HHS Public Access

Author manuscript

*Brain Lang*. Author manuscript; available in PMC 2017 June 01.

Published in final edited form as:

*Brain Lang*. 2016 ; 157-158: 14–24. doi:10.1016/j.bandl.2016.04.010.

# Matching Heard and Seen Speech: An ERP Study of Audiovisual Word Recognition

**Natalya Kaganovich**[*,1,2], **Jennifer Schumaker**[1], and **Courtney Rowland**

[1]Department of Speech, Language, and Hearing Sciences, Purdue University, 715 Clinic Drive, West Lafayette, IN 47907-2038

[2]Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2038

## Abstract

Seeing articulatory gestures while listening to speech-in-noise (SIN) significantly improves speech understanding. However, the degree of this improvement varies greatly among individuals. We examined a relationship between two distinct stages of visual articulatory processing and the SIN accuracy by combining a cross-modal repetition priming task with ERP recordings. Participants first heard a word referring to a common object (e.g., pumpkin) and then decided whether the subsequently presented visual silent articulation matched the word they had just heard. Incongruent articulations elicited a significantly enhanced N400, indicative of a mismatch detection at the pre-lexical level. Congruent articulations elicited a significantly larger LPC, indexing articulatory word recognition. Only the N400 difference between incongruent and congruent trials was significantly correlated with individuals' SIN accuracy improvement in the presence of the talker's face.

## 1. Introduction

Seeing a talker's face considerably improves speech-in-noise (SIN) perception in both children and adults (Barutchu et al., 2010; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954; Tye-Murray, Spehar, Myerson, Sommers, & Hale, 2011), with facial speech gestures providing both redundant and complementary information about the content of the auditory signal. Indeed, recent studies show that a decrease in the SIN ratio leads to greater visual fixations on the mouth of the speaker (Yi, Wong, & Eizenman, 2013) and stronger synchronizations between the auditory and visual motion/motor brain regions (Alho et al., 2014). Importantly, however, the degree to which individuals benefit from visual speech cues varies significantly (Altieri & Hudock, 2014; Grant, Walden, & Seitz, 1998). Reasons for such variability may be many. As an example, Grant and colleagues

[*]Corresponding author: Department of Speech, Language, and Hearing Sciences, Purdue University, Lyles Porter Hall, 715 Clinic Drive, West Lafayette IN 47907-2038, Phone (765)494-4233, Fax (765)494-0771, kaganovi@purdue.edu.

proposed that variability in the processing of either auditory or visual modality as well as in the audiovisual integrative mechanisms may independently contribute to the degree of improvement for audiovisual as compared to auditory only speech (Grant et al., 1998).

In this study, we focused on individual variability in matching auditory words with their silent visual articulations – the skill that is at the heart of audiovisual speech perception - and asked which aspects of such matching process play a role in improved SIN perception when seeing the talker's face. Just like auditory words, visual articulations unfold over time, and their processing is incremental in nature. Viewers may detect mismatches between auditory and articulatory information in the observed facial movements on a sub-lexical level (i.e., based on syllabic and/or phonological processing) well before the completion of the entire articulatory sequence associated with a particular word. However, because many word articulations differ only in the final segments (e.g., beam vs. beet), the unequivocal decision about a match requires that the entire sequence of facial speech gestures associated with a word is completed and coincides with the articulatory word recognition. Hypothetically, either or both stages of processing facial articulatory gestures could play a role in improving SIN perception. Because facial speech gestures typically precede the onset of sound (e.g., Conrey & Pisoni, 2006; Grant, van Wassenhove, & Poeppel, 2004; McGrath & Summerfield, 1985; but see also Schwartz & Savariaux, 2014; van Wassenhove, Grant, & Poeppel, 2007), they allow listeners to make predictions about the incoming linguistic information. Higher sensitivity to correspondences between facial speech gestures and sub-lexical units may enable more accurate predictions about the auditory signal and/or a detection of a mismatch between one's prediction and the actual sound. On the other hand, within the context of discourse, the main semantic information is carried by words. It is possible, therefore, that only the recognition of the entire articulatory sequence as a word would result in greater SIN benefit.

The notion that the degree of SIN improvement in the presence of the talker's face may depend on the level of linguistic analysis is supported by earlier research. For example, Grant and Seitz (Grant & Seitz, 1998) reported that their measures of audiovisual benefit for SIN during nonsense syllable and sentence perception did not correlate. In a similar vein, the study by Stevenson and colleagues (Stevenson et al., 2015) showed that healthy elderly adults benefit from visual speech cues during a SIN task as much as younger adults when presented with individual phonemes but show marked deficits when presented with individual words. While better understanding of how facial speech gestures facilitate SIN perception at different linguistic levels is needed, the above studies suggest that the mechanisms engaged at each level may be at least partially distinct.

In order to examine unique contributions of matching auditory and visual speech information at the sub-lexical and lexical level to the SIN accuracy, we combined a cross-modal repetition priming task with event-related potentials recordings (ERPs). The ERP technique's excellent temporal resolution allows one to tease apart perceptual and cognitive processes that jointly shape behavioral performance. We were, therefore, able to evaluate ERP responses associated with audiovisual matching at the sub-lexical level separately from the ERP responses associated with articulatory word recognition and correlate both measures with individuals' performance on the SIN task.

In the cross-modal repetition-priming task, participants first heard a word referring to a common object (such as a pumpkin) and then had to decide whether the subsequently presented visual silent articulation matched the word they had just heard. In half of trials, the presented articulation matched the heard words (congruent trials), and in another half it did not (incongruent trials). The important aspect of this paradigm is that in absolute terms, no trial contained a true repetition of the same physical stimulus since the first word was always presented in the auditory modality only and the second word in the visual modality only. On congruent trials, the seen articulation was expected to be perceived as a match to the auditory word and lead to the articulatory word recognition. On incongruent trials, a mismatch between the expected and the observed articulation would be detected. The ERP components associated with word repetition (including cross-modal presentations) – the N400 and the late positive complex (LPC) – have been well-studied and allow for clear predictions and interpretation of the results as described below.

The N400 ERP component is a negative waveform deflection that peaks at approximately 400 ms post-stimulus onset in young healthy adults and has a centro-parietal distribution. This component is thought to index the ease with which long-term semantic representations may be accessed during processing (for reviews, see Duncan et al., 2009; Holcomb, Anderson, & Grainger, 2005; Kutas & Federmeier, 2011; Kutas & Van Petten, 1988, 1994). However, and more germane to the topic of the current study, the N400 amplitude is also sensitive to phonological correspondences between prime and target words in priming tasks (Praamstra, Meyer, & Levelt, 1994; Praamstra & Stegeman, 1993), with greater negativity to phonological mismatches. Importantly, a study by Van Petten and colleagues demonstrated that the onset of the N400 component precedes the point at which words can be reliably recognized (Van Petten, Coulson, Rubin, Plante, & Parks, 1999), suggesting that this component is elicited as soon as enough information has been processed to determine that the incoming signal either matches or mismatches the expected one.

Based on the above properties of the N400 component, we predicted that incongruent visual articulations would elicit larger N400s compared to congruent visual articulations. Additionally, because all incongruent word pairs differed at the word onset, we expected that the N400 amplitude increase to incongruent articulations would reflect a relatively early process of detecting an expectancy violation, likely prior to the articulatory word recognition. Lastly, if sensitivity to audiovisual correspondences at the sub-lexical level plays a role in SIN perception, we expected that individuals with greater N400 differences between incongruent and congruent trials would show better improvement on the SIN task when seeing the talker's face.

The LPC ERP component belongs to a family of relatively late positive deflections in the ERP waveform that may vary in distribution and amplitude depending on the task used. Of particular relevance to our paradigm is the sensitivity of this component to word repetition (for reviews, see Friedman & Johnson, 2000; Rugg & Curran, 2007). More specifically, the LPC is larger (i.e., more positive) to repeated as compared to not repeated words (e.g., Neville, Kutas, Chesney, & Schmidt, 1986; Paller & Kutas, 1992), suggesting that it indexes some aspects of the recognition process. We hypothesized that the LPCs to congruent articulations should have larger amplitude than the LPCs to incongruent articulations, which

were not expected to result in the articulatory word recognition on a regular basis. Furthermore, if recognition of the entire articulatory sequence as a specific word is important for SIN, we expected that those individuals with the largest LPC differences between congruent and incongruent articulations would show the best improvements on the SIN task when seeing the talker's face.

