BRILL

Multisensory Research 31 (2018) 39–56

MULTISENSORY RESEARCH

brill.com/msr

Electrophysiological Indices of Audiovisual Speech Perception: Beyond the McGurk Effect and Speech in Noise

Julia Irwin [1,3], **Trey Avery** [1], **Lawrence Brancazio** [1,3], **Jacqueline Turcios** [1,3], **Kayleigh Ryherd** [1,2] **and Nicole Landi** [1,2,*]

[1] Haskins Laboratories, New Haven, CT, USA

[2] University of Connecticut, Storrs, CT, USA

[3] Southern Connecticut State University, New Haven, CT, USA

Received 27 September 2016; accepted 15 May 2017

Abstract

Visual information on a talker's face can influence what a listener hears. Commonly used approaches to study this include mismatched audiovisual stimuli (e.g., McGurk type stimuli) or visual speech in auditory noise. In this paper we discuss potential limitations of these approaches and introduce a novel visual phonemic restoration method. This method always presents the same visual stimulus (e.g., /ba/) dubbed with a matched auditory stimulus (/ba/) or one that has weakened consonantal information and sounds more /a/-like). When this reduced auditory stimulus (or */a/*) is dubbed with the visual /ba/, a visual influence will result in effectively 'restoring' the weakened auditory cues so that the stimulus is perceived as a /ba/. An oddball design in which participants are asked to detect the */a/* among a stream of more frequently occurring /ba/s while either a speaking face or face with no visual speech was used. In addition, the same paradigm was presented for a second contrast in which participants detected /pa/ among /ba/s, a contrast which should be unaltered by the presence of visual speech. Behavioral and some ERP findings reflect the expected phonemic restoration for the /ba/ vs. */a/* contrast; specifically, we observed reduced accuracy and P300 response in the presence of visual speech. Further, we report an unexpected finding of reduced accuracy and P300 response for both speech contrasts in the presence of visual speech, suggesting overall modulation of the auditory signal in the presence of visual speech. Consistent with this, we observed a mismatch negativity (MMN) effect for the /ba/ vs. /pa/ contrast only that was larger in absence of visual speech. We discuss the potential utility for this paradigm for listeners who cannot respond actively, such as infants and individuals with developmental disabilities.

* To whom correspondence should be addressed. E-mail: Nicole.Landi@yale.edu

DOI:10.1163/22134808-00002580

Keywords
Audiovisual speech perception, ERP, phonemic restoration

1. Introduction

When a talker produces speech, motor movements (or articulation) are visible on the speaker's face. This visible speech can provide information for the listener about what was said. Typical speech and language development is thought to take place in this audiovisual (AV) context, fostering native language acquisition (Bergeson and Pisoni, 2004; Desjardins *et al*., 1997; Lachs *et al*., 2001; Legerstee, 1990; Lewkowicz and Hansen-Tift, 2012; Meltzoff and Kuhl, 1994). Perception and production of speech can be influenced by both auditory and visual signals. For example, sighted speakers produce vowels that are further apart in articulatory space than those of blind speakers, indicating that the ability to see speech influences how it is ultimately produced (Menard *et al*., 2009). Moreover, children who produce sound substitutions are less visually influenced when viewing articulation of sounds that they cannot produce (Desjardins *et al*., 1997).

Visual information has been shown to assist listeners in the identification of speech in auditory noise, creating a 'visual gain' over heard speech alone (Sumby and Pollack, 1954; also see Erber, 1975; Grant and Seitz, 2000; Macleod and Summerfield, 1987; Payton *et al*., 1994; Ross *et al*., 2007). Significantly, visible articulatory information can impact heard speech even when the auditory signal is in clear listening conditions, that is, where there is no attendant background noise. A striking demonstration of the influence of visual information on heard speech is the classic experiment of MacDonald and McGurk (1978), where a speaker was videotaped producing consonant vowel consonant vowel (CVCV, such as /baba/) syllables with a different auditory syllable dubbed over the video. Listeners watching these dubbed productions sometimes reported hearing consonants that combined the places of articulation of the visual and auditory tokens (e.g., a visual /ba/ + an auditory /ga/ would be heard as /bga/), 'fused' the two places (e.g., a visual /ga/ + auditory /ba/ would be heard as /da/), or reflected the visual place information alone (visual /va/ + auditory /ba/ would be heard as /va/) (McGurk and MacDonald, 1976). This effect is known as the McGurk Effect (e.g., Brancazio *et al*., 2006), McGurk–MacDonald Effect (e.g., Colin *et al*., 2002) or McGurk Illusion (e.g., Alsius et al., 2005; Brancazio and Miller, 2005; Green, 1994; Rosenblum, 2008; Soto-Faraco and Alsius, 2009; Walker *et al*., 1995; Windmann, 2004).

Electrophysiological measures such as electroencephalography (EEG) and event-related potentials (ERP) have recently been used to study AV speech

perception. These techniques provide excellent temporal resolution, allowing for sensitive assessment of timing in response to AV stimuli (e.g., Klucharev, Möttönen, and Sams, 2003; Molholm *et al*., 2002; Pilling, 2009; Saint-Amour *et al*., 2007). Specifically, a number of studies have looked at components sensitive to early auditory and visual features in the auditory N1 and P2 during processing of AV speech. The auditory N1/P2 complex is elicited by auditory stimuli and can be modulated be the stimulus properties of sounds, including auditory speech (e.g., Pilling, 2009; Tremblay *et al*., 2001; Van Wassenhove *et al*., 2005). Van Wassenhove *et al*. (2005) and Pilling (2009) both report that congruent visual speech information presented with the auditory signal attenuates the amplitude of the N1/P2 auditory event-related potential (ERP) response, resulting in lower peak amplitude and a shortening of their latency. In general, EEG/ERP studies reveal that the combination of auditory and visual speech appears to dampen amplitude and speed processing of the speech signal. Further, using current density reconstructions of ERP data, Bernstein *et al*. (2008) proposed a potential spatio-temporal audiovisual speech processing circuit based for AV speech processing. Bernstein *et al*. (2008) reported very early (less than 100 ms) simultaneous activation of the supramarginal gyrus (SMG), the angular gyrus (AG), the intraparietal sulcus, the inferior frontal gyrus and the dorsolateral prefrontal cortex in adults watching speaking faces. At later time points (160 to 220 ms) Bernstein and colleagues observed only a more focal SMG/AG activation in the left hemisphere (Bernstein *et al*., 2008). These source localization findings are broadly consistent with fMRI findings (which have more precise spatial but poor temporal resolution) that localize AV speech processing to the STG, SMG and IFG.

