, Alwan, Keating, Auer, & Bernstein, 2002; Jiang, Auer, Alwan, Keating, & Bernstein, 2007; K. G. Munhall et al., 2004; H. Yehia et al., 1998; H. C. Yehia et al., 2002). However, restricting the masker to the mouth region reduced computing time and thus experiment duration since maskers were generated in real time. Moreover, previous studies demonstrate that interference produced by incongruent audiovisual speech (similar to McGurk effects) can be observed when only the mouth is visible (Thomas & Jordan, 2004), and that such effects are almost entirely abolished when the lower half of the face is occluded (Jordan & Thomas, 2011).

Second, we chose to test the effects of audiovisual asynchrony allowing the visual speech signal to lead by 50 and 100 ms. These values were chosen to be well within the audiovisual speech temporal integration window for the McGurk effect (V. van Wassenhove et al., 2007). It may have been useful to test visual-lead SOAs closer to the limit of the integration window (e.g., 200 ms), which would produce less stable integration. Similarly, we could have tested audio-lead SOAs where even a small temporal offset (e.g., 50 ms) would push the limit of temporal integration. We ultimately chose to avoid SOAs at the boundary of the temporal integration window because less stable audiovisual integration would lead to a reduced McGurk effect, which would in turn introduce noise into the classification procedure. Specifically, if the McGurk fusion rate were to drop far below 100% in the Clear-AV (unmasked) condition, it would be impossible to know whether non-fusion trials in the Masked-AV condition were due to presence of the masker itself or, rather, to a failure of temporal integration. We avoided this problem by using SOAs that produced high rates of fusion (i.e., "not-APA" responses) in the Clear-AV condition (SYNC = 95%, V-Lead50 = 94%, V-Lead100 = 94%). Moreover, we chose adjust the SOA in 50 ms steps because this step size constituted a three-frame shift with respect to the video, which was presumed to be sufficient to drive a detectable change in the classification.

Third, we added 62 dBA of noise to auditory speech signals (+6 dB SNR) throughout the experiment. As mentioned above, this was done to increase the likelihood of fusion by increasing perceptual reliance on the visual signal (Alais & Burr, 2004; Shams & Kim, 2010) so as to drive fusion rates as high as possible, which had the effect of reducing the noise in the classification procedure. However, there was a small tradeoff in terms of noise introduced to the classification procedure – namely, adding noise to the auditory signal caused auditory-only identification of APA to drop to 90%, suggesting that up to 10% of "not-APA" responses in the Masked-AV condition were judged as such purely on the basis of auditory error. If we assume that participants' responses were unrelated to the visual stimulus on 10% of trials (i.e., those trials in which responses were driven purely by auditory error), then 10% of trials contributed only noise to the classification analysis. Nevertheless, we obtained a reliable classification even in the presence of this presumed noise source, which only underscores the power of the method.

Fourth, we chose to collect responses on a 6-point confidence scale that emphasized identification of the nonword APA (i.e., the choices were between APA and Not-APA). The major drawback of this choice is that we do not know precisely what participants perceived on fusion (Not-APA) trials. A 4-AFC calibration study conducted on a different group of participants showed that our McGurk stimulus was overwhelmingly perceived as ATA (92%). A simple alternative would have been to force participants to choose between APA (the true identity of the auditory signal) and ATA (the presumed percept when McGurk fusion is obtained), but any participants who perceived, for example, AKA on a significant number of trials would have been forced to arbitrarily assign this to APA or ATA. We chose to use a simple identification task with APA as the target stimulus so that any response involving some visual interference (AKA, ATA, AKTA, etc.) would be attributed to the Not-APA category. There is some debate regarding whether percepts such as AKA or AKTA represent true fusion, but in such cases it is clear that visual information has influenced auditory perception. For the classification analysis, we chose to collapse confidence ratings to binary APA/not-APA judgments. This was done because some participants were more liberal in their use of the '1' and '6' confidence judgments (i.e., frequently avoiding the middle of the scale). These participants would have been over-weighted in the analysis, introducing a between-participant source of noise and counteracting the increased within-participant sensitivity afforded by confidence ratings. In fact, any between-participant variation in criteria for the different response levels would have introduced noise to the analysis.

A final issue concerns the generalizability of our results. In the present study, we presented classification data based on a single voiceless McGurk token, spoken by just one individual. This was done to facilitate collection of the large number of trials needed for a reliable classification. Consequently, certain specific aspects of our data may not generalize to other speech sounds, tokens, speakers, etc. These factors have been shown to influence the outcome of, e.g., gating studies (Troille, Cathiard, & Abry, 2010). However, the main findings of the current study seem likely to hold across additional contexts. Crucially, we have demonstrated a viable new method for classification of the visual speech features that influence auditory signal identity over time, and this method can be extended or modified in

future research. Refinements to the technique will likely allow for reliable classification in fewer trials and thus across a greater number of tokens and speakers.

