Published in final edited form as:

Lang Cogn Neurosci. 2017 ; 32(9): 1102–1118. doi:10.1080/23273798.2017.1283428.

Tolerance for audiovisual asynchrony is enhanced by the spectrotemporal fidelity of the speaker's mouth movements and speech

Antoine J Shahin, Stanley Shen, and Jess R Kerlin

Center for Mind and Brain, University of California, Davis, CA, 95618

Abstract

We examined the relationship between tolerance for audiovisual onset asynchrony (AVOA) and the spectrotemporal fidelity of the spoken words and the speaker's mouth movements. In two experiments that only varied in the temporal order of sensory modality, visual speech leading (exp1) or lagging (exp2) acoustic speech, participants watched intact and blurred videos of a speaker uttering trisyllabic words and nonwords that were noise vocoded with 4-, 8-, 16-, and 32-channels. They judged whether the speaker's mouth movements and the speech sounds were *in-sync* or *out-of-sync*. Individuals perceived synchrony (tolerated AVOA) on more trials when the acoustic speech was more speech-like (8 channels and higher vs. 4 channels), and when visual speech was intact than blurred (exp1 only). These findings suggest that enhanced spectrotemporal fidelity of the audiovisual (AV) signal prompts the brain to widen the window of integration promoting the fusion of temporally distant AV percepts.

Keywords

Audiovisual integration; Audiovisual onset asynchrony; Degraded speech; Spectrotemporal fidelity

1. INTRODUCTION

Speech is central to human communication. Often the speech we hear is degraded or distorted, for instance, during video conferencing. These acoustic adversities are especially daunting for people with hearing loss (HL). The brain, however, employs strategies to offset these challenges. One such strategy involves using visual cues (i.e., speech-reading, also known as lip-reading) to aid comprehension during noisy speech (Bernstein, Auer, & Takayanagi, 2004; Grant & Seitz, 2000; MacLeod & Summerfield, 1987; Middelweerd & Plomp, 1987; Sumbly & Pollack, 1954). However, the link between the information conveyed by mouth movements and the unfolding speech signal is not a simple relationship. This association varies with the type of phoneme and with the complexity of the acoustics and information conveyed by mouth movements (Cappelletta & Harte, 2012; Jesse &

Corresponding author: Antoine J Shahin, UC Davis Center for Mind and Brain, 267 Cousteau Place, Davis CA 95618, ajshahin@ucdavis.edu, Tel: 530 297 4407, Fax: 530 297 4400.

Conflict of Interest statement: The authors declare no conflicts of interest.

Massaro, 2010; Vatakis, Maragos, Rodomagoulakis, & Spence, 2012; Vatakis & Spence, 2006). For example, in visual speech, certain phonemes are more visible than others (e.g., /b/ vs. /k/), and even visible phonemes are not always unique, as /b/ cannot be distinguished from /m/ or /p/ by speech-reading alone. In such cases, other contextual cues (e.g., the rest of the word or sentence) can help constrain the perceptual outcome (Barutchu, Crewther, Kiely, Murphy, & Crewther, 2008; Summerfield, 1992; Tye-Murray, Sommers, & Spehar, 2007; Van Engen, Phelps, Smiljanic, & Chandrasekaran, 2014). For example, prior knowledge of the word “boy” helps distinguish it from the nonwords “poy” or “moy” (Summerfield, 1992).

Despite its shortcomings, speech-reading can be helpful in adverse listening situations and for individuals with hearing loss (Desai, Stickney, & Zeng, 2008; Grant & Seitz, 2000; Grant, Walden, & Seitz, 1998; Kaiser, Kirk, Lachs, & Pisoni, 2003; Sumby, 1954). For instance, Kaiser et al. (2003) examined auditory-only (A-only) and audiovisual (AV) speech perception in single and multi-talker environments in individuals with normal hearing (NH) and those with cochlear implants (CIs). They found that in a single-talker setting, the benefit of visual information to speech perception (AV vs. A-only) was greater in individuals with CIs than in NH listeners. This suggests that speech-reading may help restore phonetic representations of the degraded CI speech. Consistent with this premise, other studies that used speech with missing segments reported that speech-reading enhances the perception of continuity and intelligibility of the degraded speech (Benard & Baken, 2015; Bhat, Pitt, & Shahin, 2014; Grant et al., 1998; Shahin, Bishop, & Miller, 2009; Shahin, Kerlin, Bhat, & Miller, 2012).

Given vision's possible influence on restoration of degraded speech representations, in this study we sought to examine the impact of the spectrotemporal fidelity of audiovisual stimuli on AV integration. The metric of AV integration we examine here is tolerance for audiovisual onset asynchrony (AVOA) in spoken language (Bhat, Miller, Pitt, & Shahin, 2015). In Bhat et al (2015), AVOA tolerance was quantified by measuring individual thresholds of maximum asynchrony at which temporally misaligned mouth movements and corresponding speech sounds were perceived as synchronous. AVOA tolerance can be assessed when visual speech leads acoustic speech, or vice versa, and is rooted in the unity assumption that individuals perceive signals from two sensory modalities that share several qualities as one multisensory event (Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence, 2007; Welch & Warren, 1980). Such perceptual audiovisual unity is particularly robust between mouth movements and speech sounds (Vatakis & Spence, 2007, 2008).

AVOA tolerance is influenced by stimulus duration, loudness, spatial discrepancies and task demands (Boenke, Deliano, & Ohl, 2009; Mégevand, Molholm, Nayak, & Foxe, 2013; Powers, Hillock, & Wallace, 2009; Stevenson, Wilson, Powers, & Wallace, 2013; Vatakis, Navarra, Soto-Faraco, & Spence, 2008; Vatakis & Spence, 2006; Zampini, Guest, Shore, & Spence, 2005). Additionally, it is related to the place and manner of articulation of consonants and the salience of tongue and lip dynamics (front/back tongue, round/unrounded lips) of vowels (Vatakis et al., 2012). One consistent finding is that individuals tolerate greater AVOA when visual events lead auditory events as opposed to the reverse (Bhat et al., 2015; Conrey & Pisoni, 2006; Eg, Griwodz, Halvorsen, & Behne, 2015;

Magnotti, Ma, & Beauchamp, 2013; Miller & D'Esposito, 2005; Stevenson et al., 2013), although this effect tends to be less asymmetrical for speech stimuli (Stevenson et al., 2013). Moreover, this asymmetry effect is not observed in single-modality scenarios (auditory-auditory). Grant, Wassenhove, & Poeppel (2004) showed that two speech tokens that were derived from the same speech stimulus by band-pass filtering at different narrow bands exhibited symmetry in asynchrony tolerance when one token led and lagged the other token. Furthermore, AVOA tolerance is a function of sound complexity (Vatakis & Spence, 2006). For example, listeners are more tolerant of AVOA for music and speech than for other sounds (Dixon & Spitz, 1980; Stevenson et al., 2013; Vatakis & Spence, 2006), and for sentences than syllables (Vatakis & Spence, 2006).

Bhat et al., (2015) proposed a neural model (termed the Dynamic Reweighting Model) for how the AV integration window behaves as a function of the spectrotemporal fidelity of audiovisual stimuli. They posited that the spectrotemporal information of AV stimuli could modulate the neural pathways for processing the speech signal and in turn cause the AV window of integration to resize. For example, as the acoustic speech signal becomes more spectrotemporally meaningful, neural processing is shifted, or reweighted, toward high-level auditory networks associated with phonetic and linguistic level processing and wider windows of integration (Keceli, Okamoto, & Kakigi, 2015). These networks include the superior temporal sulcus-gyrus and middle temporal gyrus (MTG) (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004; Beauchamp, Nath, & Pasalar, 2010; Heinrich, Carlyon, Davis, & Johnsrude, 2008, 2011; Hickok, 2009; Hickok & Poeppel, 2000; Miller & D'Esposito, 2005; Pasley et al., 2012; Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002; Rauschecker & Scott, 2009; Rauschecker, Tian, & Hauser, 1995; Tian, Reser, Durham, Kustov, & Rauschecker, 2001). Concurrent with this upward shift is the weakening of activity at low-level networks in the primary auditory cortex, which encodes simple acoustic features such as acoustic onsets (Shahin et al., 2009). This Dynamic Reweighting Model was further motivated by the results of Bhat et al. (2015) showing reduced auditory evoked potentials to onsets of words that were perceived as synchronous as opposed to asynchronous with the corresponding mouth movements. A key aspect of this reweighting is that it also modifies the connections between the auditory and visual systems. Neural connections strengthen between the visual cortex and high-level auditory networks, and weaken between the visual cortex and low-level auditory networks. The collective outcome is a reduced temporal resolution of the audiovisual system and an increase in tolerance for AVOA, enabling the integration of more temporally distant audio-audio or audio-visual features in spoken language. This reweighting can also be driven by the fidelity of the visual stimulus, such that as the visual stimulus becomes more spectrotemporally informative or meaningful, the visual cortex directs reweighting along the auditory system toward high-level auditory networks leading to widening of the window of AV integration.