## 2. Method

### 2.1 Participants

Twenty-two college-age adults participated in the study for pay. They had normal hearing (tested at 500, 1000, 2000, 3000, and 4000 Hz at 20 dB SPL), normal or corrected to normal visual acuity, and normal non-verbal intelligence (Brown, Sherbenou, & Johnsen, 2010). According to the Laterality Index of the Edinburgh Handedness Questionnaire, two participants were ambidextrous, and the rest were right-handed (Cohen, 2008; Oldfield, 1971). All gave their written consent to participate in the experiment. The study was approved by the Institutional Review Board of Purdue University, and all study procedures conformed to The Code of Ethics of the World Medical Association (Declaration of Helsinki) (1964).

### 2.2 Stimuli and Experimental Design

The study consisted of two experiments. In the first (referred to henceforth as the Matching task), participants decided whether visual only articulation matched the word they had just heard. Each trial consisted of the following events (see Figure 1). Participants first saw a color picture of a common object/person (e.g., toys, mailman, etc.). While the image was still on the screen and 1000 ms after its appearance, participants heard the object named (e.g., they heard a female speaker pronounce a word "toys" or "mailman," etc.). The image continued to stay on the screen for another 1000 ms after the offset of the sound and then disappeared. A blank screen followed for another 1000 ms. Next, a video of a female talker was presented. It consisted of a static image of the talker's face taken from the first frame of the video (1000 ms), followed by a silent articulation of a word, followed by the static image of the talker's face taken from the last frame of the video (1000 ms). In half of all trials, the talker's articulation matched the previously heard word (congruent trials; for example, participants saw the talker articulate "toys" after hearing the word "toys" earlier), while in another half, the talker's articulation clearly mismatched the previously heard word (incongruent trials; for example, participants saw the talker say "bus" after hearing the word "toys" earlier). Another blank screen followed for 1000 ms before a response window started. The onset of a response window was signaled by the appearance of the screen with "Same?" written across it. The response window lasted for 2000 ms, during which participants had to determine whether or not the articulation they saw matched the word they had heard. The disappearance of the "Same?" screen concluded the trial. Trials were separated by a temporal period randomly varying between 1000 and 1500 ms. Responses were collected via a response pad (RB-530, Cedrus Corporation), with the response hand counterbalanced across participants.

Each participant completed 96 trials (48 congruent and 48 incongruent). Ninety-six words from the MacArthur Bates Communicative Developmental Inventories (Words and Sentences) (Fenson et al., 2007) were used as stimuli. The Inventory is a parent-filled questionnaire that is designed to assess vocabulary development in children between 16 and 30 months of age. It consists of words that children of this age can be expected to produce. All words contained 1–2 morphemes and were 1 to 2 syllables in length with two exceptions – "elephant" and "teddy bear." Words contained between 1 and 8 phonemes, with diphthongs counted as 1 phoneme. We used the MacArthur Bates Inventories because we planned to use this paradigm not only with adults but also with children. However, the absolute majority of the selected words are also frequent words in adults' vocabulary. According to the Subtlex-US database of word frequencies (Brysbaert & New, 2009; *The SUBTL Word Frequency*, 2009), the mean frequency of the used words was 67.862 per million, with frequencies ranging from 0.55 per million (for "jello") to 557.12 per million (for "girl"). The frequency for the compound "teddy bear" was not available in the Subtlex-US database; however, its individual words – "teddy" and "bear" – had a frequency of 15.9 and 57.41 per million respectively. Notably, only 2 out of 96 words had a frequency of less than 1 per million ("sprinkler" and "jello"). The influence of the word frequency, length, and complexity was controlled by counterbalancing word presentation in congruent and incongruent trials across participants (see below).

Each of the words was matched with a color picture from the Peabody Picture Vocabulary Test (pictures were used with the publisher's permission) (Dunn & Dunn, 2007) that exemplified the word's meaning (for example, a picture of toys was matched with the word "toys"). The sole goal of pictures was to serve as fixation points in order to better maintain participants' attention and minimize eye movements. For incongruent trials, 48 pairs of words were created in such a way that their visual articulation differed significantly during the word onset. In most cases (35 out of 48 pairs), this was achieved by pairing words in which the first consonants differed visibly in the place of articulation (e.g., belt vs. truck). In 6 pairs, the first vowels of the words differed in the shape and the degree of mouth opening (e.g., donkey vs. candy). In the remaining 7 pairs, the first sounds were a labial consonant in one word (i.e, required a mouth closure (e.g., pumpkin)) and a vowel (i.e., required a mouth opening (e.g., airplane)) in another word. Heard and articulated words in incongruent pairs had no obvious semantic relationship. Two lists containing 48 congruent and 48 incongruent heard vs. articulated word presentations were created in such a way that articulations that were congruent in list A were incongruent in list B. As a result, across all participants, we collected behavioral and ERP responses to the same articulations, which were perceived as either congruent or incongruent. Lastly, 10 different versions of list A and 10 different versions of list B were created by randomizing the order of 96 trials. Each participant completed only one version of one list (e.g., participant 1 did list A version 1; participant 2 did list B version 1; participant 3 did list A version 2, participant 4 did list B version 2, etc.) Version 1 of lists A and B is shown in the Appendix. This task was combined with electroencephalographic (EEG) recordings (see below), for which the first video frame with noticeable articulation movements served as time 0 for averaging.

In order to determine how many of the silent articulations could be recognized by our participants in the incongruent condition and to evaluate their lip-reading abilities (which are

often thought to contribute to SIN perception), we selected 20 silent articulations from the list of 96 used and asked each participant (in a separate session) to provide their best guess as to what word they thought the speaker was producing. The list of 20 words used for this task is shown in Table 1. In order to select words that reflected the diversity of lexical items used for the main task, this set of words included both one- and two-syllable words and contained items that started with either a labial (closed mouth) or an alveolar (open mouth) sound. No cues to the words' identity were provided. This part of the Matching task is referred to henceforth as the lip-reading component of the Matching task.

All words were pronounced by a female talker dressed in a piglet costume. Participants were told that the paradigm was designed to be child-friendly. The actor's mouth area was left free of makeup except for bright lipstick and therefore did not in any way obscure natural muscle movements of the lower face during articulation. Words were recorded with a Marantz digital recorder (model PMD661) and an external microphone (Shure Beta 87) at a sampling rate of 44,100 Hz. Sound files were edited in the Praat software (Boersma & Weenink, 2011) so that the onset and offset of sound were preceded by 50 ms of silence. Final sound files were root-mean-square normalized to 70 dB. All videos were recorded with the Canon Vixia HV40 camcorder. The video's frame per second rate was 29.97. The audio track of the video recording was then removed in Adobe Premier Pro CS5 (Adobe Systems Incorporated, USA). Articulation portions of videos ranged from 1133 ms (for "car") to 1700 ms (for "sandbox").The audio recording with the Marantz recorder happened simultaneously with the video recording. Therefore, all auditory words presented as primes were true matches to the silent articulations presented as targets. Stimulus presentation and response recording was controlled by the Presentation program (www.neurobs.com). Each video was edited so that it started one frame prior to the onset of the first noticeable articulation movement.

In the second experiment (referred to henceforth as the speech-in-noise (SIN) task), participants were asked to listen to the same 96 words used in the Matching task. However, this time these words were embedded in a two-talker babble masker. The masker consisted of two female voices reading popular children's stories. One sample was 3 minutes and 8 seconds long (by talker 1), and the other was 3 minutes and 28 seconds long (by talker 2). Both samples were manually edited in Praat to remove silent pauses greater than 300 ms and then repeated without discontinuity. The streams from the two talkers were root-mean-square normalized to 75 dB, mixed, and digitized using a resolution of 32 bits and a sampling rate of 24.414 kHz. The 96 stimuli words were root-mean-square normalized to 70 dB, resulting in the −5 dB signal-to-noise ratio.

A schematic representation of the SIN trial is shown in Figure 2. This task had 2 conditions – auditory only (A) and audiovisual (AV), which were administered on two separate days. The order of A and AV conditions was counterbalanced across participants, but each participant did both. The babble masker started 3 seconds prior to the first trial and was presented continuously until the end of the experiment. In the AV condition, participants saw videos of a talker producing each of 96 words. Each video was preceded and followed by a static image of a talker with a closed mouth, which lasted for 1,000 ms. In the A condition, the same static images of the talker were present; however, the video portion was replaced

with an image of the talker with her mouth open (see Figure 2). The appearance of the open-mouth picture in the A condition thus cued participants to the fact that the onset of the target word was imminent without providing any visual cues to its identity. Previous research shows that visual cues that reliably predict the onset of the auditory signal significantly improve the latter's detection threshold (ten Oever, Schroeder, Poeppel, van Atteveldt, & Zion-Golumbic, 2014). The inclusion of the cue to the target word onset in the A condition aimed to make the attentional demands of the A and AV conditions more similar and to ensure that the remaining differences would be due to the presence of the articulatory movements in the AV condition. Word presentations in both conditions were separated by 3 seconds, during which participants provided their verbal response about what they had heard. When unsure, participants were encouraged to give their best guess or to say "I don't know." No EEG recordings were collected during this task.