Given that deficits in AV processing have been associated with complex neurodevelopmental disorders such as autism spectrum disorders (ASD) and specific language impairment (SLI; e.g., Bebko *et al*., 2006; Foss-Feig *et al*., 2010; Iarocci *et al*., 2010; Irwin *et al*., 2011; Kaganovich *et al*., 2014, 2016; Smith and Bennetto, 2007), better understanding of the neural bases for typical and atypical AV speech perception will be useful for identifying potential biomarkers that could indicate communication disorders.

While there has been a sustained focus on speech in noise and mismatched (or McGurk type) AV tasks in both behavioral and electrophysiological studies, each has some limitations. Critically, the McGurk Effect creates a percept where what is heard is a separate syllable (or syllables) from either the visual or auditory signal, which generates conflict between the two modalities (Brancazio, 2004). Visual influence for these 'illusory' percepts can vary greatly by individual (Nath and Beauchamp, 2012; Schwartz, 2010). Further, McGurk type percepts are rated as less good exemplars of the category (e.g., a poorer example of a 'ba') than tokens where the visual and auditory stimuli specify the same syllable (Brancazio, 2004). Poorer exemplars of a category could lead

to decision-level difficulties in executive functioning (potentially problematic for all perceivers, and an established area of weakness for those with ASD — Eigsti and Shapiro, 2003). Finally, auditory noise may be especially disruptive for individuals with developmental disabilities (Alcántara *et al*., 2004; Irwin *et al*., 2011). While these methods may already be suboptimal for typically developing perceivers, if one wants to examine visual influence on heard speech in clinical populations, both noisy speech and McGurk type stimuli may provide additional challenges in interpreting findings. In particular, individuals with developmental delays, particularly in populations who may have significant challenges in reliably responding verbally, with aversive stimuli (such as noise) and with tasks that require executive functioning. It is possible, then, that group differences reported in the literature on AV speech processing, such as those seen in children with ASD are more a function of challenges with the standard tasks and not with speech processing per se.

To overcome these potential limitations, we have begun to assess the influence of visible articulatory information on heard speech with a novel measure that involves neither noise nor auditory and visual category conflict that can serve as an alternative to assessing audiovisual speech processing in clinical populations (also see Jerger *et al*., 2014). This new paradigm uses restoration of weakened auditory tokens with visual stimuli. There are two types of stimuli presented to the listener: clear exemplars of an auditory token (intact /ba/), and reduced tokens in which the auditory cues for the consonant are substantially weakened so that the consonant is not detected (reduced /ba/, which is more /a/ like, from this point on referred to as /*a*/). The auditory stimuli are created by synthesizing speech based on a natural production of a consonant vowel syllable (e.g., /ba/) and systematically flattening the formant transitions to create the /*a*/. Video of the speaker's face does not change (always producing /ba/), but the auditory stimuli (/ba/ or /*a*/) vary. Thus, in this example, when the /*a*/ stimulus is dubbed with the visual /ba/, a visual influence will result in effectively 'restoring' the weakened auditory cues so that the stimulus is perceived as a /ba/, akin to a visual phonemic restoration effect (Kashino, 2006; Samuel, 1981; Warren, 1970). Notably, in this design, the visual information for the same phoneme /ba/ supplements insufficient auditory information to assess the influence of visual information on what is heard (Brancazio *et al*., 2015).

The current paper examines both behavioral data and neural signatures (ERP) of audiovisual processing with this novel visual phonemic restoration method. The inclusion of ERP provides a more direct measure of neural discrimination in addition to the more traditional behavioral approach (here, an identification button press) to assess whether the neurobiological information is consistent with the behavioral data. We do this by examining two ERP components that are associated with discrimination and elicited in response

to detection of an infrequently occurring stimulus, the P300 and the mismatch negativity (MMN). An oddball paradigm was used to elicit P300 and MMN responses to the infrequently occurring /*a*/ (deviant) embedded within the more frequently occurring intact /ba/ (standards). A control contrast condition was also included; this condition involved an infrequently presented /pa/ (deviant) paired with a more frequently occurring /ba/ (standard). For each speech contrast condition, all speech tokens are paired with a face producing /ba/ or a face with a pixelated mouth containing motion but no visual speech. Behaviorally, we expect lower accuracy for the /*a*/ in the presence of a speaking face relative to the pixelated face; however, we expect that participants will easily discriminate the /pa/ stimulus in both an auditory only (pixelated) condition and an audiovisual condition (as auditory /pa/ paired with a speaker producing /ba/ leads to the perception of /pa/). For the ERP findings we predicted that P300 and MMN effects would be reduced for the /*a*/ vs. /ba/ contrast in the presence of a speaking face relative to the pixelated face, consistent with the behavioral predictions. Likewise, we predict little or no modulation of the P300 or MMN effects for the /pa/ vs. /ba/ contrast as a function of face context.

2. Method

All data was collected at matching EEG facilities at Haskins Laboratories in New Haven, CT, USA, and at the University of Connecticut in Storrs, CT, USA. Identical equipment and experimental procedures were used at both data collection sites.