We tested the feasibility of the Dynamic Reweighting Model in the current study by manipulating the visual and acoustic spectrotemporal fidelity of audiovisual speech and implementing AV asynchrony in two experiments: visual speech leading acoustic speech (V-A, exp. 1) and acoustic speech leading visual speech (A-V, exp. 2). Subjects watched videos of a speaker uttering trisyllabic words and nonwords. In both experiments, subjects made

judgements on whether they perceived the paired audio and visual stimuli as *in-sync* or *out-of-sync*. AVOA tolerance was assessed based on the percentage of trials that were perceived *in-sync*.

To test for the influence of spectrotemporal fidelity of visual speech on AVOA tolerance, the videos were blurred for half of the trials. The blurring level was chosen to effectively reduce the phonetic information, hence speech intelligibility, provided by fine-grained mouth movements, while still preserve the overall mouth movement onsets and offsets to a large extent (Beauchamp, Nath, & Pasalar, 2010; Dias, Cook, & Rosenblum, 2016). By varying the fidelity of the visual stimuli, we aimed to examine the top-down influence of vision on auditory perception of degraded speech and how that modulates the window of AV integration.

To test for the influence of the spectrotemporal fidelity of acoustic speech, the acoustic signals were noise vocoded at different levels. The advantage of vocoding is that it can degrade the quality of temporal fine structure associated with formant trajectories and transitions (important to phonetic identity) while keeping the speech envelope contour, including amplitude onset information, constant. By varying the acoustic fidelity, we aimed to emulate the acoustic experience of individuals with CIs and to examine how bottom-up representations affect the window of AV integration.

It is important to point out that the main purpose of the above spectrotemporal manipulations (preserving stimulus onset while varying its fine structure information) was to ensure that any changes in AVOA tolerance would not be due to physical differences in stimulus onsets, but rather due to changes in the spectrotemporal fidelity, hence meaningfulness, of stimuli.

In line with the Dynamic Reweighting Model proposed by Bhat et al. (2015), we hypothesized that AVOA tolerance should increase along with increased spectrotemporal fidelity of the audiovisual stimuli. Thus, individuals should perceive more trials as *in-sync* for the intact vs. blurred videos and more acoustic clarity (more vocoding channels) vs. less acoustic clarity.

2. MATERIALS AND METHODS

2.1. Subjects

2.1.1. Main experiments—AVOA tolerance was measured when visual speech led acoustic speech (V-A, exp1) and when acoustic speech led visual speech (A-V EXP2). Eighteen normal hearing, native/fluent English speakers, participated in each of the two experiments (V-A, age: $21.9_{\text{mean}} \pm 3.6_{\text{sd}}$ years, 6 male, 12 right handed, 14 native speakers; A-V, age: $21.8_{\text{mean}} \pm 4.2_{\text{sd}}$ years, 5 male, 16 right handed, 16 native speakers). English fluency, for non-native English speakers, was defined as having spoken English continuously for a minimum of 10 years. All subjects reported normal hearing and normal/corrected vision. Subjects signed an informed consent form in accordance with the ethical guidelines of the Institutional Review Board of the University of California, Davis, and Declaration of Helsinki. Subjects were compensated monetarily for participation.

2.1.2. Supplementary experiments

Response bias to asynchrony: Eight subjects (age: $21.6_{\text{mean}} \pm 2.9_{\text{sd}}$ years, 6 female, 7 right handed, 1 ambidextrous, 6 native speakers) participated in this experiment.

Intelligibility: Eight subjects (age: $21.6_{\text{mean}} \pm 2.9_{\text{sd}}$ years, all female, right handed and native speakers) participated in this experiment.

2.2. Stimuli

Speech material included eighty word and nonword pairs. The words were tri-syllabic nouns/adjectives chosen from the MRC Psycholinguistic Database (Coltheart, 1981) (http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm), excluding unfamiliar words (MRC familiarity rating, or FAM, below 300). For each word, a corresponding nonword was designed by substituting the first two phonemes with other phonemes of the same class (consonant or vowel). For example, the word “messenger” was transformed into “gossenger”. Moreover, the two initial phonemes distinguishing words and nonwords were balanced in terms of phoneme type and place of articulation (POA) to avoid confounding effects on AVOA tolerance (Vatakis et al., 2012). Initial phonemes were consonants in 60 words and 61 nonwords and were vowels in 20 words and 19 nonwords. Secondary phonemes were consonants in 19 words and 20 nonwords, and were vowels in 61 words and 60 nonwords. As for POA, the initial consonants had the following POAs ratios between words and nonwords: 6/6 bilabial (/p/, /b/, /m/, /w/), 5/4 labio-dental (/f/, /v/), 40/42 lingua-alveolar (e.g., /t/, /d/, /s/, /z/, /tʃ/), 9/5 lingua-palatal (/ʃ/, /ʒ/, /r/, /j/), and 1/3 glottal (/h/). The second phone POA was also balanced between words and nonwords.

Video and audio of a professionally trained female vocalist (mean $f_0 = 203$ Hz) uttering the words and nonwords were recorded using a Panasonic AGDVX100A digital camera at 30 frames per second and Shure KSM studio microphone with sampling rate of 48 kHz. Each audiovisual word/nonword was trimmed into an 85-frame video clip using Adobe Premiere (Adobe System, San Jose, CA), ensuring that each clip began and ended in silence with the lips in a still, neutral position. Mono audio and silent video were extracted from each clip and saved separately.

2.2.1. Preparation of acoustic stimuli—We used Matlab R2014b (Mathworks, Natick, MA) to prepare the audio and visual stimuli with in-house code. The audio files were vocoded at four different fidelity levels, using 4-, 8-, 16- and 32-channels, following the general procedure described by Shannon, Zeng, Kamath, Wygonski, & Ekelid (1995). We deemed the 32-channel vocoded speech to be an adequate approximation of intact speech. Low-noise noise (LNN) (Pumplin, 1985) – a type of noise that has minimal envelope fluctuations over time, was used for the carriers. For each word/nonword, the intact signal and a segment of LNN of the same duration were filtered through the same filter bank, designed using the designfilt.m function. The filter bank spanned the range from 80 Hz to 8000 Hz, with frequency bands equally spaced on a logarithmic scale. For each band, a Butterworth bandpass filter was created with minimum order that could satisfy the specified frequency response: lower and higher stopbands 25 dB below the passband, with less than 1 dB passband ripple. The Hilbert envelope of the bandpass filtered speech signal was

extracted using the Hilbert transform (Léger, Reed, Desloge, Swaminathan, & Braid, 2015) followed by lowpass filtering at 20 Hz with a 15th order Butterworth filter, before it was applied to the bandpass filtered LNN. The amplitude-modulated LNN was filtered again to eliminate energy leakage outside the designated band. In each step described above (and below), to ensure zero phase shift, signals were filtered in both forward and reverse directions. This series of cascaded filtering resulted in a slope as steep as 1400 dB/octave in the transition region, avoiding spectral overlap between neighboring bands, hence preserving the intelligibility of signals (Moore, 2007). The root mean square (RMS) of the filtered noise in each band was normalized to that of the filtered speech signal in the same frequency band, before the RMS normalized signals in all the bands were summed up to create the vocoded speech.

After vocoding, the overall envelope of the vocoded signal was replaced with that of the original speech signal through point-by-point division and multiplication. Both envelopes were extracted in the same way: first, the Hilbert envelope was computed, before being filtered with a 15th order Butterworth LP filter with cutoff frequency at 23 Hz; second, to eliminate negativity, the signal was full-wave rectified and filtered again through a 10th order Butterworth LP filter with cutoff frequency at 46Hz; finally, the processed signal was upshifted by a constant of 0.005 (empirically selected to fit the audio signals) to avoid numerical issues - discontinuity when calculating the reciprocal of a tiny value. This additional step of envelope replacement eliminated possible confounding effects due to the distortion of the speech envelope caused by vocoding. As a result, the envelopes of the stimuli in the four conditions were almost identical.

The following measures were taken to ensure identical onset timing of the acoustic signals across different fidelity levels: prior to vocoding, zero padding of 1000 samples (21 ms) was applied to the start and the end of the original speech signal, which also alleviated potential edge effect of filtering; subsequent to envelope replacement, sine-squared ramps were applied to where zero padding was previously applied (phase ranging from 0 to $\pi/2$ at the beginning of the signal, and from $\pi/2$ to π at the end). Note that the zero padding samples on the edges would always become non-zero after filtering, due to the Butterworth filters used in vocoding.