All testing occurred over 3 sessions administered on 3 different days. The lip-reading component of the Matching task and one of the SIN conditions (either A or AV) were administered during the first session, the Matching task during the second session, and the second SIN condition during the third session. Because the same words were used in the Matching task and in the SIN task, most participants' sessions were separated by at least 7 days to minimize the possible effect of stimulus repetition.

### 2.3 ERP Recordings and Data Analysis

During the Matching task, the EEG data were recorded from the scalp at a sampling rate of 512 Hz using 32 active Ag-AgCl electrodes secured in an elastic cap (Electro-Cap International Inc., USA). Electrodes were positioned over homologous locations across the two hemispheres according to the criteria of the International 10-10 system (American Electroencephalographic Society, 1994). The specific locations were as follows: midline sites Fz, Cz, Pz, and Oz; mid-lateral sites FP1/FP2, AF3/AF4, F3/F4, FC1/FC2, C3/C4, CP1/CP2, P3/P4, PO3/PO4, and O1/O2; and lateral sites F7/F8, FC5/FC6, T7/T8, CP5/CP6, and P7/P8; and left and right mastoids. EEG recordings were made with the Active-Two System (BioSemi Instrumentation, Netherlands), in which the Common Mode Sense (CMS) active electrode and the Driven Right Leg (DRL) passive electrode replace the traditional "ground" electrode (Metting van Rijn, Peper, & Grimbergen, 1990). During recording, data were displayed in relationship to the CMS electrode and then referenced offline to the average of the left and right mastoids (Luck, 2005). The Active-Two System allows EEG recording with high impedances by amplifying the signal directly at the electrode (BioSemi, 2013; Metting van Rijn, Kuiper, Dankers, & Grimbergen, 1996). In order to monitor for eye movement, additional electrodes were placed over the right and left outer canthi (horizontal eye movement) and below the left eye (vertical eye movement). Horizontal eye sensors were referenced to each other, while the sensor below the left eye was referenced to FP1 in order to create electro-oculograms. Prior to data analysis, EEG recordings were filtered between 0.1 and 30 Hz. Individual EEG records were visually inspected to exclude trials containing excessive muscular and other non-ocular artifacts. Ocular artifacts were corrected by applying a spatial filter (EMSE Data Editor, Source Signal Imaging Inc., USA) (Pflieger, 2001). ERPs were epoched starting at 200 ms pre-stimulus and ending at 1800 ms post-stimulus onset. The 200 ms prior to the stimulus onset served as a baseline. The onset of

articulation was used as time zero for averaging. On average, 45 clean trials (range 40–48) were collected from each participant in congruent (*SD*=2) and incongruent (*SD*=1.9) conditions.

The N400 component was measured as the mean amplitude between 300 and 550 ms post-stimulus onset in agreement with a multitude of earlier studies (for a comprehensive review, see Kutas & Federmeier, 2011). The reported latency of the LPC component, however, varies significantly from study to study. In order to select the LPC mean amplitude measurement window more objectively, we compared ERPs to congruent and incongruent articulations in a series of t-tests conducted on consecutive data points between 660 ms post-stimulus onset (the onset of the LPC component based on the visual inspection of the grand average ERP waveform) and the end of the epoch (1800 ms post-stimulus onset) at the PZ site (which showed the largest LPC amplitude (see Figure 3). The false discovery rate (FDR) method was used to correct for multiple comparisons, with family-wise error rate set to 0.05 (Groppe, Urbach, & Kutas, 2011). Prior to t-tests, the ERP data were down sampled to 100 Hz, which resulted in one measurement point for each 10 ms of recording. This analysis revealed that only one time period had more than 2 consecutive data pairs that did not survive the FDR correction – namely, between 1740 and 1800 ms post-stimulus onset. Therefore, we selected the time window between 660 and 1740 ms for measuring the mean amplitude of the LPC component across the entire set of the mid-line and mid-lateral electrode sites.

Paired samples t-tests were used to evaluate whether congruent and incongruent conditions of the Matching Task differed in the number of correct responses, incorrect responses, misses, and in reaction time. Repeated-measures ANOVAs were used to evaluate the mean amplitude of the N400 and the LPC ERP components. Both components had a broad scalp distribution. Preliminary analyses indicated that neither component showed a laterality effect; therefore, mid-lateral and midline sites were analyzed together in one ANOVA. It contained the factors of congruence (congruent vs. incongruent), anterior to posterior scalp distribution (AF and F sites, FC and C sites, CP and P sites, PO and O sites), and site (with 5 sites – 4 mid-lateral and one midline – in each of the anterior to posterior sections). In all statistical analyses, significant main effects with more than two levels were evaluated with a Bonferroni post-hoc test. In such cases, the reported *p* value indicates the significance of the Bonferroni test, rather than the adjusted alpha level. When omnibus analysis produced a significant interaction, it was further analyzed with step-down ANOVAs, with factors specific to any given interaction. Mauchly's test of sphericity was used to check for the violation of sphericity assumption in all repeated-measures tests that included factors with more than two levels. When the assumption of sphericity was violated, we used the Greenhouse-Geisser adjusted *p*-values to determine significance. Accordingly, in all such cases, adjusted degrees of freedom and the epsilon value (ε) are reported. Effect sizes, indexed by the partial eta squared statistic ($\eta_p^2$), are reported for all significant repeated-measures ANOVA results.

### 2.4 Correlations

In order to examine a relationship between the ERP indices of visual speech processing on the one hand and behavioral measures of visual speech perception (i.e., SIN improvement in the presence of the talker's face and accuracy on the lip-reading component of the Matching task) on the other hand, we conducted correlation analyses separately for the N400 and the LPC components. The ERP value used for each participant was the mean difference voltage (incongruent-congruent for N400 and congruent-incongruent for LPC) averaged over all sites with a significant congruency effect.

Additionally, because previous research shows that repeated presentations of words lead to the reduction of the LPC component, we examined the relationship between the LPC mean amplitude elicited by congruent articulations (which may have been perceived as a word repetition) and the same behavioral measures of visual speech processing as described above.

All correlation analyses reflect planned comparisons with unidirectional predictions (as described above). Therefore, one-tail t-tests were used to determine significance.

## 3. Results

### 3.1 Matching Task

**3.1.1 Behavioral Results—**Participants performed the Matching task with high accuracy. They were correct on 96.88% of congruent trials (*SD*=3.06) and on 97.54% of incongruent trials (*SD*=2.38). They failed to respond in only 0.36% of cases for congruent trials (*SD*=0.73) and in 0.45% of cases for incongruent trials (*SD*=1.18). Finally, the reaction time was also very comparable for congruent and incongruent presentations (598.81 ms (*SD*=177.4) and 616.68 ms (*SD*=211.2), respectively). Paired-samples t-tests were used to evaluate potential differences between performance on congruent and incongruent trials. None of the comparisons were significant: number of accurate responses ($t(21)=-0.734$, $p=0.471$), misses ($t(21)=-0.439$, $p=0.665$), incorrect responses ($t(21)=1.041$, $p=0.31$), and reaction time ($t(21)=-1.253$, $p=0.224$). All $p$ values are two-tailed. The lip-reading component was challenging for most participants. The mean accuracy on this task was 20.8% (range 0% to 45%, *SD*=12.9). Although the lowest score was 0%, only two participants performed this poorly.

**3.1.2 ERP Results—**ERPs elicited by congruent and incongruent trials are shown in Figure 3. Panel A directly overlays ERPs elicited by congruent and incongruent articulations. To better isolate the N400 and LPC components, we subtracted ERPs elicited by congruent articulations from the ERPs elicited by incongruent articulations. The resultant difference wave is shown in Panel B. Both components of interest are marked on the Pz site. Below we summarize all significant findings related to the effect of congruency and its interactions with other factors.

Analysis of the N400 mean amplitude revealed that this component was larger to incongruent compared to congruent articulations ($F(1,21)=8.043$, $p=0.01$, $\eta_p^2=0.277$). The effect of congruency interacted with the anterior to posterior distribution

($F$(1.368,28.735)=7.977, $p$=0.005, $\eta_p^2$=0.275, ε=0.456). Follow-up tests showed that the amplitude of N400 was larger to incongruent trials only over CP/P and PO/O regions (congruency: AF/F, $F$(1,21)<1; FC/C, $F$(1,21)=2.442, $p$=0.133; CP/P, $F$(1,21)=10.159, $p$=0.004, $\eta_p^2$=0.326; PO/O, $F$(1,21)=16.625, $p$=0.001, $\eta_p^2$=0.442).