2.1. Participants

Participants were recruited through the University of Connecticut psychology participant pool and through printed posters and online postings. A total of 50 adults participated for research credit or as unpaid volunteers. Five adults were removed from the analysis because they did not have sufficient numbers of usable trials (see data preprocessing). Of the 45 participants, all were right handed and reported normal hearing. The sample was comprised of 30 females and 15 males, mean age = 20.07 years (SD = 3.49 years). Reported race and ethnicities included 34 Caucasian participants, five Asian/Pacific Islander participants, three African American participants, one Hispanic participant, and two mixed-race participants.

2.2. Audiovisual Stimuli and Experimental Paradigm

The stimuli were created as follows. First, we videotaped and recorded an adult male speaker of English producing the syllable /ba/, and using Praat (Boersma and Weenink, 2013), we extracted acoustic parameters for the token (including formant trajectories, amplitude contour, voicing and pitch contour). Critically,

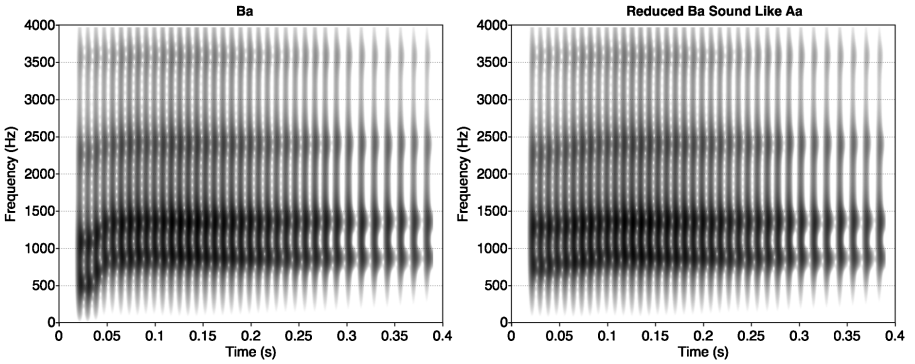

Figure 1. First panel, top left, spectrogram of synthesized/ba/. Second panel, top right, edited synthesized /ba/ with reduced initial formants for the consonant, referred to as /*a*/.

the token had rising formant transitions for F1, F2, and to a lesser extent F3, characteristic of /ba/. To create our /ba/ stimulus, we synthesized a new token of /ba/ based on these values (auditory /ba/ stimulus available as Sound S1 in online supplementary materials). To create our /*a*/ stimulus, we then modified the synthesis parameters (auditory /a/ stimulus available as Sound S2 in online supplementary materials). Specifically, we changed the onset values for F1 and F2 to reduce the extent of the transitions and lengthened the transition durations for F1, F2, and F3, and then synthesized a new stimulus. Specifically, for the full /ba/, the transitions were 34 ms long and F1 rose from 500 to 850 Hz; F2 rose from 1150 to 1350 Hz; and F3 rose from 2300 to 2400 Hz. For the /*a*/, the transitions were 70 ms long and F1 rose from 750 to 850 Hz; F2 rose from 1300 to 1350 Hz; and F3 rose from 2300 to 2400 Hz (see spectrograms in Fig. 1). Finally, to create /pa/, we modified the original /ba/ parameters for amplitude and voicing for the early portion of the stimulus to create a small burst and an aspirated, unvoiced segment, and again synthesized a new stimulus. The voice onset time (VOT) for the synthesized /pa/ was 70 ms.

To produce AV stimuli, the three synthesized stimuli were dubbed onto a visual token of the model speaker producing /ba/, with the acoustic onsets synchronized with the visible opening of the mouth. Finally, to produce PX stimuli, we created a version of the visual token in which the mouth portion was pixelated so that the articulatory movement could not be perceived (although variation in the pixelation indicated movement). Again, the three synthesized stimuli were dubbed onto the pixelated stimulus. See online Supplementary Video 1 (PX) for an example of the pixelated face stimulus.

Instructions and a practice trial were presented prior to the start of the experiment. Within the full EEG session, the experiment was blocked into two face context conditions (speaking face and pixelated face and two speech contrast conditions (ba/pa and ba/a, see Fig. 2), creating four total blocks. Each

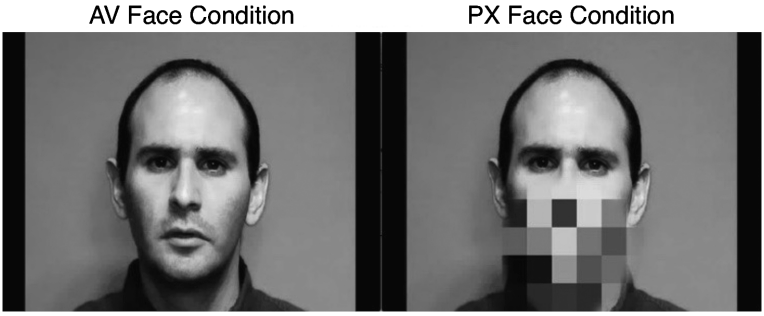

Figure 2. Left panel, audiovisual face condition, showing the visible articulation of the speaker. Right panel, pixelated face condition, showing the speaker's face, but obscuring the mouth.

face context by speech contrast block was 9 min, and contained 100 trials. For the first two blocks, the speaking face was fully visible and in the second two blocks the area around the mouth was pixelated to obscure features of the mouth but head movement is still visible during production of the speech sounds. This presentation order was intentional in order to ensure that the phonemic restoration effect was tested without exposure to the contrast of the full /ba/ and /a/ tokens in the clear. An 85/15 oddball design was used for presentation of the speech contrast stimuli, with */a/* serving as the deviant in the b/a contrast condition and /pa/ serving as the deviant in the ba/pa contrast condition. Participants were played the deviant sound (*/a/* or /pa/) before each block to remind them what they were listening for, and instructed to press the response button only after the presentation of that deviant and to otherwise remain as still as possible. Total experiment time was approximately 45 min depending on length of breaks and amount of EEG net rehydration between blocks.