Conclusions

Our visual masking technique successfully classified visual cues that contributed to audiovisual speech perception. We were able to chart the temporal dynamics of fusion at a high resolution (60 Hz). The results of this procedure revealed details of the temporal relationship between auditory and visual speech that exceed those available in typical physical or psychophysical measurements. We demonstrated unambiguously that temporally-leading visual speech information can influence auditory signal identity (in this case, the identity of a consonant), even in a VCV context devoid of consonant-related preparatory gestures. However, our measurements also suggested that temporally-overlapping visual speech information was equally if not more informative than temporally-leading visual information. In fact, it seems that the influence exerted by a particular visual cue has as much or more to do with its informational content as it does with its temporal relation to the auditory signal. However, we did find that the set of visual cues that contributed to audiovisual fusion varied depending on the temporal relation between the auditory and visual speech signals, even for stimuli that were perceived identically (in terms of phoneme identification rate). We interpreted these result in terms if a conceptual model of audiovisual-speech integration in which dynamic visual features are extracted and integrated proportional to their salience, informational content, and temporal proximity to the auditory signal. This model is not inconsistent with the notion that visual speech predicts the identity of upcoming auditory speech sounds, but suggests that ‘prediction’ is akin to simple activation and maintenance of dynamic visual features that influence estimates of auditory signal identity over time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This investigation was supported by the National Institute on Deafness and Other Communication Disorders Award DC009659 to G.H. Vivian Lu, Ereka Gordon, and Nayeem Siddique were student research assistants who contributed to this project.

References

Abry, C.; Lallouache, MT.; Cathiard, MA. Speechreading by humans and machines. Springer; 1996. How can coarticulation models account for speech sensitivity to audio-visual desynchronization?; p. 247-255.

Adams SG, Weismer G, Kent RD. Speaking rate and speech movement velocity profiles. Journal of Speech, Language, and Hearing Research. 1993; 36(1):41–54.

Ahumada A, Lovell J. Stimulus Features in Signal Detection. The Journal of the Acoustical Society of America. 1971; 49(6B):1751–1756. doi: doi:http://dx.doi.org/10.1121/1.1912577.

Alais D, Burr D. The ventriloquist effect results from near-optimal bimodal integration. Current Biology. 2004; 14(3):257–262. [PubMed: 14761661]

Andersson U, Lidestam B. Bottom-up driven speechreading in a speechreading expert: the case of AA (JK023). Ear and hearing. 2005; 26(2):214–224. [PubMed: 15809546]

Arai T, Greenberg S. The temporal properties of spoken Japanese are similar to those of English. Paper presented at the EUROSPEECH. 1997

Arnal LH, Morillon B, Kell CA, Giraud AL. Dual neural routing of visual facilitation in speech processing. J Neurosci. 2009; 29(43):13445–13453.10.1523/JNEUROSCI.3194-09.2009 [PubMed: 19864557]

Arnal LH, Wyart V, Giraud AL. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci. 2011; 14(6):797–801. [PubMed: 21552273]

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci. 2004; 7(11):1190–1192.10.1038/nn1333 [PubMed: 15475952]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289–300.

Bernstein, LE.; Auer, ET.; Moore, JK. Audiovisual Speech Binding: Convergence or Association?. In: Calvert, GA.; Spence, C.; Stein, BE., editors. Handbook of Multisensory Processing. Cambridge, MA: MIT Press; 2004. p. 203-223.

Bernstein LE, Liebenthal E. Neural pathways for visual speech perception. Frontiers in neuroscience. 2014; 8

Bever TG, P D. Analysis by synthesis: A (re-) emerging program of research for language and vision. Biolinguistics. 2010; 4(2-3):174–200.

Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology. 2000; 10(11):649–658. [PubMed: 10837246]

Campbell R. The processing of audio-visual speech: empirical and neural bases. Philos Trans R Soc Lond B Biol Sci. 2008; 363(1493):1001–1010.10.1098/rstb.2007.2155 [PubMed: 17827105]

Campbell R, Dodd B. Hearing by eye. Quarterly Journal of Experimental Psychology. 1980; 32(1):85–99. [PubMed: 7367580]

Cathiard, MA.; Lallouache, MT.; Abry, C. Speechreading by humans and machines. Springer; 1996. Does movement on the lips mean movement in the mind?; p. 211-219.

Cathiard, MA.; Tiberghien, G.; Tseva, A.; Lallouache, MT.; Escudier, P. Visual perception of anticipatory rounding during acoustic pauses: a cross-language study. Paper presented at the Proceedings of the 12th Int. Congress of Phonetic Sciences; Aixen-Provence, France. 1991.

Cathiard M, Lallouache M, Mohamadi T, Abry C. Configurational vs. temporal coherence in audio-visual speech perception. Paper presented at the Proceedings of the 13th International Congress of Phonetic Sciences. 1995

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA. The natural statistics of audiovisual speech. PLoS computational biology. 2009; 5(7):e1000436. [PubMed: 19609344]

Conrey B, Pisoni DB. Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. The Journal of the Acoustical Society of America. 2006; 119(6):4065–4073. [PubMed: 16838548]

Denison RN, Driver J, Ruff CC. Temporal structure and complexity affect audiovisual correspondence detection. Frontiers in Psychology. 2012; 3

Dixon NF, Spitz L. The detection of auditory visual desynchrony. Perception. 1980; 9(6):719–721. [PubMed: 7220244]

Eckstein MP, Ahumada AJ. Classification images: A tool to analyze visual strategies. Journal of Vision. 2002; 2(1):i.

Eg R, Behne DM. Perceived synchrony for realistic and dynamic audiovisual events. Name: Frontiers in Psychology. 2015; 6:736.

Elliott R. Simple visual and simple auditory reaction time: A comparison. Psychonomic Science. 1968; 10(10):335–336.

Erber NP. Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech, Language, and Hearing Research. 1969; 12(2):423–425.