The mean amplitude of the LPC component was significantly larger to congruent compared to incongruent articulations ($F$(1,21)=9.044, $p$=0.007, $\eta_p^2$=0.301). Similarly to the analysis of the N400 component, the effect of congruency interacted with the anterior to posterior distribution factor ($F$(2.015,42.311)=8.319, $p$=0.001, $\eta_p^2$=0.284, ε=0.672), with larger LPC to congruent articulations over the FC/C ($F$(1,21)=6.786, $p$=0.017, $\eta_p^2$=0.244), CP/P ($F$(1,21)=11.378, $p$=0.003, $\eta_p^2$=0.351) and PO/O ($F$(1,21)=16.584, $p$=0.001, $\eta_p^2$=0.441) regions, but not over the AF/F ($F$(1,21)=2.238, $p$=0.15) region.

### 3.2 Speech-In-Noise Task

The SIN data from 2 participants were not available because they failed to complete all testing sessions. Participants' accuracy was on average 61.6% in the A condition ($SD$=7.7) and 90.5% in the AV condition ($SD$=4.2), with a 28.9% ($SD$=8.7) improvement in the SIN perception in the presence of the talker's face. Importantly, each participant demonstrated a marked enhancement in the AV condition without reaching the ceiling level.

### 3.3 Correlations

Two participants with missing SIN data were excluded from correlation analyses. Additionally, we examined the presence of outliers in our data by means of the standardized DFBeta function in the SPSS Statistics program. This function tests the influence of individual cases on a regression model. When the regression model is stable, excluding any one case should not have a significant influence on the outcome. Cases with the standardized DFBeta values over 1 were considered to have a significant influence over the model and were excluded from analyses (Field, 2009). Based on this criterion, only 1 case was excluded from the analysis between the N400 difference measure and the degree of the SIN improvement in the AV condition (with the DFBeta value of −1.34). The results of the correlation analyses are shown in Figure 4.

Correlation analyses yielded two significant findings. First, individuals with larger N400 differences between incongruent and congruent trials improved more on the SIN task in the AV compared to the A condition ($r$=-0.444, $p$=0.028, 95% confidence interval= −0.747-0.013). Second, individuals with smaller LPC to congruent articulations were more accurate on the lip-reading component of the matching task ($r$=−0.404, $p$=0.031, 95% confidence interval=−0.705-0.021).

Lastly, in order to determine whether our two behavioral measures of visual speech perception were related, we conducted a correlation analysis between the degree of the SIN improvement in the AV condition and lip-reading accuracy. We found that these measures were *not* related ($r$=0.111, $p$=0.321, 95% confidence interval=−0.349–0.528).

## 4. Discussion

We examined individual variability in two distinct cognitive processes associated with matching facial articulatory movements with auditory words - namely, detecting auditory/articulatory mismatches at the pre-lexical level and articulatory word recognition. We then evaluated how each of these stages of processing contributes to enhancement in SIN perception in the presence of the talker's face. We reported two key electrophysiological findings. First, those articulations that were incongruent with the preceding auditory words elicited significantly larger N400s than congruent articulations, reflecting the detection of the auditory-articulatory mismatch. Second, congruent articulations elicited significantly larger LPCs compared to incongruent articulations, indicative of the articulatory word recognition. Most importantly, only the amplitude of the N400 (measured as a difference between incongruent and congruent articulations) was significantly correlated with individuals' improvement on SIN in the presence of the talker's face.

Given the relatively early onset of the N400 effect in our data (especially in view of the considerable length of visual articulations), this ERP component likely reflects the detection of a mismatch between articulatory gestures and sub-lexical speech units (such as syllables or individual phonemes). In our data set, all visual articulations that did not match auditory words differed from expected articulations during the word onset, which also helps explain the N400's relatively early latency. What perceptual or cognitive mechanisms underlie individual variability in detecting the auditory-articulatory mismatches requires future studies. One possibility is that the nature of the task encouraged participants to mentally visualize the articulation of the word they heard at the onset of each trial and then compare it to the seen articulation. If so, individual differences in the N400 mean amplitude may reflect differences in the strength of long-term memory traces for how speech sounds look when articulated, with stronger traces leading to a greater neural response to the observed mismatch. During an SIN task, strong memory traces for articulatory gesture-phoneme correspondences may facilitate lexical processing by selectively activating only those lexical representations that match the observed articulation. Individuals with weaker traces would not be expected to benefit from the presence of the talker's face to the same degree.

Alternatively, individual variability in the N400 amplitude may reflect the strength or accuracy of the auditory representation of the heard word maintained in the working memory while observing silent articulations. Such auditory representations are thought to be maintained by means of sub-vocal motor processes (Hickok, Okada, & Serences, 2009). They may then be mapped onto observed visual articulations as the latter unfold. However, even in this case, the accurate matching of speech sounds onto articulatory gestures presupposes some prior knowledge of how speech sounds look when articulated. This knowledge would then facilitate individuals' SIN perception.

Somewhat surprisingly, the N400 difference between congruent and incongruent articulations was not at all related to participants' lip-reading ability as measured by the lip-reading component of the Matching task. Nor was the lip-reading skill related to the SIN performance. As mentioned in the Introduction, the study by Van Petten and colleagues (Van Petten, Coulso, Rubin, Plante, & Parks, 1999) showed that the N400 component is generated

as soon as sufficient sensory information is processed to determine whether or not the incoming signal matches the expectation. The lip-reading task, on the other hand, requires a definitive identification of the word based entirely on its silent articulation, which likely leads to further neural processing. More generally, our results suggest that the neural mechanisms engaged during the initial detection of a mismatch between the expected and the seen articulation are at least partially different from the neural mechanisms engaged during lip-reading. This finding is in agreement with earlier reports showing that different aspects of audiovisual processing may rely on disparate brain regions (e.g., Callan, Jones, & Callan, 2014; Calvert, 2001; Erickson et al., 2014), with distinct areas of the premotor cortex activated by audiovisual and visual only speech perception (Callan et al., 2014)

As predicted, we also found that the LPC amplitude was significantly larger to congruent as compared to incongruent articulations, indicative of word recognition during congruent trials. In studies in which words are presented in their written form, the latency of this component is typically measured between 500 and 900 ms post-stimulus onset (e.g., Friedman & Johnson, 2000; Neville et al., 1986; Paller & Kutas, 1992). In our results, the LPC difference between congruent and incongruent articulations was markedly prolonged, suggesting that most of the articulation video had to be processed before recognition took place. This finding may reflect the fact that our stimuli unfolded over time (which is in contrast to written word presentations of earlier studies). But it may also reflect the fact that recognizing an articulation is more difficult than recognizing a written word, in part because the same mouth shape and/or mouth movement may be associated with multiple speech sounds (Tye-Murray, Sommers, & Spehar, 2007).

While the finding of the overall larger amplitude of the LPC component to congruent articulations agreed with our original hypotheses, its relationship to the SIN performance did not. Neither the LPC difference between congruent and incongruent articulations nor its mean amplitude to congruent articulations was related to the SIN accuracy increase in the AV condition. As mentioned above, the LPC component occurred very late in our data, suggesting that a full recognition of visual articulations as being congruent with the previously heard word does not happen until after most of the articulation video has been watched, at which processing point visual speech cues may no longer be of much relevance for improvement in the SIN perception.

The mean amplitude of the LPC component to congruent articulations was, however, related to the accuracy on the lip-reading task. More specifically, individuals with overall smaller LPC were better lip-readers. This finding appears counterintuitive on first read. However, one possible explanation is that the LPC component in our data reflects two different variables that influence its amplitude in opposite ways – namely, congruency and repetition. Support for this interpretation comes in part from work by Olichney and colleagues (Olichney et al., 2006; Olichney et al., 2013) who used a semantic category matching task. In this task, participants were first given a category, such as “a breakfast food.” They then had to decide whether subsequently presented words matched (were congruent with) or did not match (were incongruent with) the given category. Approximately half of the used words were repeated, while others were not. The authors reported that congruent words elicited a

significantly larger LPC compared to incongruent words. However, when the same congruent words were presented multiple times, they elicited significantly reduced LPCs.