After each single audio file was processed, the entire set of audio stimuli was normalized to the same dBA level with the equal loudness contour option in Adobe Audition 3.0 (Adobe Systems, San Jose, CA).

Note audio processing was implemented on the non-zero utterance part of the original audio file, which contained silence at both ends. The finalized audio files would maintain the same length and the same perfect alignment with the matching video files.

2.2.2 Preparation of visual stimuli—The visual stimuli had two fidelity levels, intact (clear) and blurred. The intact videos were cropped to the size of $504_{\text{width}} \times 192_{\text{height}}$ pixels such that the mouth was centered and the eyes excluded (Figure 1). For the blurred videos, the intact videos were filtered with a Gaussian kernel (filter size = 49×49 pixels, standard deviation = 12 pixels) using the Matlab function `fspecial.m`[^{note}]. This blurring level was

deemed by the experimenters to appropriately preserve the global movements of the mouth, such as opening and closing, while minimizing the finer spatial details, such as placement of the tongue relative to the teeth, thus degrading the phonetic cues. Though the level of blurring was designed to limit the impact of blurring on the distortion of mouth movement onsets, we acknowledge that full preservation of visual temporal information is not possible with any level of blurring.

2.3. Equipment and Procedure

2.3.1. Main experiments (V-A and A-V)—Individuals sat about 85 cm in front of a 24-inch Dell monitor with a sound bar (Vizio, model S2920W-C0) below the monitor. Only one loudspeaker of the sound bar was activated, thus the sound bar was horizontally shifted so that this loudspeaker was placed in the center. The videos were presented using Presentation Software (v.18.1, Neurobehavioral Systems, Albany, CA). The Audio card used was RME Hammerfall HDSP 9652 PCI ADAT Audio Interface, tested to ensure a consistent < 2 ms latency when using Exclusive Mode in Presentation. The video card used was NVIDIA Quadro K620 2GB. The onset of each acoustic stimulus was time-locked relative to the expected monitor refresh (“flip”) of the first frame of the video, with an accuracy of 1 ms. The timing accuracy of acoustic and video onsets output by Presentation were determined using a Tektronix TBS1052B oscilloscope connected to a microphone and a photodiode while presenting click and flash dummy stimuli. The size of the video on the monitor was 27.5 cm horizontal (15 cm actual ear to ear width) and 10 cm vertical.

Figure 1 depicts the experimental design of a trial. Subjects watched videos of the speaker and made synchrony judgments using their left middle and index fingers on whether they perceived the mouth movements and sounds as *in-sync* or *out-of-sync*. The subjects were explicitly instructed that the words could have degraded quality and that their task was to make their synchrony judgements regardless of whether they understood the words. They were instructed to delay their response until they saw a green arrow below the video. The green arrow appeared 2700 ms after the start of the video, and lasted until the end of the trial. Requiring participants to delay their response ensured that the whole utterance of the visual and acoustic speech was available before the subject reported a decision. Within each experiment, all the trials were presented with a fixed AVOA, with no synchronous trials. In the case when visual speech led acoustic speech (V-A, Exp 1) the acoustic signal was always delayed by 360 ms with respect to the natural recorded audiovisual alignment. When acoustic speech led visual speech (A-V, Exp 2) the acoustic signal always led by 314 ms. The misalignment values for these tasks were based on the 50% synchronous thresholds reported by (Bhat et al., 2015). They used 230 monosyllabic intact word videos and an adaptive procedure (one-up one-down, 2-alternative forced choice, response) to determine the average threshold for each temporal order of AV events (V-A or A-V) across all participants.

Accurate AVOA timing (± 1 ms) was realized by: (1) time-locking each stimulus onset to the start of a particular refresh period of the monitor; (2) adjusting for audio and video hardware

note ¹The designation of parameters followed this heuristic formula as implemented in Matlab R2015a by default: $hsize = 2 \times \text{ceil}(2 \times \sigma) + 1$, so that visual fidelity decreases monotonically as σ increases.

delays, (3) adopting the parallel stimulus event function in Presentation, which coordinates the timing of audio and video stimuli in parallel. Logfiles, combined with hardware testing, confirmed effective timing control. Note, despite accurate timing control of the audio onset relative to the video onset, the resolution of the visual stimulus was inherently limited by the video recording rate of 30 Hz (33.3 ms/frame).

The experiment consisted of 8 blocks (14 min each) with each block containing 160 trials presented in an event-related design. For each subject and prior to the experiment, pseudo-random stimulus sequences (word/nonword token $\times$ visual condition $\times$ acoustic condition) for the 8 blocks were generated collectively using in-house Matlab code. The algorithm utilizing the logic of circulant matrices ensured that: (1) Trials in all the $2 \times 2 \times 4$ combinatorial conditions were available in each block; (2) different levels of each independent variable (word/nonword, visual fidelity and acoustic fidelity) were optimally counterbalanced in each block; (3) each word/nonword token appeared in each audiovisual condition (2×4) exactly once throughout the experiment, with 8 occurrences spread across 8 blocks.

Each trial was about 5.1 seconds in duration, starting with a static picture of the first frame extracted from the video, and ending with the static picture of the last frame after the video stopped. Naturally, the onsets of mouth movements and acoustic signals were temporally variable across words and nonwords with respect to the start of the video, which was further jittered by ± 150 ms relative to the trial onset. Individuals had around 2 seconds to respond following AV word presentations. The words had an average sound intensity level of 57.3 dBA and average peak sound intensity level of 62.4 dBA (slow time weighting). These measurements were made using a Larson Davis SoundTrack LxT1 sound level meter situated at head level 85 cm from the monitor, based on the mean level of 40 randomly selected words and nonwords (concatenated into a long audio file and played through Presentation). Responses were logged by Presentation software. Participants took a few minutes' break between blocks.

The words and nonwords had average durations of $996 \pm 108_{\text{sd}}$ ms and $1003 \pm 147_{\text{sd}}$ ms, respectively. A paired t-test revealed no significant difference between their durations ($t_{(79)} = 0.55$, $p = 0.58$).

2.3.2. Supplementary experiments

Response bias to asynchrony: This supplementary experiment was conducted to examine whether the AVOA tolerance results of the main experiments were mainly driven by a bias toward *in-sync* responses due to differences in intelligibility of auditory and/or visual stimuli (see next section). To probe the potential bias, using the similar audiovisual setting as those of the main experiment 1 (V-A condition), we used two reference asynchrony levels in addition to the previous level (360 ms) – perfectly synchronous (0 ms shift associated with original recorded alignment) and extremely asynchronous (550 ms shift). If there was no criterion shift between conditions (e.g., intact vs. blurred, 4 channel vs. 16 channels), we expected to observe near-ceiling percentages of *in-sync* responses for the perfectly synchronous trials, differential percentages for the ambiguous trials (360 ms), and near-bottom percentages for the extreme asynchronous (550 ms) trials. The methods for this task

also differed from main experiment 1 in that only the real words ($n = 80$) and two vocoding levels (4- and 16- channels) were used. The stimuli were presented in an event-related design in 12 blocks, with an additional practice session of 10 trials (0 ms and 550 ms delay only). The task was identical to the main experiments (respond *in-sync* or *out-of-sync*)

Intelligibility: This experiment was conducted to determine the baseline intelligibility of the AV word stimuli ($n = 80$, 2 visual fidelity levels, $\times 4$ acoustic fidelity levels, as described in the main experiments). Participants in this task had not previously been exposed to the word stimuli. All 80 audiovisual words were presented with zero ms AV shift to each participant, with ten words assigned to each condition (8×10). Participants were asked to type the word they heard after each trial. Participants pressed Enter to view their typed response, and the “j” key to retype the response if a typo was made. Pressing Enter twice would finalize the response and start the next word presentation. Words were presented only once to each participant to avoid memory-related judgements. The words assigned to each condition rotated across subjects such that each word was presented in each of the 8 conditions over the course of the experiment.

2.4. Data Analysis

2.4.1. Main experiments—Each subject’s *in-sync* and *out-of-sync* responses, collapsed across 8 blocks, were segregated into 16 conditions that comprised all combinations of the experimental parameters: word/nonword, intact/blurred, and 4 levels of vocoding. AVOA tolerance was calculated as the percentage of *in-sync* trials relative to all trials (*in-sync* and *out-of-sync*) within each of the 16 conditions, excluding trials without a valid response. For trials with more than one response, only the first response was considered valid.

2.4.2. Supplementary experiments

Response bias to asynchrony: Each subject’s *in-sync* and *out-of-sync* responses, were segregated into three asynchrony levels and four conditions (intact/blurred and 4-/16-channel). AVOA tolerance was calculated the same way as the main experiments.