Escudier P, Benoît C, Lallouache T. Visual perception of anticipatory rounding gestures. The Journal of the Acoustical Society of America. 1990; 87(S1):S126–S127.

Fiset D, Blais C, Arguin M, Tadros K, Ethier-Majcher C, Bub D, Gosselin F. The spatio-temporal dynamics of visual letter recognition. Cogn Neuropsychol. 2009; 26(1):23–35. [PubMed: 18979274]

Golumbic EMZ, Poeppel D, Schroeder CE. Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. Brain and Language. 2012; 122(3):151–161. [PubMed: 22285024]

Gosselin F, Schyns PG. Bubbles: a technique to reveal the use of information in recognition tasks. Vision research. 2001; 41(17):2261–2271. [PubMed: 11448718]

Gracco V. Timing factors in the coordination of speech movements. The Journal of Neuroscience. 1988; 8(12):4628–4639. [PubMed: 3199197]

Gracco VL, Lofqvist A. Speech motor coordination and control: evidence from lip, jaw, and laryngeal movements. The Journal of Neuroscience. 1994; 14(11):6585–6597. [PubMed: 7965062]

Grant KW, Greenberg S. Speech intelligibility derived from asynchronous processing of auditory-visual information. Paper presented at the AVSP 2001-International Conference on Auditory-Visual Speech Processing. 2001

Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. The Journal of the Acoustical Society of America. 2000; 108(3):1197–1208. [PubMed: 11008820]

Grant KW, Walden BE. Evaluating the articulation index for auditory–visual consonant recognition. The Journal of the Acoustical Society of America. 1996; 100(4):2415–2424. [PubMed: 8865647]

Grant KW, Wassenhove Vv, Poeppel D. Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. Speech Communication. 2004; 44(1):43–53.

Green KP. The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. Hearing by eye. 1998; II:3–26.

Green KP, Norrix LW. Perception of/r/and/l/in a stop cluster: Evidence of cross-modal context effects. Journal of Experimental Psychology: Human Perception and Performance. 2001; 27(1):166. [PubMed: 11248931]

Greenberg S. Speaking in shorthand–A syllable-centric perspective for understanding pronunciation variation. Speech Communication. 1999; 29(2):159–176.

Greenberg S. A multi-tier framework for understanding spoken language. Listening to speech: An auditory perspective. 2006:411–433.

Jesse A, Massaro DW. The temporal distribution of information in audiovisual spoken-word identification. Attention, Perception, & Psychophysics. 2010; 72(1):209–225.

Jiang J, Alwan A, Keating PA, Auer ET, Bernstein LE. On the relationship between face movements, tongue movements, and speech acoustics. EURASIP Journal on Applied Signal Processing. 2002; 11:1174–1188.

Jiang J, Auer ET, Alwan A, Keating PA, Bernstein LE. Similarity structure in visual speech perception and optical phonetic signals. Perception & Psychophysics. 2007; 69(7):1070–1083. [PubMed: 18038946]

Jones JA, Jarick M. Multisensory integration of speech signals: the relationship between space and time. Experimental Brain Research. 2006; 174(3):588–594. [PubMed: 16900363]

Jordan TR, Sergeant P. Effects of distance on visual and audiovisual speech recognition. Language and Speech. 2000; 43(1):107–124.

Jordan TR, Thomas SM. When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audiovisual speech perception. Attention, Perception, & Psychophysics. 2011; 73(7):2270–2285.

Kayser C, Petkov CI, Logothetis NK. Visual modulation of neurons in auditory cortex. Cerebral Cortex. 2008; 18(7):1560–1574. [PubMed: 18180245]

Kim J, Davis C. Investigating the audio–visual speech detection advantage. Speech Communication. 2004; 44(1):19–30.

King A, Palmer A. Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. Experimental Brain Research. 1985; 60(3):492–500. [PubMed: 4076371]

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C. What's new in Psychtoolbox-3. Perception. 2007; 36(14):1.1–16.

Kollia HB, Gracco VL, Harris KS. Articulatory organization of mandibular, labial, and velar movements during speech. The Journal of the Acoustical Society of America. 1995; 98(3):1313–1324. [PubMed: 7560504]

Lander K, Capek C. Investigating the impact of lip visibility and talking style on speechreading performance. Speech Communication. 2013; 55(5):600–605.

Löfqvist A, Gracco VL. Interarticulator programming in VCV sequences: Lip and tongue movements. The Journal of the Acoustical Society of America. 1999; 105(3):1864–1876. [PubMed: 10089609]

Löfqvist A, Gracco VL. Control of oral closure in lingual stop consonant production. The Journal of the Acoustical Society of America. 2002; 111(6):2811–2827. [PubMed: 12083216]

Luo H, Liu Z, Poeppel D. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS biology. 2010; 8(8):e1000445. [PubMed: 20711473]

MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology. 1987; 21(2):131–141. [PubMed: 3594015]

Magnotti JF, Ma WJ, Beauchamp MS. Causal inference of asynchronous audiovisual speech. Frontiers in Psychology. 2013; 4

Maier JX, Di Luca M, Noppeney U. Audiovisual asynchrony detection in human speech. Journal of Experimental Psychology: Human Perception and Performance. 2011; 37(1):245. [PubMed: 20731507]

Massaro, DW. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Erlbaum Associates; 1987.