Olichney's findings of reduced LPC to repeated word presentations is of direct relevance to our paradigm because congruent articulations might have been perceived by our participants as "repeated" articulations if they formed any expectation of what the articulation should look like following the auditory word presentation. It then stands to reason that individuals with smaller LPCs to congruent articulations did better on the lip-reading component of the Matching task since both reflect the ability to identify/recognize words based on observed articulation. Within the current paradigm, comparing ERPs elicited by the first and the repeated presentations of the same articulation was not possible because each articulation was presented only once. However, the suggested interpretation of our LPC results lends itself to a testable prediction that individuals with smaller LPCs to congruent articulations in the current study would also show the largest reduction in this component if the same visual articulation is repeated multiple times.

As with all studies, our paradigm has its limitations. Our task required participants' explicit judgment about whether or not visual articulations matched preceding words. It remains to be seen whether similar results could be obtained in a task that examines matching facial articulatory movement to auditory words in a more implicit manner. A number of neuroimaging studies suggest that simply hearing speech automatically activates the listener's motor speech representations (Möttönen, Dutton, & Watkins, 2013; Pulvermüller et al., 2006; Wilson, Saygin, Sereno, & Iacoboni, 2004). However, whether representations for how words look when articulated may also be automatically activated is yet to be investigated. Additionally, in our design, congruent and incongruent articulations differed during word onset in order to obtain the largest temporal separation between the ERP component associated with detecting the articulatory mismatch (N400) and the ERP component associated with word recognition (LPC). An interesting question to be addressed during future studies is whether sensitivity to auditory-articulatory mismatches at a later portion of a word's articulation would still be predictive of individuals' SIN improvement in the presence of the talker's face or whether this effect is limited to word onsets. Lastly, in most repetition priming paradigms, behavioral measures of reaction time and accuracy reveal greater ease of processing repeated as compared to non-repeated information. However, we did not see either the shortening of the reaction time or the increase of accuracy to congruent articulations. The likely reason for this is that the behavioral response was intentionally delayed until after the target articulation was completed to avoid the contamination of EEG recordings with the motor response-related ERPs. Such delay must have been sufficient to mask any congruency effects we could have obtained if participants were instructed to respond as soon as possible after seeing the articulation.

In sum, we have described a novel paradigm that examines two temporally distinct stages in matching facial articulatory gestures with auditory speech - namely, detecting audiovisual correspondences at the pre-lexical level and articulatory word recognition. We showed that it is the earlier process - the detection of audiovisual correspondences at the phonemic and/or syllabic level - that is correlated with the SIN perception. The ability to match visual articulatory gestures with auditory speech information is one of the key skills necessary for

successful audiovisual speech perception. This study may, therefore, serve as a baseline for evaluating the processing of facial articulatory movements in various populations for whom atypical audiovisual speech perception has been reported, such as autism (Foxe et al., 2013; Guiraud et al., 2012; Saalasti et al., 2012; Stevenson et al., 2014; Taylor, Isaac, & Milne, 2010), dyslexia (Bastien-Toniazzo, Stroumza, & Cavé, 2010), specific language impairment (Boliek, Keintz, Norrix, & Obrzut, 2010; Hayes, Tiippana, Nicol, Sams, & Kraus, 2003; Kaganovich, Schumaker, Macias, & Anderson, in press; Leybaert et al., 2014; Meronen, Tiippana, Westerholm, & Ahonen, 2013; Norrix, Plante, & Vance, 2006; Norrix, Plante, Vance, & Boliek, 2007), and phonological disorders (Dodd, McIntosh, Erdener, & Burnham, 2008).

## Acknowledgments

This research was supported in part by the R03DC013151 grant from the National Institute on Deafness and Other Communicative Disorders, National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Institute on Deafness and Other Communicative Disorders or the National Institutes of Health. We are grateful to Kevin Barlow for creating stimulus presentation programs, to Steven Hnath and Samantha Hoover for help with video materials, and to James Hengenius for assistance with statistical analyses. We are also thankful to Michele Rund, Rachel Buckser, Jessica Huemmer, Olivia Niemiec, and Kelly Sievert for help with various stages of this project.

## References

Alho J, Lin F-H, Sato M, Tiitinen H, Sams M, Jääskeläinen IP. Enhanced neural synchrony between left auditory and premotor cortex is associated with successful phonetic categorization. Frontiers in Psychology. 2014; 5

Altieri N, Hudock D. Hearing impairment and audiovisual speech integration ability: a case study report. Frontiers in Psychology. 2014; 5

American Electroencephalographic Society. Guideline thirteen: Guidelines for standard electrode placement nomenclature. Journal of Clinical Neurophysiology. 1994; 11:111–113. [PubMed: 8195414]

Barutchu A, Danaher J, Crewther SG, Innes-Brown H, Shivdasani MN, Paolini AG. Audiovisual integration in noise by children and adults. Journal of Experimental Child Psychology. 2010; 105:38–50. [PubMed: 19822327]

Bastien-Toniazzo M, Stroumza A, Cavé C. Audio-visual perception and integration in developmental dyslexia: An exploratory study using the McGurk effect. Current Psychology Letters: Behaviour, Brain and Cognition. 2010; 25(3):1–15.

BioSemi. Active Electrodes. 2013. Retrieved from http://www.biosemi.com/active_electrode.htm

Boersma, P.; Weenink, D. Praat: doing phonetics by computer (version 5.3) [Computer program]. 2011. Retrieved from http://www.praat.org (Version 5.1)

Boliek CA, Keintz C, Norrix LW, Obrzut J. Auditory-visual perception of speech in children with leaning disabilities: The McGurk effect. Canadian Journal of Speech-Language Pathology and Audiology. 2010; 34(2):124–131.

Brown, L.; Sherbenou, RJ.; Johnsen, SK. Test of Nonverbal Intelligence. 4th. Austin, Texas: Pro-Ed: An International Pubilsher; 2010.

Brysbaert M, New B. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods. 2009; 41(4):977–990. [PubMed: 19897807]

Callan DE, Jones JA, Callan A. Multisensory and modality specific processing of visual speech in different regions of the prefrontal cortex. Frontiers in Psychology. 2014; 5

Calvert GA. Crossmodal processing in the human brain: Insights from functional neuroimaging studies. Cerebral Cortex. 2001; 11:1110–1123. [PubMed: 11709482]

Cohen, MS. Handedness Questionnaire. 2008. Retrieved from http://www.brainmapping.org/shared/Edinburgh.php#

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

Dodd B, McIntosh B, Erdener D, Burnham D. Perception of the auditory-visual illusion in speech perception by children with phonological disorders. Clinical Linguistics and Phonetics. 2008; 22(1):69–82. [PubMed: 18092221]

Duncan CC, Barry RJ, Connolly JF, Fischer C, Michie PT, Näätänen R, Van Petten C. Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clinical Neurophysiology. 2009; 120:1883–1908. [PubMed: 19796989]

Dunn, LM.; Dunn, DM. Peabody Picture Vocabulary Test. 4th. Pearson; 2007.

Erickson LC, Zielinski BA, Zielinski JEV, Liu G, Turkeltaub PE, Leaver AM, Rauschecker JP. Distinct cortical locations for integration of audiovisual speech and the McGurk effect. Frontiers in Psychology. 2014; 5

Fenson, L.; Marchman, V.; Thal, DJ.; Dale, PS.; Reznick, JS.; Bates, E. MacArthur-Bates Communicative Development Inventories (CDI) Words and Sentences. Brookes Publishing Co.; 2007.

Field, A. Discovering statistics using SPSS. 3. Washington, DC: Sage; 2009.

Foxe JJ, Molholm S, Del Bene VA, Frey H-P, Russo NN, Blanco D, Ross LA. Severe multisensory speech integration deficits in high-functioning school-aged children with Autism Spectrum Disorder (ASD) and their resolution during adolescence. Cerebral Cortex. 2013

Friedman D, Johnson R Jr. Event-related potential (ERP) studies of memory encoding and retrieval: A selective review. Microscopy Research and Technique. 2000; 51:6–28. [PubMed: 11002349]

Grant KW, Seitz AR. Measures of auditory-visual integration in nonsense syllables and sentences. Journal of the Acoustical Society of America. 1998; 104(4):2438–2450. [PubMed: 10491705]

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

Grant KW, Walden BE, Seitz P. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. Journal of the Acoustical Society of America. 1998; 103(5):2677–2690. [PubMed: 9604361]

Groppe DM, Urbach TP, Kutas M. Mass univariate analysis of event related brain potentials/fields I: A critical tutorial review. Psychophysiology. 2011; 48:1711–1725. [PubMed: 21895683]

Guiraud JA, Tomalski P, Kushnerenko E, Ribeiro H, Davies K, Charman T, Team tB. Atypical audiovisual speech integration in infants at risk for autism. PLOS ONE. 2012; 7(5):e36428. [PubMed: 22615768]

Hayes EA, Tiippana K, Nicol TG, Sams M, Kraus N. Integration of heard and seen speech: a factor in learning disabilities in children. Neuroscience Letters. 2003; 351:46–50. [PubMed: 14550910]

Hickok G, Okada K, Serences JT. Area Spt in the human planum temporale supports sensory-motor integration for speech processing. Journal of Neurophysiology. 2009; 101:2725–2732. [PubMed: 19225172]

Holcomb PJ, Anderson J, Grainger J. An electrophysiological study of cross-modal repetition priming. Psychophysiology. 2005; 42:493–507. [PubMed: 16176372]

Kaganovich N, Schumaker J, Macias D, Anderson D. Processing of audiovisually congruent and incongruent speech in school-age children with a history of Specific Language Impairment: a behavioral and event-related potentials study. Developmental Science. in press.