2.3. EEG Data Collection

EEG data was collected with an Electrical Geodesics Inc. (EGI) netamps 300 high-impedance amplifier, using 128 Ag/AgCl electrodes embedded in soft sponges woven into a geodesic array. The EEG sensor nets were soaked for up to 10 min prior to use in a warm potassium chloride solution (2 teaspoons of potassium chloride, 1 liter of water purified by reverse osmosis, and 3 cc of Johnson and Johnson baby shampoo to remove oils from the scalp). The high-density hydrocel geodesic sensor nets and associated high impedance amplifiers have been designed to accept impedance values ranging as high as 100 kΩ, which permits the sensor nets to be applied in under 10 min and without scalp abrasion, recording paste, or gel (e.g., Ferree *et al.*, 2001; Pizzagalli, 2007). Impedance for all electrodes was kept below 40 kΩ throughout the experimental run (impedances were re-checked between blocks). Online

recordings at each electrode used the vertex electrode as the reference and were later referenced to the average reference.

EEG was continuously recorded using Netstation 4.5.7 on a MacPro running OS X 10.6.8 while participants completed experimental tasks. Stimuli were presented using E-Prime (PST) version 2.0.8.90 on a Dell Optiplex 755 (Intel Core 2 Duo at 2.53 GHz and 3.37 GB RAM) running Windows XP. Audio stimuli were presented from an audio speaker centered 85 cm above the participant connected to a Creative SB X-Fi audio card. Visual stimuli were presented at a visual angle of 23.62 degrees [video was 9.44 inches (24 cm) wide and 7.87 inches (20) cm tall] on Dell 17 inch flat panel monitors 60 cm from the participant connected to an Nvidia GeForce GT 630 video card. Speech sounds were presented free-field at 65 decibels, measured by a sound pressure meter.

2.4. ERP Data Processing

Initial processing was conducted using Netstation 4.5.7. EEG data were band pass filtered at 0.3 to 30 Hz [Passband Gain: 99.0 % (−0.1 dB), Stopband Gain: 1.0 % (−40.0 dB), Rolloff: 2.00 Hz] and segmented by condition, 100 ms pre-stimulus to 800 ms post-stimulus. Eye blinks and vertical eye movements were examined with electrodes located below and above the eyes (channels 8, 126, 25, 127). Horizontal eye movements were measured using channels 125 and 128, located at positions to the left and right of the eyes. Artifacts were automatically detected and manually verified for exclusion from additional analysis (bad channel $>$ 200 μV, eye blinks $>$ 140 μV and eye movement $>$ 55 μV). For every channel, 50% or greater bad segments was used as the criteria for marking the channel bad; for every segment, greater than 20 bad channels was used as a criterion for marking a segment bad. Participants with fewer than 20 (25%) out of a possible 80 usable trials in any condition were excluded from analysis, leaving 45 (out of a total of 50) participants in the sample. The average usable trial count across all conditions had a mean of 58.39 (SD = 15.92) and each experiment had similar amounts of usable data AV mean = 59.36 (SD = 14.83) and PX mean = 57.42 (SD = 16.64). Collapsing standards and deviants there were similar quantities of usable trials in the grand average mean standards = 55.78 (SD = 15.86) and mean deviants = 61.00 (SD = 15.98).

Bad channels (fluctuations over 100 μV) were spline interpolated from nearby electrodes. Data were baseline corrected using a 100-ms window prior to onset of all stimuli. Data were re-referenced from vertex recording to an average reference of all 128 channels. For ERP analysis, only standard /ba/ sounds before each of the deviant (/*a*/ or /pa/) and deviants with accurate behavioral responses were included.

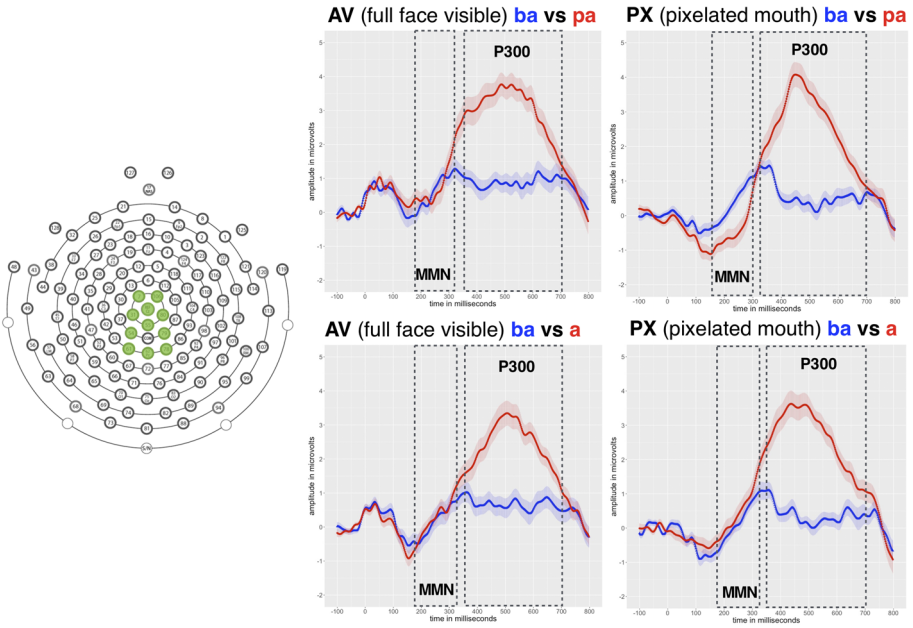

Figure 3. Left panel: EEG electrode montage used for ERP waveform plots. Top center (AV) and right (PX): MMN and P300 response to standard /ba/ and deviant /pa/; bottom center (AV) and right (PX): MMN and P300 response to standard /ba/ and deviant /*a*/. Shading around the waveform represents standard error from the grand mean.