Intelligibility: The accuracy of the typed response to each trial (Correct or Incorrect) was judged by a lab member, blinded to the fidelity manipulation of each trial. Responses were judged as correct if the spelling of the response closely matched the phonetic spelling of the target word, and the phonetics did not match a known alternative word, e.g., a response of “dissresion” for the target “discretion” would be marked correct, but “dissection” would be marked as wrong. Percent mean accuracy was calculated for each subject as the number of correct responses for each condition divided by the total number of word presentations for that particular condition.

2.5. Statistics

2.5.1. Main experiments—Effects on AVOA tolerance (*in-sync* percentages) were evaluated by a $2 \times 2 \times 4$ repeated-measure ANOVA with the independent variables being lexicality (word/nonword), visual fidelity (intact/blurred) and acoustic fidelity (vocoding level), respectively. Separate ANOVAs were conducted for exp1 (V-A) and exp2 (A-V), because these two tasks were performed by different pools of subjects. Post hoc analyses

were based on Scheffe's test. Only contrasts with p-values of 0.05 or less were considered significant. p-values of the main ANOVA (main effects and interactions) were adjusted for sphericity violations using the Greenhouse-Geisser method. Effect sizes are denoted by partial eta square (η_p^2). Statistics were performed by Statistica v.12 (Dell Software, Tulsa OK).

2.5.2. Supplementary experiments

Response bias to asynchrony: Effects of AVOA tolerance (*in-sync* percentages) were evaluated by a $3 \times 2 \times 4$ repeated-measure ANOVA with the independent variables being asynchrony level (0, 360, 550 ms), visual fidelity and acoustic fidelity. All other statistics conform with those of the main experiments (section 2.5.1).

Intelligibility: Intelligibility was evaluated by a 2×4 repeated-measure ANOVA with the independent variables being visual fidelity (intact/blurred) and acoustic fidelity (4-, 8-, 16-, 32 channel). All other statistics conform with those of the main experiments.

3. RESULTS

3.1. Main experiments

3.1.1. Visual speech leads acoustic speech (V-A, main Exp 1)—Results of this experiment are shown in Figure 2. The ANOVA revealed main effects of lexicality ($F_{(1, 17)} = 6.2$, $p = 0.023$; $\eta_p^2 = 0.26$), visual fidelity ($F_{(1, 17)} = 35.6$, $p = 0.000015$; $\eta_p^2 = 0.68$) and acoustic fidelity ($F_{(3, 51)} = 18.4$, $p = 0.00018$; $\eta_p^2 = 0.52$). There were no interactions between the variables ($p > 0.05$). The lexicality effect was attributed to greater *in-sync* perception (AVOA tolerance) for words than nonwords (13 out of 18 subjects exhibited this effect). The visual fidelity main effect was attributed to greater *in-sync* perception for the intact than blurred videos (17 out of 18 subjects exhibited this effect). Finally, the acoustic fidelity main effect was attributed to greater *in-sync* perception for the 8-, 16- and 32-channel relative to the 4-channel acoustic stimuli (Scheffe's test $p < 0.005$; Figure 2). Post hoc tests did not reveal differences between the 8-, 16- and 32-channel vocoded stimuli ($p > 0.25$). Note that 16 or more out of 18 subjects showed this effect of acoustic fidelity (16/18 for the contrast between 8-,16- vs. 4-channels and 17/18 for the contrast between the 32- and 4-channels).

3.1.2. Acoustic speech leads visual speech (A-V, main Exp 2)—Results of this experiment are shown in Figure 3. The ANOVA revealed a main effect of lexicality ($F_{(1, 17)} = 4.52$, $p = 0.048$; $\eta_p^2 = 0.21$), a main effect of visual fidelity, ($F_{(1, 17)} = 6.7$, $p = 0.02$; $\eta_p^2 = 0.28$), and a main effect of acoustic fidelity ($F_{(3, 51)} = 8.4$, $p = 0.005$; $\eta_p^2 = 0.33$). There was also an interaction between all the variables ($F_{(3, 51)} = 3.8$, $p = 0.015$; $\eta_p^2 = 0.18$). The lexicality main effect occurred because subjects exhibited greater *in-sync* perception (AVOA tolerance) for words than nonwords (12 out of 18 subjects exhibited this effect). The visual fidelity main effect reflected greater *in-sync* perception for blurred than intact videos (13 out of 18 subjects exhibited this effect). Note that this result is opposite to that observed when visual speech led acoustic speech (Figure 2). The acoustic fidelity effect occurred because subjects exhibited greater *in-sync* perception for the 8-, 16- and 32-channel than for the 4-

channel acoustic stimuli (Scheffe's test $p < 0.05$; Figure 3). Post hoc tests did not reveal a difference between the 8-, 16- and 32-channel stimuli ($p > 0.8$). Note that 13 or more out of 18 subjects showed this effect of acoustic fidelity (15/18 for the contrast between 8- vs. 4-channels, and 13/18 for the contrast between the 16- and 4-channels and 14/18 for the contrast between the 32- and 4-channels).

The interaction between all variables was partly driven by the contrast between the 8-channel words and nonwords, in which individuals exhibited a stronger *in-sync* perception for words when the videos were intact than blurred.

3.1.3. Adaptation and fatigue—One unintended consequence of conducting studies probing AVOA tolerance is the effect of adaptation (Bhat et al., 2015; Navarra et al., 2005; Van der Burg, Alais, & Cass, 2013). When individuals are exposed to more asynchronous AV stimuli, they become more tolerant to AVOA. This certainly applies to the current study. Also, fatigue during the 2-hour long task may have affected the outcome. To address these concerns – whether individual adaptation/fatigue may have skewed the results – we divided the individual sessions of each of the main experiments into the first and the second parts (4 blocks each). First, to probe whether there was a shift due to adaptation or fatigue we conducted similar ANOVAs similar to the main experiments (sections 3.1.1 and 3.1.2) with the added variable adaptation/fatigue (ANOVA variables: 2 adaptation/fatigue $\times$ 2 lexicality $\times$ 2 visual fidelity $\times$ 4 acoustic fidelity). Indeed, there was a main effect of adaptation/fatigue that led to greater AVOA tolerance in the second part of the experiment compared to the first part (Figure 4, V-A: $F_{(1, 17)} = 5.31$, $p = 0.034$; $\eta_p^2 = 0.23$; A-V: $F_{(1, 17)} = 5.33$, $p = 0.034$; $\eta_p^2 = 0.24$). Note the close results of the V-A and A-V conditions, which suggests a consistency of adaptation. All other main effects were identical to those from previous ANOVAs (sections 3.1.1 and 3.1.2).

Second, to assess whether the previous results would hold despite adaptation/fatigue, separate ANOVAs were conducted for the first and second parts of the main experiments and for the V-A and A-V conditions. 1) For the first half of the V-A condition, the main effect of lexicality was marginal ($p = 0.073$), but the main effects of visual fidelity and acoustic fidelity were significant ($p < 0.001$) (compare to Fig. 2); 2) For the second half of the V-A condition the main effects of lexicality ($p < 0.01$), visual fidelity and acoustic fidelity ($p < 0.001$), observed in Figure 2, were all significant; 3) For the first half of the A-V condition, the main effect of lexicality was marginal ($p = 0.066$), but the main effects of visual fidelity ($p < 0.05$) and acoustic ($p < 0.001$) fidelity were significant (compare to Fig. 3); 4) For the second half of the A-V condition, the main effects of lexicality ($p < 0.05$), and acoustic fidelity ($p < 0.001$) were significant, but the main effect of visual fidelity was marginal ($p = 0.058$) (compare to Fig. 3).

In short, while the significance of a few main effects became marginal on either side of the split, the main effects of acoustic and visual fidelity, and lexically were all present both early and late in each of the main experiments (3.1.1 and 3.1.2). Note, the marginal effects observed for lexicality (V-A and A-V) in the first part and for visual fidelity (A-V contrast) in the second part of the experiments may also be attributed to the reduced signal-to-noise ratio due to apportioning the experiment sessions (fewer trials).

3.2. Supplementary experiments

3.2.1. Response bias to asynchrony—An ANOVA revealed main effects of asynchronous level ($F_{(2, 14)} = 26.5$, $p = 0.00056$, $\eta_p^2 = 0.79$), visual fidelity ($F_{(1, 7)} = 14.8$, $p = 0.006$, $\eta_p^2 = 0.67$), and acoustic fidelity ($F_{(1, 7)} = 7.8$, $p = 0.026$, $\eta_p^2 = 0.52$). These main effects are consistent with the results of the main experiment (exp1), a greater *in-sync* perception for the smaller vs. larger asynchronous levels (0 ms vs. 360 or 550 ms), for intact vs. blurred videos, and for 16-channel vs. 4-channel vocoded words. However, the purpose of this experiment was to establish an interaction between the asynchronous level and visual and acoustic fidelity. Indeed, there were interactions between asynchronous level and visual fidelity ($F_{(2, 14)} = 4.97$, $p = 0.029$, $\eta_p^2 = 0.41$) and acoustic fidelity ($F_{(2, 14)} = 4.3$, $p = 0.051$, $\eta_p^2 = 0.38$). These interactions are attributed to a greater *in-sync* perception for the intact and 16-channel stimuli than the blurred and 4-channel stimuli, respectively, occurring only for the 360 ms asynchronous level (Figure 5). These results lessen the likelihood that the original results (Figs. 2 & 3) are mainly driven by a bias to respond *in-sync* to the more intelligible stimuli (see next section).