Massaro DW, Cohen MM, Smeele PM. Perception of asynchronous and conflicting visual and auditory speech. The Journal of the Acoustical Society of America. 1996; 100(3):1777–1786. [PubMed: 8817903]

McClean MD. Patterns of orofacial movement velocity across variations in speech rate. Journal of Speech, Language, and Hearing Research. 2000; 43(1):205–216.

McGrath, M. An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces. University of Nottingham; 1985.

McGurk H, MacDonald J. Hearing lips and seeing voices. 1976

Meredith MA, Nemitz JW, Stein BE. Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. The Journal of Neuroscience. 1987; 7(10):3215–3229. [PubMed: 3668625]

Miller GA, Nicely PE. An analysis of perceptual confusions among some English consonants. The Journal of the Acoustical Society of America. 1955; 27(2):338–352.

Moradi S, Lidestam B, Rönnberg J. Gated audiovisual speech identification in silence vs. noise: effects on time and accuracy. Frontiers in Psychology. 2013; 4

Munhall K, Kroos C, Jozan G, Vatikiotis-Bateson E. Spatial frequency requirements for audiovisual speech perception. Perception & Psychophysics. 2004; 66(4):574–583. [PubMed: 15311657]

Munhall KG, Gribble P, Sacco L, Ward M. Temporal constraints on the McGurk effect. Perception & Psychophysics. 1996; 58(3):351–362. [PubMed: 8935896]

Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E. Visual prosody and speech intelligibility head movement improves auditory speech perception. Psychological Science. 2004; 15(2):133–137. [PubMed: 14738521]

Munhall KG, Tohkura Y. Audiovisual gating and the time course of speech perception. The Journal of the Acoustical Society of America. 1998; 104(1):530–539. [PubMed: 9670544]

Munhall KG, Vatikiotis-Bateson E. Spatial and temporal constraints on audiovisual speech perception. The handbook of multisensory processes. 2004:177–188.

Musacchia G, Sams M, Nicol T, Kraus N. Seeing speech affects acoustic information processing in the human brainstem. Experimental Brain Research. 2006; 168(1-2):1–10. [PubMed: 16217645]

Navarra J, Vatakis A, Zampini M, Soto-Faraco S, Humphreys W, Spence C. Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. Cognitive Brain Research. 2005; 25(2):499–507. [PubMed: 16137867]

Neely KK. Effect of visual factors on the intelligibility of speech. The Journal of the Acoustical Society of America. 1956; 28(6):1275–1277.

Parush A, Ostry DJ, Munhall KG. A kinematic study of lingual coarticulation in VCV sequences. The Journal of the Acoustical Society of America. 1983; 74(4):1115–1125. [PubMed: 6643833]

Poeppel D. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Communication. 2003; 41(1):245–255.

Poeppel D, Idsardi WJ, van Wassenhove V. Speech perception at the interface of neurobiology and linguistics. Philosophical Transactions of the Royal Society B: Biological Sciences. 2008; 363(1493):1071–1086.

Power AJ, Mead N, Barnes L, Goswami U. Neural entrainment to rhythmically presented auditory, visual, and audio-visual speech in children. Frontiers in Psychology. 2012; 3

Rosenblum LD, Saldaña HM. An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology: Human Perception and Performance. 1996; 22(2):318. [PubMed: 8934846]

Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex. 2007; 17(5):1147–1153. [PubMed: 16785256]

Saltzman EL, Munhall KG. A dynamical approach to gestural patterning in speech production. Ecological Psychology. 1989; 1(4):333–382.

Sánchez-García C, Alsius A, Enns JT, Soto-Faraco S. Cross-modal prediction in speech perception. PLoS ONE. 2011; 6(10):e25198. [PubMed: 21998642]

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A. Neuronal oscillations and visual amplification of speech. Trends in Cognitive Sciences. 2008; 12(3):106–113. [PubMed: 18280772]

Schwartz JL, Savariaux C. No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag. PLoS Comput Biol. 2014; 10(7):e1003743. doi:citeulike-article-id:13320829 doi:10.1371/journal.pcbi.1003743. [PubMed: 25079216]

Shams L, Kim R. Crossmodal influences on visual perception. Physics of life reviews. 2010; 7(3):269–284. [PubMed: 20447880]

Smeele, PMT. Perceiving speech: Integrating auditory and visual speech. TU Delft, Delft University of Technology; 1994.

Soto-Faraco S, Alsius A. Conscious access to the unisensory components of a cross-modal illusion. Neuroreport. 2007; 18(4):347–350. [PubMed: 17435600]

Soto-Faraco S, Alsius A. Deconstructing the McGurk–MacDonald illusion. Journal of Experimental Psychology: Human Perception and Performance. 2009; 35(2):580. [PubMed: 19331510]

Stein BE, Meredith MA, Wallace MT. The visually responsive neuron and beyond: multisensory integration in cat and monkey. Progress in brain research. 1993; 95:79–90. [PubMed: 8493355]

Stein BE, Stanford TR. Multisensory integration: current issues from the perspective of the single neuron. Nature Reviews Neuroscience. 2008; 9(4):255–266. [PubMed: 18354398]

Stekelenburg JJ, Vroomen J. Neural correlates of multisensory integration of ecologically valid audiovisual events. Journal of Cognitive Neuroscience. 2007; 19(12):1964–1973. [PubMed: 17892381]

Stevenson RA, Altieri NA, Kim S, Pisoni DB, James TW. Neural processing of asynchronous audiovisual speech perception. Neuroimage. 2010; 49(4):3308–3318. [PubMed: 20004723]

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America. 1954; 26(2):212–215.