Kutas M, Federmeier KD. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review In Psychology. 2011; 62:621–647.

Kutas M, Van Petten C. Event-related brain potential studies of language. Advances in Psychophysiology. 1988; 3:139–187.

Kutas, M.; Van Petten, C. Psycholinguistics electrified: Event-related brain potential investigations. In: Gernsbacher, MA., editor. Handbook of Psycholinguistics. San Diego, CA: Academic Press, Inc.; 1994. p. 83-143.

Leybaert J, Macchi L, Huyse A, Champoux F, Bayard C, Colin C, Berthommier F. Atypical audio-visual speech perception and McGurk effects in children with specific language impairment. Frontiers in Psychology. 2014; 5

Luck, SJ. An Introduction to the Event-Related Potential Technique. Cambridge, MA: The MIT Press; 2005.

McGrath M, Summerfield Q. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. Journal of the Acoustical Society of America. 1985; 77(2):678–685. [PubMed: 3973239]

Meronen A, Tiippana K, Westerholm J, Ahonen T. Audiovisual speech perception in children with developmental language disorder in degraded listening conditions. Journal of Speech, Language, and Hearing Research. 2013; 56:211–221.

Metting van Rijn, AC.; Kuiper, AP.; Dankers, TE.; Grimbergen, CA. Low-cost active electrode improves the resolution in biopotential recordings; Paper presented at the 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Amsterdam, The Netherlands. 1996.

Metting van Rijn AC, Peper A, Grimbergen CA. High-quality recording of bioelectric events. Part 1: Interference reduction, theory and practice. Medical and Biological Engineering and Computing. 1990; 28:389–397. [PubMed: 2277538]

Möttönen R, Dutton R, Watkins KE. Auditory-motor processing of speech sounds. Cerebral Cortex. 2013; 23(5):1190–1197. [PubMed: 22581846]

Neville HJ, Kutas M, Chesney G, Schmidt AL. Event-related brain potentials during initial encoding and recognition memory of congruous and incongruous words. Journal of Memory and Language. 1986; 25:75–92.

Norrix LW, Plante E, Vance R. Auditory-visual speech integration by adults with and without language-learning disabilities. Journal of Communication Disorders. 2006; 39:22–36. [PubMed: 15950983]

Norrix LW, Plante E, Vance R, Boliek CA. Auditory-visual integration for speech by children with and without specific language impairment. Journal of Speech, Language, and Hearing Research. 2007; 50:1639–1651.

Oldfield RC. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia. 1971; 9:97–113. [PubMed: 5146491]

Olichney JM, Iragui VJ, Salmon DP, Riggins BR, Morris SK, Kutas M. Absent event-related potential (ERP) word repetition effects in mild Alzheimer's disease. Clinical Neurophysiology. 2006; 117:1319–1330. [PubMed: 16644278]

Olichney JM, Pak J, Salmon DP, Yang J-C, Gahagan T, Nowacki R, Iragui-Madoz VJ. Abnormal P600 word repetition effect in elderly persons with preclinical Alzheimer's disease. Cognitive Neuroscience. 2013; 4(3–4):143–151. [PubMed: 24090465]

Paller KA, Kutas M. Brain potentials during memory retrieval provide neurophysiological support for the distinction between conscious recollection and priming. Journal of Cognitive Neuroscience. 1992; 4(4):375–392. [PubMed: 23968130]

Pflieger, ME. Theory of a spatial filter for removing ocular artifacts with preservation of EEG; Paper presented at the EMSE Workshop; Princeton University; 2001. http://www.sourcesignal.com/SpFilt_Ocular_Artifact.pdf

Praamstra P, Meyer AS, Levelt WJM. Neurophysiological manifestations of phonological processing: Latency variation of a negative EP component timelocked to phonological mismatch. Journal of Cognitive Neuroscience. 1994; 6(3):204–219. [PubMed: 23964972]

Praamstra P, Stegeman DF. Phonological effects on the auditory N400 event-related brain potential. Cognitive Brain Research. 1993; 1:73–86. [PubMed: 8513242]

Pulvermüller F, Huss M, Kherif F, Moscoso del Prado Martin F, Hauk O, Shtyrov Y. Motor cortex maps articulatory features of speech sounds. Proceedings of the National Academy of Sciences. 2006; 103(20):7865–7870.

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

Rugg MD, Curran T. Event-related potentials and recognition memory. Trends in Cognitive Sciences. 2007; 11(6):251–257. [PubMed: 17481940]

Saalasti S, Kätsyri J, Tiippana K, Laine-Hernandez M, von Wendt L, Sams M. Audiovisual speech perception and eye gaze behavior of adults with Asperger syndrome. Journal of Autism and Developmental Disorders. 2012; 42:1606–1615. [PubMed: 22068821]

Schwartz J-L, 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:e1003743. [PubMed: 25079216]

Stevenson RA, Nelms CE, Baum SH, Zurkovsky L, Barense MD, Newhouse PA, Wallace MT. Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition. Neurobiology of Aging. 2015; 36:283–291. [PubMed: 25282337]

Stevenson RA, Siemann JK, Schneider BC, Eberly HE, Woynaroski TG, Camarata SM, Wallace MT. Multisensory temporal integration in autism spectrum disorders. The Journal of Neuroscience. 2014; 34(3):691–697. [PubMed: 24431427]

The SUBTL Word Frequency. 2009

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

Taylor N, Isaac C, Milne E. A comparison of the development of audiovisual integration in children with Autism Spectrum Disorders and typically developing children. Journal of Autism and Developmental Disorders. 2010; 40:1403–1411. [PubMed: 20354776]

Oever S, Schroeder CE, Poeppel D, van Atteveldt N, Zion-Golumbic E. Rhythmicity and cross-modal temproal cues facilitate detection. Neuropsychologia. 2014; 63:43–50. [PubMed: 25128589]

Tye-Murray N, Sommers MS, Spehar B. Auditory and visual lexical neighborhoods in audiovisual speech perception. Trends in Amplification. 2007; 11(4):233–241. [PubMed: 18003867]

Tye-Murray N, Spehar B, Myerson J, Sommers MS, Hale S. Cross-modal enhancement of speech detection in young and older adults: Does signal content matter? Ear and Hearing. 2011; 32(5): 650–655. [PubMed: 21478751]

Van Petten C, Coulso S, Rubin S, Plante E, Parks M. Time course of word identification and semantic integration in spoken language. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1999; 25(2):394–417.

Van Petten C, Coulson S, Rubin S, Plante E, Parks M. Time course of word identification and semantic integration in spoken language. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1999; 25(2):394–417.

van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. Neuropsychologia. 2007; 45:598–607. [PubMed: 16530232]

Wilson SM, Saygin AP, Sereno MI, Iacoboni M. Listening to speech activates motor areas involved in speech production. Nature Neuroscience. 2004; 7:701–702. [PubMed: 15184903]

Yi A, Wong W, Eizenman M. Gaze patterns and audiovisual speech enhancement. Journal of Speech, Language, and Hearing Research. 2013; 56:471–480.