3.2.2. Intelligibility—As the spectrotemporal fidelity of AV stimuli increases, so does intelligibility. This experiment was conducted to demonstrate this effect, which should implicate intelligibility as a factor in modulating AVOA tolerance. An ANOVA revealed main effects of visual fidelity ($F_{(1, 7)} = 13.2$, $p = 0.0084$, $\eta_p^2 = 0.65$) and acoustic fidelity ($F_{(1, 7)} = 15.3$, $p = 0.001$, $\eta_p^2 = 0.68$), that were due to enhanced intelligibility for the more spectrotemporally intact visual and acoustic stimuli (Figure 6). Note the similarity between the intelligibility results and the results of the main experiment 1 (Fig. 2).

4. DISCUSSION

4.1 Theory and brief summary of the results

In this study, we examined the influence of spectrotemporal fidelity of audiovisual speech on AVOA tolerance. The Dynamic reweighting Model of Bhat et al. (2015) reasons that an increase in the spectrotemporal fidelity of the image of the speaker's mouth movements or the spoken words will render the AV signal more speech-like and induce a reweighting (shift) in neural processing from low- to high-level, speech-specific auditory networks. This upward reweighting in turn will cause an expansion of the window of integration (via exciting high-level auditory networks) and the reduction in sensitivity to acoustic onsets (via inhibiting low-level auditory networks), resulting in reduced temporal resolution of the auditory system and greater tolerance for AVOA. This theory was motivated by extensive prior work (Dixon & Spitz, 1980; Heinrich et al., 2011; Hickok & Poeppel, 2000; Okada et al., 2010; Patterson et al., 2002; Rauschecker et al., 1995; van Wassenhove, Grant, & Poeppel, 2005; van Wassenhove, Grant, & Poeppel, 2007; Vatakis & Spence, 2006; Wessinger et al., 2001). The current findings provide some support for this model. Under both the V-A and A-V conditions (temporal order of AV events), enhanced spectrotemporal fidelity of acoustic speech led to an increase in AVOA tolerance, as predicted. However, although clear visual information led to greater AVOA tolerance than blurred visual information when visual speech led acoustic speech (consistent with our predictions), the opposite occurred when acoustic speech led visual speech (opposite to our prediction). Thus,

detailed, phonetically informative aspects of visual speech may only be influential proactively, rather than retroactively, in the promotion of wider windows of AV integration.

4.2 Contrasts with previous accounts

To our knowledge, there is only one report (Eg et al., 2015) that examined the influence of acoustic speech quality on perception of AV synchrony and three prior reports, all recent, that examined the effect of blurring (Eg & Behne, 2015; Eg et al., 2015; Magnotti et al., 2013). The results from these studies are mixed, but we choose to discuss two reports (Eg et al., 2015; Magnotti et al., 2013), because they provided evidence that may seem inconsistent with our hypothesis. We partly address them by providing theoretical explanations of the inconsistencies.

Eg et al. (2015) used videos of a news anchor reading news while varying the intensity of background pink noise. They examined AVOA tolerance for various asynchrony values, when acoustic led and lagged the visual events. Individuals perceived synchrony more often for the noisiest condition vs. the less noisy conditions when the acoustic event led the visual events (opposite our findings). The different outcomes of the study by Eg et al. (2015) and our study may be partly attributed to differences in the way the acoustic quality was degraded. In their study, the background noise distorted not only the speech as a whole but also the acoustic onsets of the speech. A condition with less audible acoustic onsets may lead to reduced sensitivity (precision) to onset timing and hence render greater tolerance for AVOA. In our design, the audibility of the acoustic onsets for all levels of degradation (vocoding) remained constant, and thus the differences in AVOA tolerance we report here are more likely facilitated by enhanced spectrotemporal fidelity of the stimuli.

The study by Magnotti et al. (2013) used monosyllabic words to assess the influence of blurring on asynchrony judgments for various asynchrony values. They found that individuals were more tolerant of AV asynchrony when visual speech was blurred than intact for both the V-A and A-V conditions. This is opposite our results for the V-A (visual speech led acoustic speech) condition, but consistent with our results for the A-V condition. They explained that physically blurring visual speech causes an increase in the observer's uncertainty of timing of the visual onsets, leading to more tolerance for AV asynchrony. Despite our contradictory results, their explanation can be reconciled with what we advocate. In accordance with the Dynamic Reweighting Model of Bhat et al. (2015), AVOA tolerance increases with weakened sensitivity to acoustic and/or visual onsets. This can be achieved: 1) physically, by reducing the clarity of mouth movements onsets or acoustic onsets (Eg et al., 2015; Magnotti et al., 2013); or 2) perceptually, by keeping stimulus onset clarity fairly intact while varying the overall spectrotemporal fidelity, hence phonetic clarity, of the unfolding visual and acoustic speech (current study). We propose that the brain suppresses sensitivity to stimulus onsets when relevant information (phonetic) prevails over less useful information (onset information) due to enhanced stimulus spectrotemporal fidelity.

Another plausible theory is that AVOA tolerance is enhanced by a more dynamic speech signal (Vatakis & Spence, 2007, 2008). Magnotti et al. (2013) used monosyllabic words, where mouth movement onsets represent a major event in the visual speech stream. In more

complex speech (trisyllabic words herein, or sentences) mouth movements reoccur. Thus, fusion of visual speech with acoustic speech may not be limited to onset information but may occur continually as the word or sentence unfolds. We reason that while coupling an acoustic onset to the first movement of the mouth might be less precise for blurred than intact visual speech (rendering greater AVOA tolerance), this coupling uncertainty breaks down once a second and a third movement of the mouth unfold. Relative to intact visual speech, individuals become more certain that visual and acoustic speech in blurred videos are asynchronous during more complex and dynamic AV speech. However, our results suggest that this effect may only be valid when visual speech leads acoustic speech (Eg & Behne, 2015).

4.3 Comparison with other models

The Dynamic Reweighting Model overlaps with and builds on the Predictive Coding Model in audiovisual research (Arnal, Wyart, & Giraud, 2011; Poeppel, Idsardi, & van Wassenhove, 2008; van Wassenhove, Grant, & Poeppel, 2005). The Predictive Coding Model posits that preceding visual input creates abstract internal representations that constrain auditory expectations, including phonetic representations. Through computations in multisensory networks (e.g., superior temporal gyrus), auditory inputs are evaluated against these abstract representations. The more correlated the auditory and abstract signals are to each other, the faster the processing. Thus, a more salient visual input is a stronger predictor of auditory input, leading to an accelerated auditory processing (van Wassenhove et al., 2005). This effect is evidenced in the reduction in the latency of the N1/P2 auditory evoked potential when synchronous visual speech precede speech sounds (van Wassenhove et al., 2005). The Dynamic Reweighting Model (Bhat et al., 2015) is based on the same principle of visually driven prediction of auditory processing, with the exception that it hypothesizes that vision directly/indirectly reinforces the encoding of phonetic representations (as opposed to evaluation against abstract representations in multisensory networks), via weighted inhibition and excitation of auditory networks.

The Dynamic Reweighting Model postulates that by suppressing the neural response to acoustic onsets, individuals become less aware of the lag between acoustic and visual speech and thus recalibrate (increase) their windows of AVOA tolerance. Other models of AV integration, however, propose that detection of lags across the senses is an integral step in a process that the brain uses to recalibrate and restore perception of synchrony (Freeman et al., 2013; Parise & Ernst, 2016). Freeman et al. (2013) further posit that recalibration of temporal thresholds of maximum asynchrony perception or AV integration (e.g., McGurk illusion) depends on the average lags amongst all neural processes involved during the task. Because different processes recruit different neural networks with different “clocks”, the AV asynchronies tolerated in one task can be considerably different from those in other tasks.

4.4. AVOA Temporal Asymmetry: Speech Intelligibility and lexicality

The influence of visual fidelity on enhancement of AVOA tolerance was only robust when visual speech led acoustic speech. While broad variability of asynchrony exists in everyday speech (Schwartz & Savariaux, 2014), phonetically informative mouth movements often precede acoustic speech (Bhat et al., 2015; Navarra et al., 2005; Van der Burg et al., 2013).