Summerfield Q. Articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: Human Perception and Performance. 1981; 7(5):1074. [PubMed: 6457109]

Summerfield, Q. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, editor. Hearing by eye: The psychology of lip-reading. Lawrence Erlbaum Associates; 1987.

Summerfield Q. Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 1992; 335(1273):71–78. [PubMed: 1348140]

Thomas SM, Jordan TR. Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. Journal of Experimental Psychology: Human Perception and Performance. 2004; 30(5):873. [PubMed: 15462626]

Thurman SM, Giese MA, Grossman ED. Perceptual and computational analysis of critical features for biological motion. Journal of Vision. 2010; 10(12):15. [PubMed: 21047747]

Thurman SM, Grossman ED. Diagnostic spatial frequencies and human efficiency for discriminating actions. Attention, Perception, & Psychophysics. 2011; 73(2):572–580.

Troille E, Cathiard MA, Abry C. Speech face perception is locked to anticipation in speech production. Speech Communication. 2010; 52(6):513–524.

Van der Burg E, Cass J, Olivers C, Theeuwes J, Alais D. Efficient visual search from nonspatial auditory cues requires more than temporal synchrony. Temporal multisensory processing and its effects on attention. 2009:63–84.

van Wassenhove V. Minding time in an amodal representational space. Philosophical Transactions of the Royal Society B: Biological Sciences. 2009; 364(1525):1815–1830.

van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A. 2005; 102(4):1181–1186. [PubMed: 15647358]

van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. Neuropsychologia. 2007; 45(3):598–607.10.1016/j.neuropsychologia.2006.01.001 [PubMed: 16530232]

Vatakis A, Maragos P, Rodomagoulakis I, Spence C. Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. Front Integr Neurosci. 2012; 6

Vinette C, Gosselin F, Schyns PG. Spatio-temporal dynamics of face recognition in a flash: It's in the eyes. Cognitive Science. 2004; 28(2):289–301.

Vroomen J, Keetels M. Perception of intersensory synchrony: a tutorial review. Attention, Perception, & Psychophysics. 2010; 72(4):871–884.

Walden BE, Prosek RA, Montgomery AA, Scherr CK, Jones CJ. Effects of training on the visual recognition of consonants. Journal of Speech, Language, and Hearing Research. 1977; 20(1): 130–145.

Wallace MT, Meredith MA, Stein BE. Multisensory integration in the superior colliculus of the alert cat. J Neurophysiol. 1998; 80(2):1006–1010. [PubMed: 9705489]

Yehia H, Rubin P, Vatikiotis-Bateson E. Quantitative association of vocal-tract and facial behavior. Speech Communication. 1998; 26(1):23–43.

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

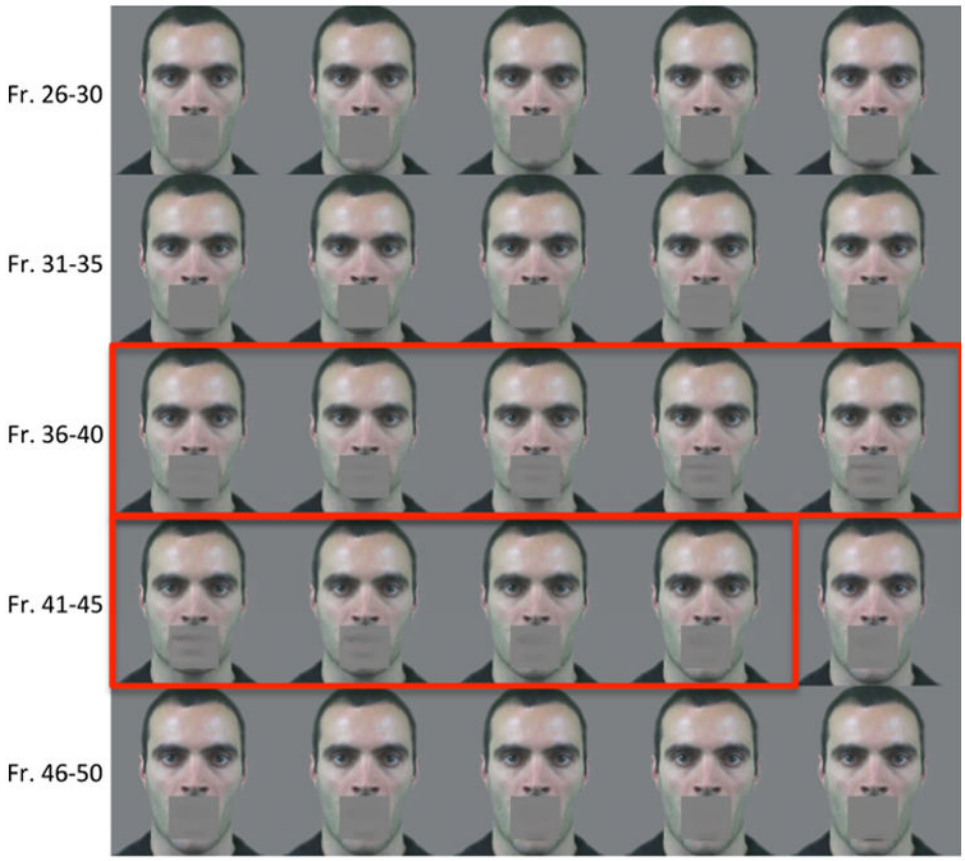

Figure 1. Twenty-five frames from an example Masked-AV stimulus

Masker alpha (transparency) values were spatiotemporally correlated such that only certain frames would be revealed on a given trial. These frames are outlined in red on the example stimulus shown here. Upon close inspection, one can see that the mouth is visible in these frames but not in others. There was a smooth, natural transition between transparency and opacity when movies were presented in real time.