All processed, artifact-free segments were averaged by condition producing a single event related potential waveform for each condition for all participants and exported for plotting and statistical analysis in R. The MMN was identified as the most negative peak between 200 and 375 ms following stimulus onset and the P300 was identified as the most positive peak between 400 and 700 ms following stimulus onset, both within a cluster of eleven central electrodes [Hydrocel GSN channels 54, 55 (CPz), 61, 62 (Pz), 67 (P03), 71, 72 (POz), 76, 77 (PO4), 78, 79; see Fig. 3]. Waveforms for each channel within the averaged cluster of electrodes are available as online Supplementary Figs S1 for the AV condition and S2 for the PX condition. Identification of the P300 and MMN was based on both visual inspection and guidelines provided by previous MMN and P300 studies, which indicate a fronto-central distribution (Alho, 1995; Polich *et al.*, 2007). Statistical analyses (repeated measures ANOVAs and *t*-tests for planned comparisons, see Sect. 3. Results) were conducted on average amplitudes that included a window of 25 ms around the peak, identified using an adaptive mean function, which identifies individual windows for each participant to account for subtle differences in waveform morphology across participants. In all waveform plots shading around waveforms represent the standard error from the mean.

3. Results

3.1. Behavioral Data

To analyze accuracy data, we ran a 2 × 2 repeated measures ANOVA with speech contrast (ba/a vs. ba/pa) and face context (AV vs. PX; audiovisual speech vs. audio only) as within subjects variables. On average, participants were able to perceive and respond to the deviant oddball target stimuli with high accuracy, mean accuracy = 92.42% (SD = 6.29). Our ANOVA revealed a main effect of speech contrast $F(1, 44) = 6.945$, $p = 0.012$, such that higher deviant detection accuracy was observed for the ba/pa speech contrast, mean accuracy (collapsed across face contrast) = 93.92% (SD = 0.14) relative to the ba/a speech contrast mean accuracy (collapsed across face contrast) = 90.56% (SD = 2.24). We also observed a main effect of face context, $F(1, 44) = 15.79$, $p < 0.001$ such that higher accuracy was observed in the PX relative to the AV condition, mean accuracy for PX (collapsed across speech contrast condition) = 97.33% (SD = 0.08) relative to the mean accuracy for the AV (collapsed across contrast condition) = 87.14% (SD = 35.70). There were no other main effects or interactions. To further explore whether the face context manipulation differentially modulated oddball detection performance, planned comparisons were run to contrast the accuracy difference for the ba/a contrast condition as a function of face context and the ba/pa contrast as a function of face context. For the ba/a contrast condition, we found a significant difference between the AV and PX conditions $t(2, 44) = -4.69$, $p < 0.001$ such that participants were more accurate in the PX condition relative to the AV condition, as expected. For the ba/pa contrast, we also found a significant difference between the AV and PX conditions $t(2, 44) = -2.46$, $p = 0.02$, although this difference was numerically smaller. Accuracy for the ba/a contrast: AV = 83% (SD = 0.19); PX = 95% (SD = 0.17), mean accuracy difference = 12%; accuracy for the ba/pa contrast: AV = 90% (SD = 0.21); PX = 96% (SD = 0.15), mean accuracy difference = 6%. See Fig. 4 for accuracy by speech contrast and face context.

3.2. ERP Data

3.2.1. P300

To examine the effects of speech contrast (ba/a vs. ba/pa), face context (AV vs. PX) and P300 response, we ran a 2 × 2 × 2 repeated measures ANOVA with speech contrast (ba/a vs. ba/pa), face context (AV vs. PX) and stimulus (standard vs. deviant) as within subjects variables. This analysis revealed the predicted main effect of stimulus, $F(1, 44) = 127.91$, $p < 0.001$, $\eta^2 = 0.743$, such that deviants were more positive than standards across all conditions. Pairwise follow-up comparisons (two-tailed t-tests) revealed confirmed significant differences between standards and deviants for all speech contrasts

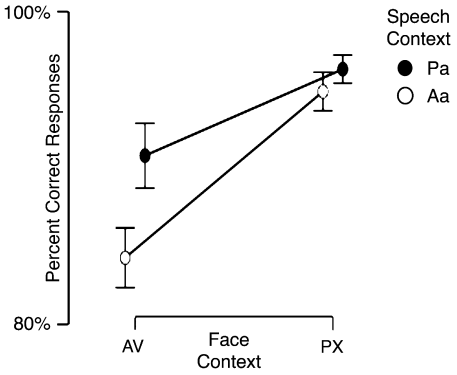

Figure 4. The percentage of correct behavioral responses (button presses) out of a possible 40 deviant trials in each face context.

and face contexts; /pa/ vs. /ba/ PX $t(44) = -7.737$, $p < 0.001$; /pa/ vs. /ba/ AV $t(44) = -8.047$, $p < 0.001$; /*a*/ vs. /ba/ PX $t(44) = -9.653$, $p < 0.001$; /*a*/ vs. /ba/ AV $t(44) = -6.813$, $p < 0.001$. We also observed a stimulus (standard/deviant) by face context interaction $F(1, 44) = 5.56$, $p = 0.023$, $\eta^2 = 0.112$, such that the amplitude difference between standards and deviants was greater when participants were viewing a pixelated face (relative to a speaking face). Further, the interaction between speech contrast and face context was marginal $F(1, 44) = 3.03$, $p = 0.089$, $\eta^2 = 0.064$, suggesting a trend for differential modulation of speech contrast by face context in overall amplitudes; however, the three-way interaction of speech contrast, face context and stimulus type was not significant $F(1, 44) = 2.356$, $p = 0.132$, $\eta^2 = 0.051$. Planned comparison t-tests are motivated by our hypothesis that face context would modulate the ba/a contrast to a greater degree than the ba/pa contrast. These contrasts (Table 1, top row) compared the size of the standard-deviant difference within each speech contrast as a function of face context (AV vs. PX). These comparisons revealed a significant difference for the ba/a contrast as a function of face context $t(44) = -2.972$, $p = 0.005$; but no effect of face context for the ba/pa contrast $t(44) = -0.431$, $p = 0.669$. Figures 4 and 5 clearly show a large amplitude difference between standards and deviants for the ba/a contrast as a function of face context, and no difference for the ba/pa contrast as a function of face context.