Because of this attribute, the asymmetry of AVOA tolerance may have evolved to maximize the influence of preceding visual content on speech intelligibility. In support, prior studies found that intelligibility of spectrally (i.e., formant) degraded speech (Grant & Greenberg, 2001) and the perception of the McGurk illusion (Munhall & Tohkura, 1998; van Wassenhove et al., 2007) were more resilient for larger AV asynchronies when visual speech led acoustic speech than the reverse. As the speed of speech increases, the importance of visual precedence for intelligibility becomes even more pronounced (Brungart, Iyer, Simpson, & Wassenhove, 2008), suggesting that this asymmetry is tuned to allow maximal comprehension of rapid conversational speech. This asymmetrical resizing of AV integration window may proximally result from separate auditory and visual initiated processes (Munhall & Tohkura, 1998; Cecere, Gross, & Thut, 2016; van Wassenhove, 2013) – a view consistent with the Dynamic Reweighting Model.

In our results, the influence of lexicality (word vs. nonword) on AVOA tolerance was modest, compared to the influence of auditory degradation (4- vs. 8-channel). While nonwords had no lexical content, the 4-channel words were often unintelligible, thus decreases in lexical content cannot be the primary cause of decreases in AVOA tolerance. Furthermore, visual clarity, when visual speech led acoustic speech, increased AVOA tolerance for both words and nonwords equally. This suggests that the wider window of AV integration found with clear visual speech results primarily from enhancement at the level of phonetic encoding (formant transition/trajectory), prior to the establishment of word meaning. Modest increase in AVOA tolerance from speech with lexical content may come from increased predictability, and thereby encoding, of incoming low level phonetic information. This interpretation coincides with the more general view that visual enhancement of spoken language perception occurs primarily via the perceptual rather than the linguistic (e.g., lexical) level (Samuel & Lieblich, 2014) (if we stipulate that phonetic cues represent perceptual objects).

4.5. Is speech special?

A debatable aspect of the current topic is whether the mechanisms observed (e.g., tolerance for AV asynchronies) can be attributed to certain special characteristics of speech (Arnal et al., 2011; Poeppel et al., 2008; van Wassenhove et al., 2005). That is, whether increased AVOA tolerance is distinctively useful during perception of spoken language, as opposed to perception of non-speech stimuli (we omit the special case of music herein, because it is beyond the scope of this article). This is grounded in the premise that spectrotemporally complex speech necessitates wider windows of audiovisual integration – to attain phonological categorization for example – than simpler non-speech sounds, for which onset information may be adequate (van Wassenhove et al., 2005; van Wassenhove, 2013; van Wassenhove, Grant, & Poeppel, 2007). Indeed, for the purpose of movement and spatial orientation, precise perception of AV asynchrony may be important for non-speech sounds, such as a flash light and a tone pairing (Stevenson et al., 2013; van Eijk, Kohlrausch, Juola, & van de Par, 2008), the sight of a bouncing ball and the sound of its impact (van Eijk et al., 2008), or a hammer hitting a peg (Dixon & Spitz, 1980; Vatakis & Spence, 2006). In contrast, cross-modal temporal *imprecision* may be more useful during speech perception. That is, not only does knowledge of the precise relative timing of visual and acoustic speech

of a speaker provide no practical benefit to the listener, but also greater AVOA tolerance allows individuals to better fuse temporally distant AV cues and thus enhance speech comprehension. In other words, AVOA tolerance in speech perception might be more evolutionarily adaptive.

Though, Vroomen & Stekelenburg (2011) challenged this view, arguing that previous accounts on speech and non-speech may have been confounded by low-level stimulus differences. Thus, in their study they controlled for low-level stimulus factors while cueing the subjects to perceive the stimuli as either speech or non-speech. They compared audiovisual temporal synchrony judgements (point of subjective simultaneity or PSS, and Just Noticeable Difference or JND) between a group that perceived sine-wave speech (SWS) tokens as speech (speech-mode group) and a group that perceived the SWS tokens as artificial computer sounds (non-speech-mode group). They did not find differences *in-synchrony* judgments between the groups, concluding that speech is not a special case. However, we also minimized the low-level stimulus differences, while changing the fidelity of the original signal. We found that as the signal becomes more speech-like individuals become more tolerant of AVOA. Thus, this issue remains open to further research and debate.

4.6. Relevance to clinical populations

By using vocoded speech, we attempted to emulate the type of speech experienced by listeners with cochlear implants. We found a substantial drop in AVOA tolerance when stimuli were vocoded with 4 channels, with no significant difference between the less degraded vocoding conditions (> 8 channels). Still, factors such as CI users' experience with degraded speech and the degree of their reliance on visual speech compared to normal hearing listeners could not be taken into account in this study. Hay-McCutcheon, Pisoni, & Hunt (2009) examined AV asynchrony perception for words in individuals with cochlear implants. Their study did not find a significant difference in tolerance of AV asynchrony between normal hearing and CI listeners. As modern cochlear implants have 8 or more channels, these findings suggest that the fidelity of modern implants is sufficient to maintain a normal window of audiovisual integration. Ba Kent & Bazo (2011) also reported no difference of asynchrony detection between normal hearing individuals and individuals with moderate to severe hearing loss. It appears that only extreme reduction of phonetic information, beyond the degradation experienced by CI listeners, leads to a reduction of AVOA tolerance.

4.7. Caveats and Considerations

In the main experiment, inter-subject variance in overall AVOA tolerance was large (range from 8% to 100% *in-sync*, standard deviation 22%). The use of a single asynchrony value across subjects entailed that some subjects were much more likely to report *in-sync* than others. This is likely due to a combination of genuine differences in perceptual experience, as well as differences in participants' exact interpretation of the task and/or task strategy. This heterogeneity between subjects could have made small differences between experimental conditions more difficult to detect.

Along these lines, subjects may have been biased toward responding *in-sync* for more informative and intelligible stimuli despite the fact that participants were naïve to our research hypothesis and were explicitly instructed to judge synchrony regardless of intelligibility. Nonetheless, such bias is less likely for the following reasons: 1) If fidelity *per se* biased participants toward an *in-sync* percept, we would have observed greater AVOA tolerance for intact visual speech in Exp 1 (V-A) and Exp 2 (A-V). The effect was only present when visual speech led acoustic speech, with the opposite effect observed when acoustic speech led visual speech; 2) Substantial effects of AVOA tolerance among conditions were absent for perfect alignment (0 ms) and extreme asynchrony (550 ms) in the supplementary experiment (Fig. 5).

We also considered the use of alternative tasks such as the Temporal Order Judgment (TOJ) of the onset. However, previous evidence supports the use of a synchrony judgment task over TOJ task for investigating perception of audiovisual synchrony (van Eijk et al., 2008). More pertinently, we deemed the synchrony judgement task to be better suited for measuring the window of integration across the whole visual and acoustic speech stream, rather than focusing on the onsets alone or otherwise forcing attention to the most discriminative cues. As we alluded above, although asynchrony may be detected initially, auditory and visual percepts may still fuse as the speech unfolds. This may dynamically alter the subjective perception of synchrony, which was the focus of the current investigation. However, the current study was not designed to identify the exact visual and auditory cues used to make *in-sync/out-of-sync* judgements, and thus we cannot rule out the possibility that some participants made judgments primarily based on the onset (or offset) of mouth movement or the onset of auditory speech, whichever happens earlier. Although differences in onsets and offsets may be phonetically uninformative, such cues may serve a functional role to successfully bind the visual and auditory components of speech. Future studies are necessary to determine whether the influence of acoustic and visual fidelity on synchrony is evenly distributed over the course of an utterance.

5. CONCLUSION

The current findings support to the hypothesis that an increase in the meaningfulness of audiovisual speech enhances AVOA tolerance. According to the Dynamic Reweighting Model proposed by Bhat et al. (2015), enhanced fidelity of acoustic and visual speech renders the signal more speech-like and causes neural processing to shift toward high-level language networks, where wider windows of integration become available. However, this neural network reweighting account still needs to be verified, possibly via neuroimaging studies.

Acknowledgments

We thank Drs. Mark Pitt and Kristina Backer on their insights regarding the experimental design. We thank Dr. Lee Miller for providing the original videos. We thank Dr. Frederic Apoux for sharing the original vocoding Matlab scripts.

Funding: This work was supported by the National Institute of Health (NIH/NIDCD) under Grant R01-DC013543 (AJS).