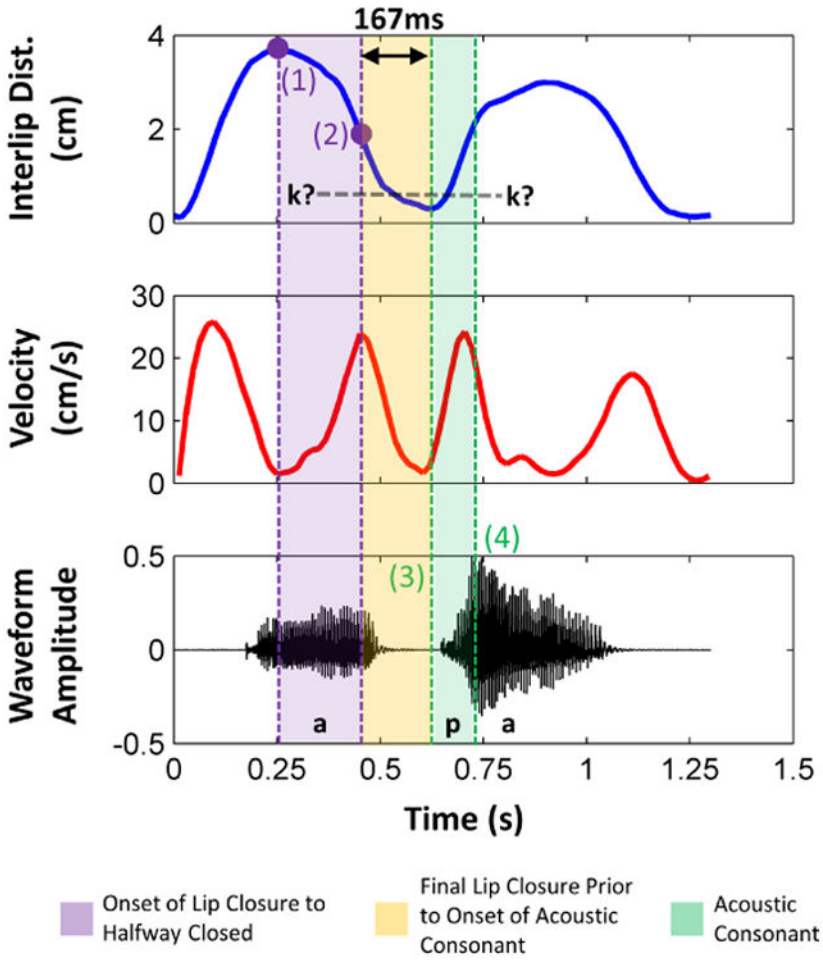

Figure 2. McGurk visual stimulus parameters, calculated following Chandrasekaran et al. (2009)

Pictured are curves showing the visual interlip distance (blue, top) and lip velocity (red, middle). These curves describe the temporal evolution of the visual AKA signal used to construct our McGurk stimuli. Also pictured is the auditory APA waveform used in the SYNC (synchronized) McGurk stimulus. Several features are marked by numbers on the graphs: (1) corresponds to the onset of lip closure during the initial vowel production; (2) corresponds to the point at which the lips were half-way closed at the offset of initial vowel production; (3) corresponds to onset of consonant-related sound energy (3dB up from the trough in the acoustic intensity contour); (4) corresponds to offset of the formant transitions in the acoustic consonant cluster. The time between (2) and (3) is the so-called 'time to voice.' The edges of the purple shaded region correspond to (1) and (2). The edges of the green shaded region correspond to (3) and (4). The yellow shaded region shows the time-to-voice. As shown in the upper panel, visual information related to /k/ may be spread across all three regions.