## Appendix. The pairing of auditory words and silent visual articulations

Note that articulations that are congruent (i.e., match the preceding auditory word) in List A are incongruent (i.e., do not match the preceding auditory word) in List B.

| | List A | | List B | |
|---|---|---|---|---|
| | **Auditory Word** | **Silent Visual Articulation** | **Auditory Word** | **Silent Visual Articulation** |
| 1 | shower | shower | candy | donkey |
| 2 | tree | lamb | cat | cat |
| 3 | jello | jello | jello | monkey |
| 4 | cat | girl | egg | egg |
| 5 | egg | pool | donut | bottle |
| 6 | donut | donut | zipper | present |
| 7 | zipper | zipper | donkey | candy |
| 8 | donkey | donkey | shirt | shirt |
| 9 | grapes | grapes | grapes | farm |
| 10 | police | apple | police | police |
| 11 | truck | belt | apple | apple |
| 12 | apple | police | truck | truck |
| 13 | monkey | monkey | monkey | jello |
| 14 | sandwich | mailman | sandwich | sandwich |
| 15 | car | car | car | fish |
| 16 | turtle | turtle | turtle | popcorn |
| 17 | squirrel | squirrel | squirrel | pretzel |
| 18 | window | window | window | sandbox |
| 19 | sled | bird | sled | sled |
| 20 | necklace | necklace | bread | duck |
| 21 | water | water | water | carrot |
| 22 | sink | sink | sink | mop |
| 23 | paint | paint | paint | woods |
| 24 | pretzel | pretzel | pretzel | squirrel |
| 25 | nail | peas | nail | nail |

| | List A | | List B | |
|---|---|---|---|---|
| | **Auditory Word** | **Silent Visual Articulation** | **Auditory Word** | **Silent Visual Articulation** |
| 26 | bird | sled | bird | bird |
| 27 | corn | corn | corn | frog |
| 28 | couch | couch | couch | moose |
| 29 | farm | farm | farm | grapes |
| 30 | airplane | pumpkin | airplane | airplane |
| 31 | popcorn | popcorn | popcorn | turtle |
| 32 | penguin | doctor | penguin | penguin |
| 33 | knife | mouth | mouth | mouth |
| 34 | arm | horse | arm | arm |
| 35 | bed | ear | bed | bed |
| 36 | present | present | present | zipper |
| 37 | sandbox | sandbox | sandbox | window |
| 38 | mop | mop | mop | sink |
| 39 | mailman | sandwich | mailman | mailman |
| 40 | lamb | tree | shower | necklace |
| 41 | candy | candy | duck | bread |
| 42 | scissors | balloon | scissors | scissors |
| 43 | pool | egg | pool | pool |
| 44 | bee | bee | bee | eye |
| 45 | chair | boat | chair | chair |
| 46 | cake | ball | cake | cake |
| 47 | boy | boy | boy | dog |
| 48 | sprinkler | muffin | sprinkler | sprinkler |
| 49 | elephant | elephant | elephant | teddy bear |
| 50 | comb | beach | comb | comb |
| 51 | jar | purse | jar | jar |

| | List A | | List B | |
|---|---|---|---|---|
| | **Auditory Word** | **Silent Visual Articulation** | **Auditory Word** | **Silent Visual Articulation** |
| 52 | horse | arm | horse | horse |
| 53 | sweater | sweater | sweater | picture |
| 54 | moose | moose | moose | couch |
| 55 | muffin | sprinkler | muffin | muffin |
| 56 | ear | bed | ear | ear |
| 57 | toys | toys | toys | bus |
| 58 | bus | bus | carrot | water |
| 59 | carrot | carrot | teacher | buttons |
| 60 | teacher | teacher | hammer | hammer |
| 61 | hammer | pizza | bus | toys |
| 62 | frog | frog | frog | corn |
| 63 | shirt | foot | necklace | shower |
| 64 | buttons | buttons | buttons | teacher |
| 65 | ball | cake | ball | ball |
| 66 | beach | comb | beach | beach |
| 67 | girl | cat | girl | girl |
| 68 | mouth | knife | knife | knife |
| 69 | peas | nail | peas | peas |
| 70 | woods | woods | woods | paint |
| 71 | picture | picture | picture | sweater |
| 72 | purse | jar | purse | purse |
| 73 | belt | truck | belt | belt |
| 74 | wolf | wolf | wolf | house |
| 75 | scarf | scarf | scarf | broom |
| 76 | teddy bear | teddy bear | teddy bear | elephant |
| 77 | house | house | house | wolf |

| | List A | | List B | |
|---|---|---|---|---|
| | **Auditory Word** | **Silent Visual Articulation** | **Auditory Word** | **Silent Visual Articulation** |
| 78 | eye | eye | eye | bee |
| 79 | dog | dog | dog | boy |
| 80 | flower | orange | flower | flower |
| 81 | doctor | penguin | doctor | doctor |
| 82 | foot | shirt | foot | foot |
| 83 | broom | broom | broom | scarf |
| 84 | tractor | pencil | tractor | tractor |
| 85 | circus | money | circus | circus |
| 86 | balloon | scissors | balloon | balloon |
| 87 | orange | flower | orange | orange |
| 88 | pencil | tractor | pencil | pencil |
| 89 | pumpkin | airplane | pumpkin | pumpkin |
| 90 | bread | bread | lamb | lamb |
| 91 | money | circus | money | money |
| 92 | bottle | bottle | bottle | donut |
| 93 | boat | chair | boat | boat |
| 94 | pizza | hammer | pizza | pizza |
| 95 | fish | fish | fish | car |
| 96 | duck | duck | tree | tree |

## Highlights

- Matching silent articulations with heard words elicits N400 and LPC ERP components
- N400 is larger to incongruent articulations and reflects pre-lexical matching
- LPC is larger to congruent articulations and indexes articulatory word recognition
- Only N400 amplitude is predictive of SIN improvement in the AV condition

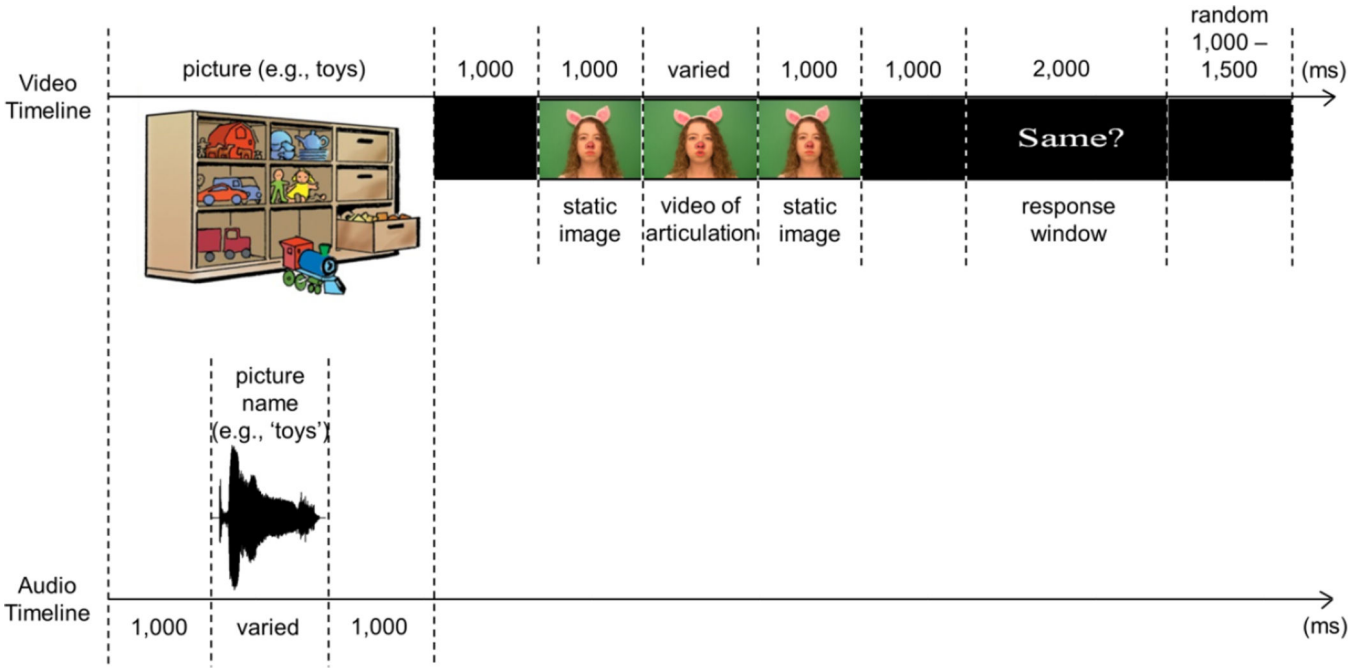


**Figure 1. Schematic representation of a trial in the Matching task**

Note that separate timelines are shown for the video and audio tracks. The video of articulation was congruent in half of all trials (e.g., participants saw the piglet silently articulate "toys" after hearing "toys" at the start of the trial) and incongruent in the other half of trials (e.g., participants saw the piglet silently articulate "bus" after hearing "toys" at the start of the trial). The onset of articulation was used as time 0 for ERP averaging.