References

- Arnal LH, Wyart V, Giraud AL. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*. 2011; 14(6):797–801. DOI: 10.1038/nn.2810 [PubMed: 21552273]
- Barutchu A, Crewther SG, Kiely P, Murphy MJ, Crewther DP. When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology*. 2008; 20(1):1–11. DOI: 10.1080/09541440601125623
- Ba kent D, Bazo D. Audiovisual Asynchrony Detection and Speech Intelligibility in Noise With Moderate to Severe Sensorineural Hearing Impairment. *Ear and Hearing*. 2011; 32(5):582–592. DOI: 10.1097/AUD.0b013e31820fca23 [PubMed: 21389856]
- Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience*. 2004; 7(11): 1190–1192. DOI: 10.1038/nn1333 [PubMed: 15475952]
- Beauchamp MS, Lee KE, Argall BD, Martin A. Integration of Auditory and Visual Information about Objects in Superior Temporal Sulcus. *Neuron*. 2004; 41(5):809–823. DOI: 10.1016/S0896-6273(04)00070-4 [PubMed: 15003179]
- Beauchamp MS, Nath AR, Pasalar S. fMRI-Guided Transcranial Magnetic Stimulation Reveals That the Superior Temporal Sulcus Is a Cortical Locus of the McGurk Effect. *Journal of Neuroscience*. 2010; 30(7):2414–2417. DOI: 10.1523/JNEUROSCI.4865-09.2010 [PubMed: 20164324]
- Beauchamp MS, Nath AR, Pasalar S. fMRI-Guided Transcranial Magnetic Stimulation Reveals That the Superior Temporal Sulcus Is a Cortical Locus of the McGurk Effect. *Journal of Neuroscience*. 2010; 30(7):2414–2417. DOI: 10.1523/JNEUROSCI.4865-09.2010 [PubMed: 20164324]
- Benard MR, Ba kent D. The effect of visual cues on top-down restoration of temporally interrupted speech, with and without further degradations. *Hearing Research*. 2015; 328:24–33. DOI: 10.1016/j.heares.2015.06.013 [PubMed: 26117407]
- Bernstein LE, Auer ET, Takayanagi S. Auditory speech detection in noise enhanced by lipreading. *Speech Communication*. 2004; 44(1–4):5–18. DOI: 10.1016/j.specom.2004.10.011
- Bhat J, Miller LM, Pitt MA, Shahin AJ. Putative mechanisms mediating tolerance for audiovisual stimulus onset asynchrony. *Journal of Neurophysiology*. 2015; 113(5):1437–1450. DOI: 10.1152/jn.00200.2014 [PubMed: 25505102]
- Bhat J, Pitt MA, Shahin AJ. Visual context due to speech-reading suppresses the auditory response to acoustic interruptions in speech. *Frontiers in Neuroscience*. 2014; 8doi: 10.3389/fnins.2014.00173
- Boenke LT, Deliano M, Ohl FW. Stimulus duration influences perceived simultaneity in audiovisual temporal-order judgment. *Experimental Brain Research*. 2009; 198(2–3):233–244. DOI: 10.1007/s00221-009-1917-z [PubMed: 19590862]
- Brungart, DS., Iyer, I., Simpson, BD., van Wassenhove, V. The effects of temporal asynchrony on the intelligibility of accelerated speech. Presented at the International Conference on Auditory-Visual Speech Processing 2008; Tangalooma Wild Dolphin Resort, Moreton Island, Queensland, Australia. 2008 Sep. p. 19-24. Retrieved from http://www.isca-speech.org/archive_open/avsp08/av08_019.html
- Cappelletta, L., Harte, N. Phoneme-to-viseme Mapping for Visual Speech Recognition. Presented at the 1st International Conference on Pattern Recognition Applications and Methods; Vilamoura, Algarve, Portugal. 2012 Feb. p. 174-176. Retrieved from <http://citeseerx.ist.psu.edu/>
- Cecere R, Gross J, Thut G. Behavioural evidence for separate mechanisms of audiovisual temporal binding as a function of leading sensory modality. *European Journal of Neuroscience*. 2016; 43(12):1561–1568. DOI: 10.1111/ejn.13242 [PubMed: 27003546]
- Coltheart M. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*. 1981; 33(4):497–505. DOI: 10.1080/14640748108400805
- 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.doi: 10.1121/1.2195091 [PubMed: 16838548]

- Desai S, Stickney G, Zeng FG. Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *The Journal of the Acoustical Society of America*. 2008; 123(1):428.doi: 10.1121/1.2816573 [PubMed: 18177171]
- Dias JW, Cook TC, Rosenblum LD. Influences of selective adaptation on perception of audiovisual speech. *Journal of Phonetics*. 2016; 56:75–84. DOI: 10.1016/j.wocn.2016.02.004 [PubMed: 27041781]
- Dixon NF, Spitz L. The detection of auditory visual desynchrony. *Perception*. 1980; 9(6):719–721. DOI: 10.1068/p090719 [PubMed: 7220244]
- Eg R, Behne DM. Perceived synchrony for realistic and dynamic audiovisual events. *Frontiers in Psychology*. 2015; 6doi: 10.3389/fpsyg.2015.00736
- Eg R, Griwodz C, Halvorsen P, Behne D. Audiovisual robustness: exploring perceptual tolerance to asynchrony and quality distortion. *Multimedia Tools and Applications*. 2015; 74(2):345–365. DOI: 10.1007/s11042-014-2136-6
- Freeman ED, Ipser A, Palmbaha A, Paunoiu D, Brown P, Lambert C, Driver J. Sight and sound out-of-synch: Fragmentation and renormalisation of audiovisual integration and subjective timing. *Cortex*. 2013; 49(10):2875–2887. DOI: 10.1016/j.cortex.2013.03.006 [PubMed: 23664001]
- Grant, KW., Greenberg, S. Speech intelligibility derived from asynchronous processing of auditory-visual information. Presented at the AVSP 2001 International Conference on Auditory-Visual Speech Processing; Aalborg, Denmark. 2001 Sep. p. 132-137. Retrieved from http://www.isca-speech.org/archive_open/avsp01/av01_132.html
- 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.doi: 10.1121/1.1288668 [PubMed: 11008820]
- Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*. 1998; 103(5):2677.doi: 10.1121/1.422788 [PubMed: 9604361]
- Grant KW, van Wassenhove V, Poeppel D. Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Communication*. 2004; 44(1–4):43–53. DOI: 10.1016/j.specom.2004.06.004
- Hay-McCutcheon MJ, Pisoni DB, Hunt KK. Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: A preliminary analysis. *International Journal of Audiology*. 2009; 48(6):321–333. DOI: 10.1080/14992020802644871 [PubMed: 19925340]
- Heinrich A, Carlyon RP, Davis MH, Johnsrude IS. Illusory vowels resulting from perceptual continuity: a functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*. 2008; 20(10):1737–1752. DOI: 10.1162/jocn.2008.20069 [PubMed: 18211243]
- Heinrich A, Carlyon RP, Davis MH, Johnsrude IS. The Continuity Illusion Does Not Depend on Attentional State: fMRI Evidence from Illusory Vowels. *Journal of Cognitive Neuroscience*. 2011; 23(10):2675–2689. DOI: 10.1162/jocn.2011.21627 [PubMed: 21268669]
- Hickok G. The functional neuroanatomy of language. *Physics of Life Reviews*. 2009; 6(3):121–143. DOI: 10.1016/j.plrev.2009.06.001 [PubMed: 20161054]
- Hickok G, Poeppel D. Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*. 2000; 4(4):131–138. DOI: 10.1016/S1364-6613(00)01463-7 [PubMed: 10740277]
- Jesse A, Massaro DW. The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception & Psychophysics*. 2010; 72(1):209–225. DOI: 10.3758/APP.72.1.209
- Kaiser AR, Kirk KI, Lachs L, Pisoni DB. Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research: JSLHR*. 2003; 46(2):390–404. DOI: 10.1044/1092-4388(2003/032)
- Keceli S, Okamoto H, Kakigi R. Hierarchical neural encoding of temporal regularity in the human auditory cortex. *Brain Topography*. 2015; 28(3):459–470. DOI: 10.1007/s10548-013-0300-3 [PubMed: 23793552]
- MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*. 1987; 21(2):131–141. DOI: 10.3109/03005368709077786 [PubMed: 3594015]