3.2.2. MMN

To examine the effects of speech contrast (more vs. less acoustically distinct), face context (audiovisual speech vs. audio only) and MMN response we ran a $2 \times 2 \times 2$ repeated measures ANOVA with contrast (ba/a vs. ba/pa), face context (AV, PX) and stimulus (standard, deviant) included as within subjects variables. This analysis revealed a main effect of stimulus $F(1, 44) = 6.712$,

Table 1.
Comparisons of standard and deviant values for P300 and MMN within speech contrast, within face context and overall. Boldface p values are significant (<0.005)

		t	df	p	Cohen's d
Within Speech Contrast Comparisons (contrast of the difference between standards and deviants within each speech contrast as a function of face context)	P300				
	ba/a: AV vs. PX	−2.972	44	**0.005**	−0.443
	ba/pa: AV vs. PX	−0.431	44	0.669	−0.064
	MMN				
	ba/a: AV vs. PX	−0.188	44	0.851	−0.028
	ba/pa: AV vs. PX	2.331	44	**0.024**	0.348
Within Face Context Comparisons (contrast of the difference between standards and deviants within each face context as a function of speech contrast)	P300				
	AV: aa vs. pa	−1.577	44	0.122	−0.235
	PX: aa vs. pa	0.480	44	0.634	0.072
	MMN				
	PX: aa vs. pa	1.343	44	0.186	0.200
	AV: aa vs. pa	−1.446	44	0.155	−0.216
Comparisons of all Standards and Deviants	P300				
	AV: ba vs. aa	−6.813	44	**<0.001**	−1.016
	PX: ba vs. aa	−9.653	44	**<0.001**	−1.439
	AV: ba vs. pa	−8.047	44	**<0.001**	−1.200
	PX: ba vs. pa	−7.737	44	**<0.001**	−1.153
	MMN				
	AV: ba vs. aa	1.787	44	0.081	0.266
	PX: ba vs. aa	1.532	44	0.133	0.228
	AV: ba vs. pa	−0.710	44	0.482	−0.106
	PX: ba vs. pa	2.815	44	**0.007**	0.420

$p = 0.013$, $\eta^2 = 0.132$, such that deviants were more negative than standards in the 200 to 375 ms following stimulus onset (see Figs 4 and 5). Follow-up pairwise comparisons revealed that this difference was significant for the ba/pa contrast in the PX face context $t(44) = 2.815$, $p = 0.007$; however this contrast was not statistically significant for any other contrast (all p's > 0.08, see Table 1). Thus, no statistically significant MMNs were elicited by the ba/a contrast. We also observed a marginal face context by stimulus interaction $F(1, 44) = 3.71$, $p = 0.061$, $\eta^2 = 0.07$, suggesting a trend for differential modulation of speech contrast by face context in overall amplitudes. Finally, we observed a significant three-way speech contrast by face context by stimulus interaction $F(1, 44) = 5.014$, $p = 0.03$, $\eta^2 = 0.07$, suggesting that the amplitude difference between standards and deviants was differentially modulated for our speech contrasts as a function of face context. Follow up t-tests revealed that the difference between standards and deviants was larger for the ba/pa contrast for the PX condition relative to the AV condition $t(44) = 2.331$,

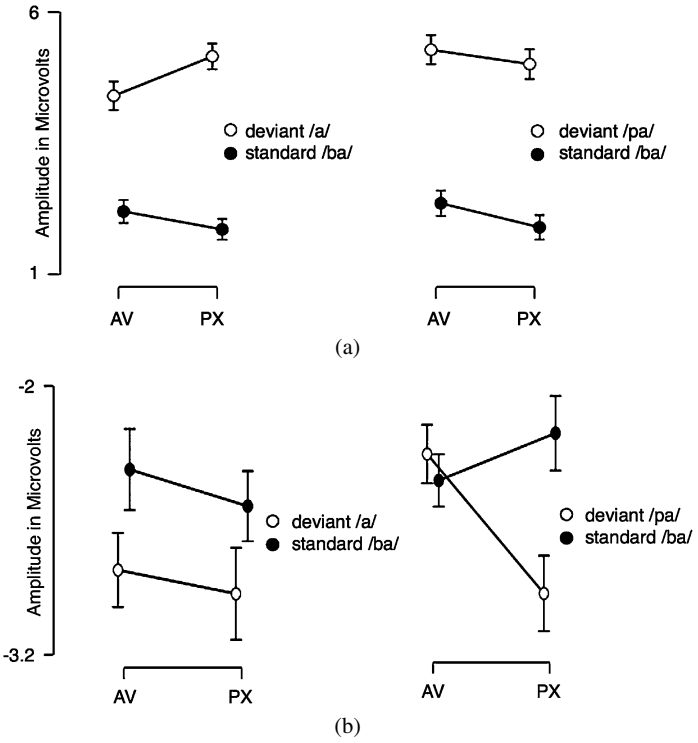

Figure 5. (a) Line graph showing P300 adaptive mean amplitudes in μV for each speech sound (/ba/, /p*a*/, and /a/) in both the AV and PX face contexts. (b) Line graph showing MMN adaptive mean amplitudes in μV for each speech sound (/ba/, /p*a*/, and /a/) in both the AV and PX face contexts.

$p = 0.024$, but no difference between standards and deviants as a function of face context in the ba/a contrast $t\,(44) = -0.881$, $p = 0.851$ (Table 1). Overall, minimal MMNs were observed for the ba/a contrast and the only significant pairwise MMN was observed for the ba/pa PX condition.