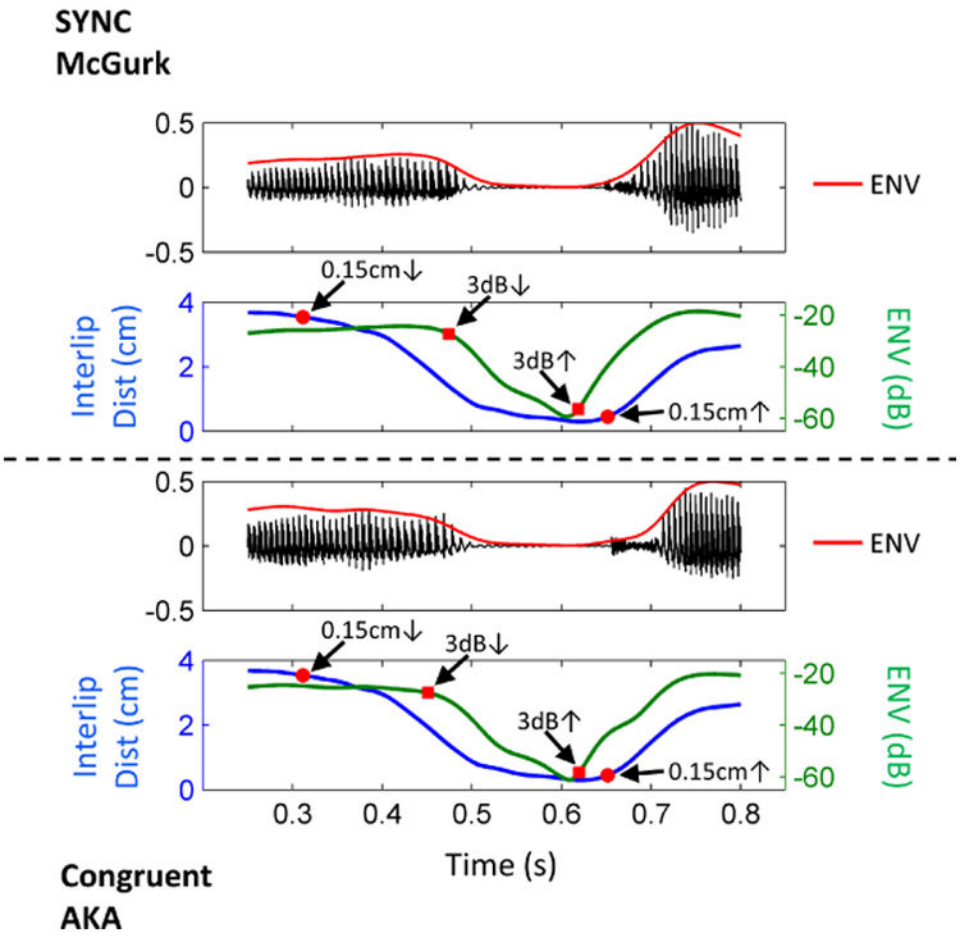

Figure 3. Audiovisual asynchrony in the SYNC McGurk stimulus, calculated following Schwartz and Savariaux (2014)

(Top). Pictured is a trace of the acoustic waveform (black) with the extracted envelope (red). Also pictured are curves showing the visual interlip distance (blue) and the acoustic envelope converted to a decibel scale (green). The time axis is zoomed in on the portion of the stimulus containing the offset of the initial vowel through the onset of the following consonant cluster. Red markers indicate the offset of the initial vowel in the visual (circle, 0.15cm↓) and auditory (square, 3dB↓) signals, and also the onset of the consonant in the visual (circle, 0.15cm↑) and auditory (square, 3dB↑) signals. Audiovisual asynchrony at vowel offset and consonant onset can be calculated by taking the difference (in time) between corresponding markers on the visual and auditory signals. (Bottom). The analysis was repeated the congruent audiovisual AKA from which the visual signal for the McGurk stimulus was drawn.

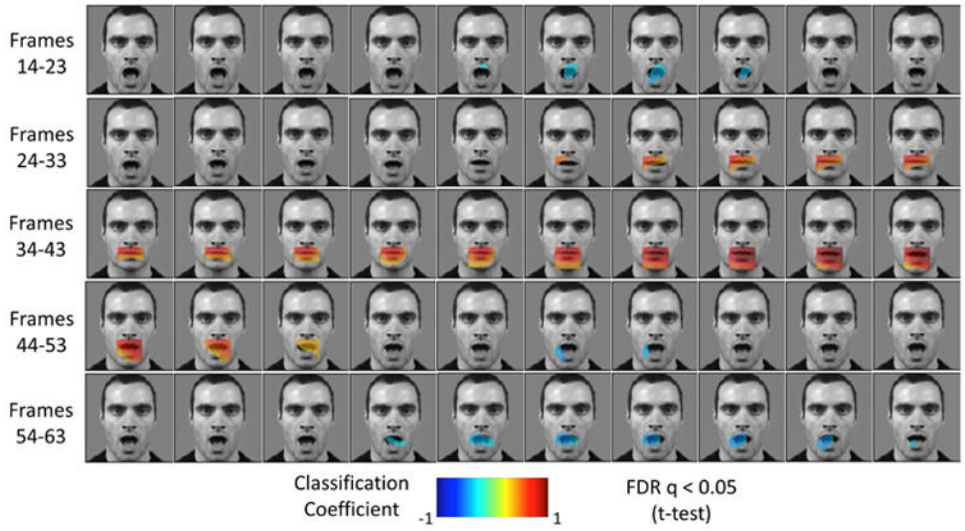

Figure 4. Results: Group classification movie for SYNC

Fifty example frames from the classification movie for the SYNC McGurk stimulus are displayed. Warm colors mark pixels that contributed significantly to fusion. When these pixels were transparent, fusion was reliably observed. Cool colors mark pixels that showed the opposite effect. When these pixels were transparent, fusion was reliably blocked. Only pixels that survived multiple comparison correction at FDR $q < 0.05$ are assigned a color.