- Magnotti JF, Ma WJ, Beauchamp MS. Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology*. 2013; 4doi: 10.3389/fpsyg.2013.00798
- Mégevand P, Molholm S, Nayak A, Foxe JJ. Recalibration of the Multisensory Temporal Window of Integration Results from Changing Task Demands. *PLoS ONE*. 2013; 8(8):e71608.doi: 10.1371/journal.pone.0071608 [PubMed: 23951203]
- Middelweerd MJ, Plomp R. The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*. 1987; 82(6):2145–2147. DOI: 10.1121/1.395659 [PubMed: 3429736]
- Miller LM, D'Esposito M. Perceptual Fusion and Stimulus Coincidence in the Cross-Modal Integration of Speech. *Journal of Neuroscience*. 2005; 25(25):5884–5893. DOI: 10.1523/JNEUROSCI.0896-05.2005 [PubMed: 15976077]
- Moore, BCJ., editor. *Cochlear Hearing Loss*. West Sussex, England: John Wiley & Sons, Ltd.; 2007. Retrieved from <http://doi.wiley.com/10.1002/9780470987889>
- 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. DOI: 10.1121/1.423300 [PubMed: 9670544]
- 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. DOI: 10.1016/j.cogbrainres.2005.07.009 [PubMed: 16137867]
- Parise CV, Ernst MO. Correlation detection as a general mechanism for multisensory integration. *Nature Communications*. 2016; 7:11543.doi: 10.1038/ncomms11543
- Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Chang EF. Reconstructing Speech from Human Auditory Cortex. *PLoS Biology*. 2012; 10(1):e1001251.doi: 10.1371/journal.pbio.1001251 [PubMed: 22303281]
- Patterson RD, Uppenkamp S, Johnsrude IS, Griffiths TD. The Processing of Temporal Pitch and Melody Information in Auditory Cortex. *Neuron*. 2002; 36(4):767–776. DOI: 10.1016/S0896-6273(02)01060-7 [PubMed: 12441063]
- Poeppl 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. DOI: 10.1098/rstb.2007.2160
- Powers AR, Hillock AR, Wallace MT. Perceptual Training Narrows the Temporal Window of Multisensory Binding. *Journal of Neuroscience*. 2009; 29(39):12265–12274. DOI: 10.1523/JNEUROSCI.3501-09.2009 [PubMed: 19793985]
- Pumplin J. Low-noise noise. *The Journal of the Acoustical Society of America*. 1985; 78(1):100.doi: 10.1121/1.392571
- Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*. 2009; 12(6):718–724. DOI: 10.1038/nn.2331 [PubMed: 19471271]
- Rauschecker J, Tian B, Hauser M. Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*. 1995; 268(5207):111–114. DOI: 10.1126/science.7701330 [PubMed: 7701330]
- Samuel AG, Lieblich J. Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology: Human Perception and Performance*. 2014; 40(4):1479–1490. DOI: 10.1037/a0036656 [PubMed: 24749935]
- 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 Computational Biology*. 2014; 10(7):e1003743.doi: 10.1371/journal.pcbi.1003743 [PubMed: 25079216]
- Shahin AJ, Bishop CW, Miller LM. Neural mechanisms for illusory filling-in of degraded speech. *NeuroImage*. 2009; 44(3):1133–1143. DOI: 10.1016/j.neuroimage.2008.09.045 [PubMed: 18977448]
- Shahin AJ, Kerlin JR, Bhat J, Miller LM. Neural restoration of degraded audiovisual speech. *NeuroImage*. 2012; 60(1):530–538. DOI: 10.1016/j.neuroimage.2011.11.097 [PubMed: 22178454]
- Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech Recognition with Primarily Temporal Cues. *Science*. 1995; 270(5234):303–304. DOI: 10.1126/science.270.5234.303 [PubMed: 7569981]

- Stevenson RA, Wilson MM, Powers AR, Wallace MT. The effects of visual training on multisensory temporal processing. *Experimental Brain Research*. 2013; 225(4):479–489. DOI: 10.1007/s00221-012-3387-y [PubMed: 23307155]
- Sumby WH. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*. 1954; 26(2):212.doi: 10.1121/1.1907309
- Sumby WH, Pollack I. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*. 1954; 26(2):212.doi: 10.1121/1.1907309
- Summerfield Q. Lipreading and Audio-Visual Speech Perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 1992; 335(1273):71–78. DOI: 10.1098/rstb.1992.0009
- Tian B, Reser D, Durham A, Kustov A, Rauschecker JP. Functional specialization in rhesus monkey auditory cortex. *Science*. 2001; 292(5515):290–293. DOI: 10.1126/science.1058911 [PubMed: 11303104]
- Tye-Murray N, Sommers M, Spehar B. Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception. *Trends in Amplification*. 2007; 11(4):233–241. DOI: 10.1177/1084713807307409 [PubMed: 18003867]
- Van der Burg E, Alais D, Cass J. Rapid Recalibration to Audiovisual Asynchrony. *Journal of Neuroscience*. 2013; 33(37):14633–14637. DOI: 10.1523/JNEUROSCI.1182-13.2013 [PubMed: 24027264]
- van Eijk RLJ, Kohlrausch A, Juola JF, van de Par S. Audiovisual synchrony and temporal order judgments: Effects of experimental method and stimulus type. *Perception & Psychophysics*. 2008; 70(6):955–968. DOI: 10.3758/PP.70.6.955 [PubMed: 18717383]
- Van Engen KJ, Phelps JEB, Smiljanic R, Chandrasekaran B. Enhancing Speech Intelligibility: Interactions Among Context, Modality, Speech Style, and Masker. *Journal of Speech Language and Hearing Research*. 2014; 57(5):1908.doi: 10.1044/JSLHR-H-13-0076
- van Wassenhove V. Speech through ears and eyes: interfacing the senses with the supramodal brain. *Frontiers in Psychology*. 2013; 4doi: 10.3389/fpsyg.2013.00388
- van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*. 2005; 102(4):1181–1186. DOI: 10.1073/pnas.0408949102
- van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*. 2007; 45(3):598–607. DOI: 10.1016/j.neuropsychologia.2006.01.001 [PubMed: 16530232]
- Vatakis A, Ghazanfar AA, Spence C. Facilitation of multisensory integration by the “unity effect” reveals that speech is special. *Journal of Vision*. 2008; 8(9):14.1–11. DOI: 10.1167/8.9.14
- 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. *Frontiers in Integrative Neuroscience*. 2012; 6:71.doi: 10.3389/fnint.2012.00071 [PubMed: 23060756]
- Vatakis A, Navarra J, Soto-Faraco S, Spence C. Audiovisual temporal adaptation of speech: temporal order versus simultaneity judgments. *Experimental Brain Research*. 2008; 185(3):521–529. DOI: 10.1007/s00221-007-1168-9 [PubMed: 17962929]
- Vatakis A, Spence C. Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*. 2006; 1111(1):134–142. DOI: 10.1016/j.brainres.2006.05.078 [PubMed: 16876772]
- Vatakis A, Spence C. Crossmodal binding: evaluating the “unity assumption” using audiovisual speech stimuli. *Perception & Psychophysics*. 2007; 69(5):744–756. DOI: 10.3758/BF03193776 [PubMed: 17929697]
- Vatakis A, Spence C. Evaluating the influence of the “unity assumption” on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica*. 2008; 127(1):12–23. DOI: 10.1016/j.actpsy.2006.12.002 [PubMed: 17258164]
- Vroomen J, Stekelenburg JJ. Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*. 2011; 118(1):75–83. DOI: 10.1016/j.cognition.2010.10.002 [PubMed: 21035795]
- Welch RB, Warren DH. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*. 1980; 88(3):638–667. DOI: 10.1037/0033-2909.88.3.638 [PubMed: 7003641]

Zampini M, Guest S, Shore DI, Spence C. Audio-visual simultaneity judgments. *Perception & Psychophysics*. 2005; 67(3):531–544. DOI: 10.3758/BF03193329 [PubMed: 16119399]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
Schematic of stimulus presentation and subject's response for a single trial.

Visual speech leads acoustic speech

Figure 2.

Bar graphs (with standard error) depicting AVOA tolerance as a function of lexicity, visual and acoustic fidelity. **A.** Main effect of lexicity. Individuals tended to tolerate AVOA more for words than nonwords. **B.** Main effect of visual fidelity. Individuals tolerated AVOA more when videos were intact than blurred. **C.** Main effect of acoustic fidelity. Individuals tolerated AVOA more for words vocoded at 8-, 16- and 32 channels than 4 channels. ns indicates non-significance ($p > 0.05$).

Acoustic speech leads visual speech

Figure 3.

Bar graphs (with standard error) depicting AVOA tolerance as a function of lexicality, visual and acoustic fidelity. **A.** Main effect of lexicality. Individuals tolerated AVOA more for words than nonwords. **B.** Main effect of visual fidelity. Individuals tended to tolerate AVOA more for blurred than intact videos. **C.** Main effect of acoustic fidelity. Individuals tolerated AVOA more for words vocoded at 8-, 16- and 32 channels than 4 channels. ns indicates non-significance ($p > 0.05$).

Figure 4. Bar graphs (with standard error) depicting AVOA tolerance as a function of the first and second halves of the experiments when visual speech leads (A) and lags (B) acoustic speech.

Figure 5.

Bar graphs (with standard error) depicting AVOA tolerance as a function of visual and acoustic fidelity for three asynchrony values (0, 360 and 550 ms), when visual speech leads (A) and lags (B) acoustic speech.

Figure 6. Bar graphs (with standard error) depicting intelligibility as a function of visual (**A**) and acoustic (**B**) fidelity.