4. Discussion

Behavioral and electrophysiological data revealed hypothesized differences related to the difficulty of discriminating auditory stimuli as a function of the face context (pixelated face versus audiovisual speech). Behaviorally, although accuracy was generally quite high across all conditions, accuracy was higher when participants were detecting a deviant /pa/ among standard /ba/s relative to when participants were detecting an /*a*/ among standard /ba/s. This finding was not unexpected given that the /ba/ and /pa/ tokens are more acoustically distinct. Further, the presence of a speaking face modulated the effects of accuracy such that participants were less accurate overall when they had to

attend to both a speaking face and the acoustic speech information. Planned comparisons revealed that the accuracy benefit for the PX face context was numerically larger for the ba/a speech contrast condition. These findings suggest that the presence of a face producing /ba/ effectively restored phonemic information for the /*a*/, making participants less able to discriminate it from the full /ba/. However, the fact that the presence of visual speech reduced accuracy for both speech contrasts suggests either that the presence of identical visual speech made the contrasts harder to discriminate because they became more perceptually similar overall, or that the level of multi-sensory processing required for AV speech reduced overall performance.

Our neurophysiological data focus on the P300, a measure of identification and discrimination that is modulated by attention and working memory, and the MMN, a pre-attentive measure of discrimination. Our data reveal a large P300, with more positive amplitudes elicited by deviant relative to standard tokens in a large cluster of central electrodes between 400 ms and 700 ms post stimulus onset. Planned comparisons revealed that the P300 effect was larger in in the absence of visual speech for the ba/a speech contrast, which suggests that the face producing /ba/ supported phonemic restoration of the /*a*/, making it less distinct from the full /ba/. Critically, this same contrast for the ba/pa contrast was not significant. This overall pattern of effects is generally consistent with our accuracy data, however one noted difference is that the face context manipulation had a larger effect on behavioral performance for the ba/pa contrast than electrophysiological response — the P300 effect for the ba/pa contrast was not significantly different between the AV and PX conditions. The most likely explanation is that behavioral performance reflects a combination of multiple neural processes as well as signal transmission from the central nervous system to the peripheral nervous system, whereas any individual electrophysiological component reflects a more isolated neural process. Indeed, when we consider the MMN response, discussed below, we see a significant change in amplitude for the ba/pa contrast as a function of face context, which supports this interpretation.

With respect the MMN, we observed a small negative deflection for deviant relative to standard tokens between 200 ms and 375 ms in the same central electrode cluster. This MMN effect was modulated by a significant three-way interaction between stimulus (standard vs. deviant), face context and speech contrast, and follow-up pairwise comparisons revealed larger MMN effects for the PX relative to the AV condition for the ba/pa speech contrast, but no change in MMN effect for the ba/a contrast as a function of face context. Indeed, overall MMNs for the ba/a speech contrast condition were extremely small, in both face context conditions. Indeed, a significant MMN effect was only present for the ba/pa contrast in the PX condition. The lack of any significant MMN effects for the ba/a contrast suggest that this contrast may be

too subtle to be detected pre-attentively. However, it is possible that an MMN would be elicited for this contrast in the context of a completely passive task, as MMN responses are not always seen in an active oddball detection task. Further, the elimination of an MMN response for the ba/pa contrast in the presence of a speaking face is consistent with prior literature that suggest a MMN reduction for AV speech relative to auditory only speech (Bernstein *et al.*, 2001). Finally, with respect to comparisons between behavioral performance and electrophysiological response, which appeared somewhat incongruous for the ba/pa speech contrast when considering the P300 in isolation, the MMN effect difference for ba/pa as a function of face context (larger for PX) suggests a pre-attentive difference at the neural level that contributes to behavioral response. As such, at a minimum we can consider behavioral performance in this task to reflect a combination of our neural response measures.

Taken together, these findings speak to the potential utility of using an audiovisual phonemic restoration technique as an alternative approach to comparing audiovisual speech and auditory only speech processing using ERP. Specifically, we have tested a new method of assessing AV speech that does not require obvious cross-category mismatch or auditory noise. This technique may be particularly useful for populations who have significant challenges with other available AV methods. In the current paper we utilized an active task to obtain both behavioral and electrophysiological data from individuals who could respond with a button press. Moving forward, this approach may be adapted as a passive ERP task for populations that cannot overtly (by button press or verbally) respond to what was heard, such as infants and young children and adults with developmental disabilities.

References

Alcántara, J. I., Weisblatt, E. J., Moore, B. C. and Bolton, P. F. (2004). Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome, *J. Child Psychol. Psychiat.* **45**, 1107–1114.

Alho, K. (1995). Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes, *Ear Hear.* **16**, 38–51.

Alsius, A., Navarra, J., Campbell, R. and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands, *Curr. Biol.* **15**, 839–843.

Bebko, J. M., Weiss, J. A., Demark, J. L. and Gomez, P. (2006). Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism, *J. Child Psychol. Psychiat.* **47**, 88–98.

Bergeson, T. R. and Pisoni, D. B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation, in: *Handbook of Multisensory Processes*, G. Calvert, C. Sence and B. E. Stein (Eds), pp. 749–771. MIT Press, Cambridge, MA, USA.

Bernstein, L. E., Ponton, C. W. and Auer Jr, E. T. (2001). Electrophysiology of unimodal and audiovisual speech perception, in: *AVSP 2001 — International Conference on Auditory–Visual Speech Processing*, Aalborg, Denmark, pp. 50–55.

Bernstein, L. E., Auer, E. T., Wagner, M. and Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing, *Neuroimage* **39**, 423–435.

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer, version 5.3.39, University of Amsterdam, Amsterdam, Netherlands. Retrieved from http://www.praat.org/, September 26, 2016.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception, *J. Exp. Psychol. Hum. Percept. Perform.* **30**, 445–463.

Brancazio, L. and Miller, J. L. (2005). Use of visual information in speech perception: evidence for a visual rate effect both with and without a McGurk effect, *Percept. Psychophys.* **67**, 759–769.

Brancazio, L., Best, C. T. and Fowler, C. A. (2006). Visual influences on perception of speech and nonspeech vocal-tract events, *Lang Speech* **49**, 21–53.