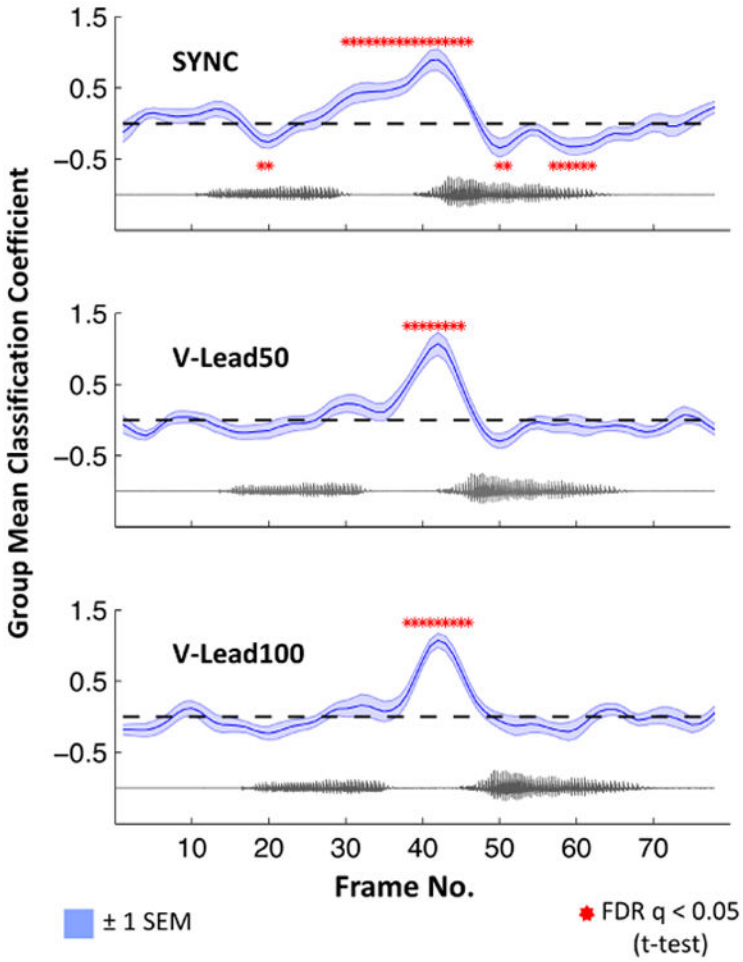

Figure 5. Results: Group classification time-courses for each McGurk stimulus

The group-mean classification coefficients are shown for each frame in the SYNC (top), V-Lead50 (middle), and V-Lead100 (bottom) McGurk stimuli. Significant frames are labeled with stars. These frames contributed reliably to McGurk fusion. Values close to zero (dotted lines) did not reliably influence perception. The waveform of the auditory signal (gray) for each stimulus is plotted beneath the classification time course (blue).

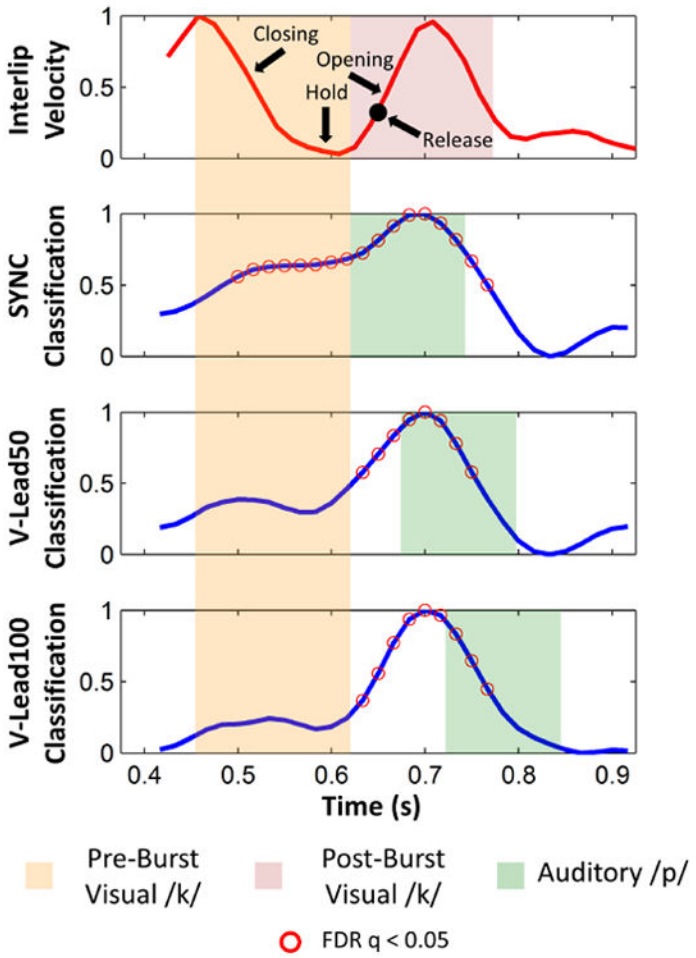

Figure 6. Classification time-courses for the SYNC, V-Lead50, and V-Lead100 McGurk stimuli (blue) are plotted along with the lip velocity function (red)

The figure is zoomed in on the time period containing frames that contributed significantly to fusion (marked as red circles). Classification time-courses have been normalized (max = 1). The onset of the yellow shaded period corresponds to lip closure following the initial vowel, and the offset corresponds to the onset of consonant-related sound energy (3dB up from the trough in the acoustic envelope). We have labeled this the 'pre-burst' visual /k/. Shaded in green is the period containing the auditory consonant /p/ from onset of sound energy to onset of vowel steady state. The green shaded region is shifted appropriately to account for auditory lags in V-Lead50 and V-Lead100. A region on the lip velocity curve is shaded pink. This region corresponds to 'post-burst' visual /k/, as estimated from the classification time-courses. Changes in the oral aperture are labeled (black) on the lip velocity function. The 'release' point marks the time at which interlip distance (not pictured) increased by 0.15cm from the trough at oral closure (note: the release of the tongue from the velum during the production of /k/ may have occurred at a different time point).