Brancazio, L., Moore, D., Tyska, K., Burke, S., Cosgrove, D. and Irwin, J. (2015). McGurk-like effects of subtle audiovisual mismatch in speech perception, presented at the *27th Annual Convention of the Association for Psychological Science*, New York, NY, USA, May 23, 2015.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F. and Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory, *Clin. Neurophysiol.* **113**, 495–506.

Desjardins, R. N., Rogers, J. and Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks, *J. Exp. Child Psychol.* **66**, 85–110.

Eigsti, I. M. and Shapiro, T. (2003). A systems neuroscience approach to autism: biological, cognitive, and clinical perspectives, *Ment. Retard. Dev. Disabil. Res. Rev.* **9**, 205–215.

Erber, N. P. (1975). Auditory–visual perception of speech, *J. Speech Hear. Disord.* **40**, 481–492.

Ferree, T. C., Luu, P., Russell, G. S. and Tucker, D. M. (2001). Scalp electrode impedance, infection risk, and EEG data quality, *Clin. Neurophysiol.* **112**, 536–544.

Foss-Feig, J. H., Kwakye, L. D., Cascio, C. J., Burnette, C. P., Kadivar, H., Stone, W. L. and Wallace, M. T. (2010). An extended multisensory temporal binding window in autism spectrum disorders, *Exp. Brain Res.* **203**, 381–389.

Grant, K. W. and Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences, *J. Acoust. Soc. Am.* **108**, 1197–1208.

Green, K. (1994). The influence of an inverted face on the McGurk effect, *J. Acoust. Soc. Am.* **95**, 3014. DOI:10.1121/1.408802.

Iarocci, G., Rombough, A., Yager, J., Weeks, D. J. and Chua, R. (2010). Visual influences on speech perception in children with autism, *Autism* **14**, 305–320.

Irwin, J. R., Tornatore, L. A., Brancazio, L. and Whalen, D. H. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Dev.* **82**, 1397–1403.

Jerger, S., Damian, M. F., Tye-Murray, N. and Abdi, H. (2014). Children use visual speech to compensate for non-intact auditory speech, *J. Exp. Child Psychol.* **126**, 295–312.

Kaganovich, N., Schumaker, J., Leonard, L. B., Gustafson, D. and Macias, D. (2014). Children with a history of SLI show reduced sensitivity to audiovisual temporal asynchrony: an ERP study, *J. Speech Lang. Hear. Res.* **57**, 1480–1502.

Kaganovich, N., Schumaker, J. and Rowland, C. (2016). Matching heard and seen speech: an ERP study of audiovisual word recognition, *Brain Lang.* **157**, 14–24.

Kashino, M. (2006). Phonemic restoration: the brain creates missing speech sounds, *Acoust. Sci. Technol.* **27**, 318–321.

Klucharev, V., Möttönen, R. and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception, *Cogn. Brain Res.* **18**, 65–75.

Lachs, L., Pisoni, D. B. and Kirk, K. I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: a first report, *Ear Hear.* **22**, 236–251.

Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds, *Infant Behav. Dev.* **13**, 343–354.

Lewkowicz, D. J. and Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech, *Proc. Natl Acad. Sci. USA* **109**, 1431–1436.

MacDonald, J. and McGurk, H. (1978). Visual influences on speech perception processes, *Atten. Percept. Psychophys.* **24**, 253–257.

MacLeod, A. and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise, *Br. J. Audiol.* **21**, 131–141.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**(5588), 746–748.

Meltzoff, A. N. and Kuhl, P. K. (1994). Faces and speech: intermodal processing of biologically relevant signals in infants and adults, in: *The Development of Intersensory Perception: Comparative Perspectives*, D. J. Lewkowicz and R. Lickliter (Eds), pp. 335–369. Erlbaum, Hillsdale, NJ, USA.

Ménard, L., Dupont, S., Baum, S. R. and Aubin, J. (2009). Production and perception of French vowels by congenitally blind adults and sighted adults, *J. Acoust. Soc. Am.* **126**, 1406–1414.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E. and Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study, *Cogn. Brain Res.* **14**, 115–128.

Nath, A. R. and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion, *Neuroimage* **59**, 781–787.

Payton, K. L., Uchanski, R. M. and Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing, *J. Acoust. Soc. Am.* **95**, 1581–1592.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception, *J. Speech Lang. Hear. Res.* **52**, 1073–1081.

Pizzagalli, D. A. (2007). Electroencephalography and high-density electrophysiological source localization, in: *Handbook of Psychophysiology*, 3rd edn., J. T. Cacioppo, L. G. Tassinary and G. G. Berntson (Eds), pp. 56–84. Cambridge University Press, Cambridge, UK.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b, *Clin. Neurophysiol.* **118**, 2128–2148.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon, *Curr. Dir. Psychol. Sci.* **17**, 405–409.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments, *Cereb. Cortex* **17**, 1147–1153.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W. and Foxe, J. J. (2007). Seeing voices: high-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion, *Neuropsychologia* **45**, 587–597.

Samuel, A. G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion, *J. Exp. Psychol. Hum. Percept. Perform.* **7**, 1124–1131.

Schwartz, J. L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent, *J. Acoust. Soc. Am.* **127**, 1584–1594.

Smith, E. G. and Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism, *J. Child Psychol. Psychiat.* **48**, 813–821.

Soto-Faraco, S. and Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion, *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 580–587.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Am.* **26**, 212–215.

Tremblay, K., Kraus, N., McGee, T., Ponton, C. and Otis, B. (2001). Central auditory plasticity: changes in the N1–P2 complex after speech-sound training, *Ear Hear.* **22**, 79–90.

Van Wassenhove, V., Grant, K. W. and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech, *Proc. Natl Acad. Sci. USA* **102**, 1181–1186.

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

Warren, R. M. (1970). Perceptual restoration of missing speech sounds, *Science* **167**, 392–393.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion, *J. Mem. Lang.* **50**, 212–230.