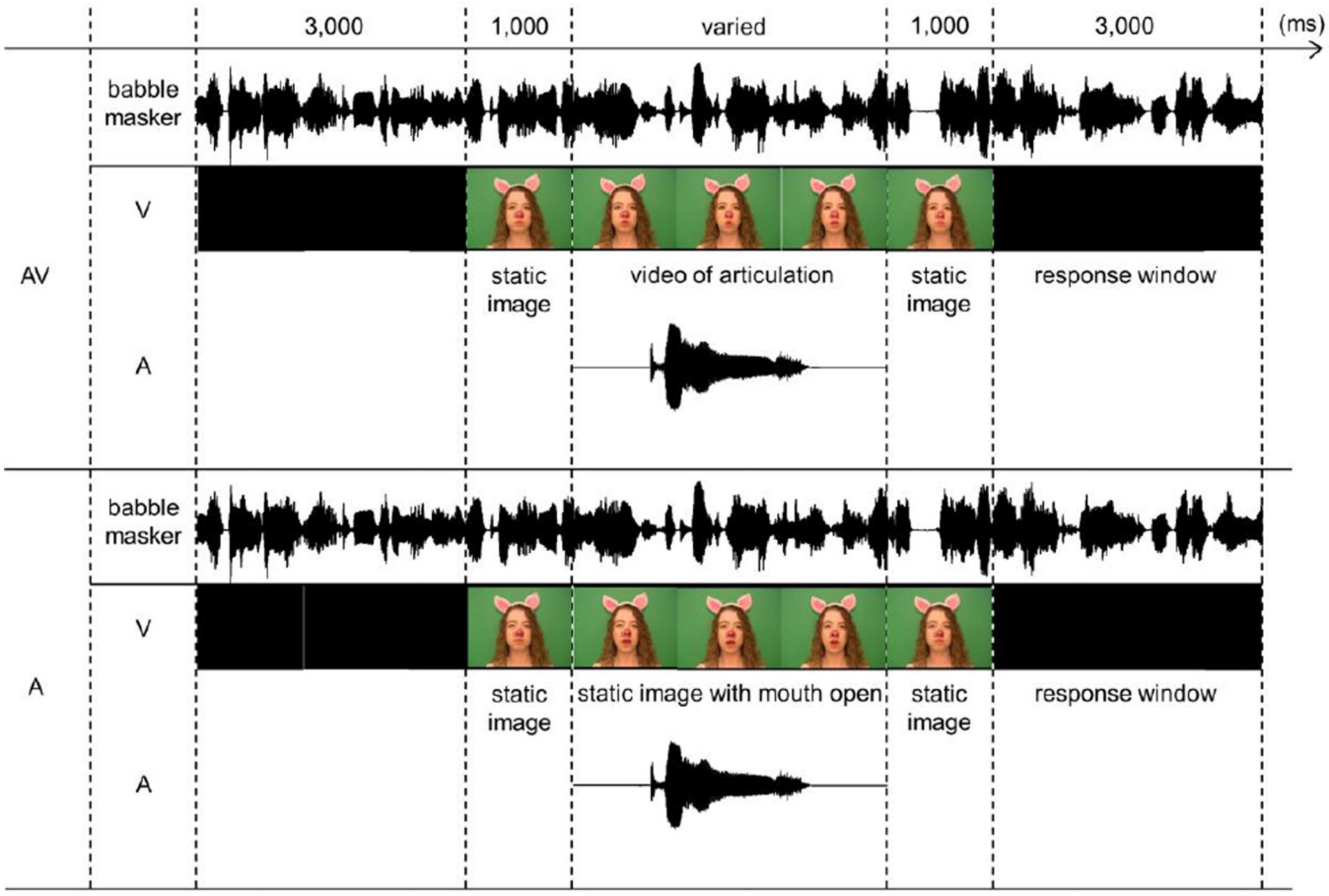


**Figure 2. Schematic representation of a trial in the Speech-In-Noise (SIN) task**

The SIN task had two conditions – the audiovisual (AV, top panel) and the auditory only (A, bottom panel). Note that separate timelines are shown for the video and audio tracks in each condition. In the AV condition, participants saw a video of the piglet articulating target words, while in the A condition the video portion was replaced with a static image of the piglet's face with her mouth open. The appearance of the open mouth picture in the A condition cued participants to the fact that the onset of the auditory word was imminent, but provided no visual cues to its identity.

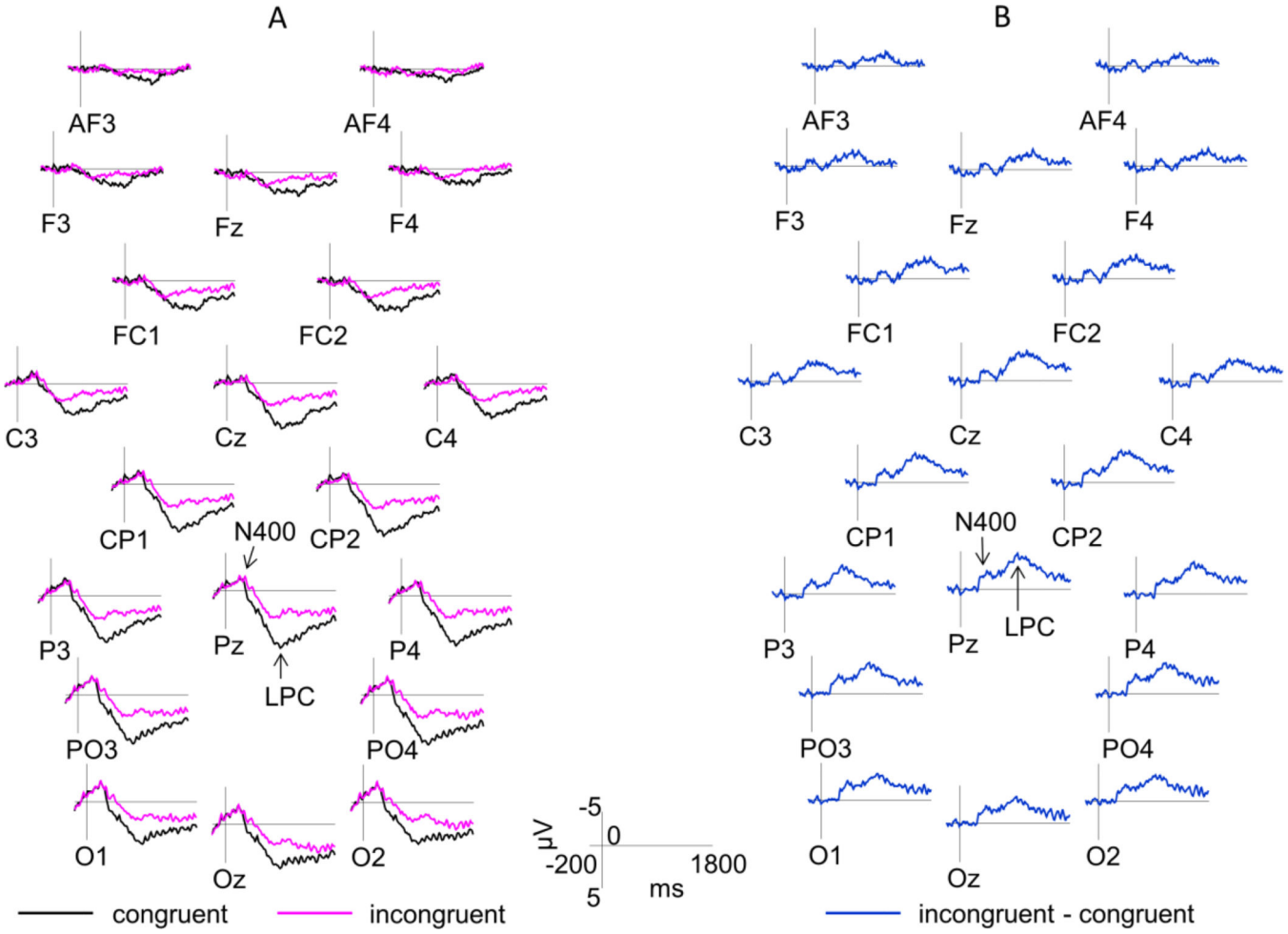


**Figure 3. ERP Results**

Panel A shows an overlap between grand average waveforms elicited by congruent and incongruent silent articulations. Panel B shows a difference waveform produced by subtracting the ERPs elicited by incongruent articulations from the ERPs elicited by congruent articulations. Because the LPC mean amplitude was more positive to congruent articulations, it appears as a negative (rather than a positive) deflection in the difference waveform. The N400 and LPC components are marked on the Pz site. Only the midline and the mid-lateral sites are shown. Negative is plotted up. Time 0 indexes the onset of articulatory movements.

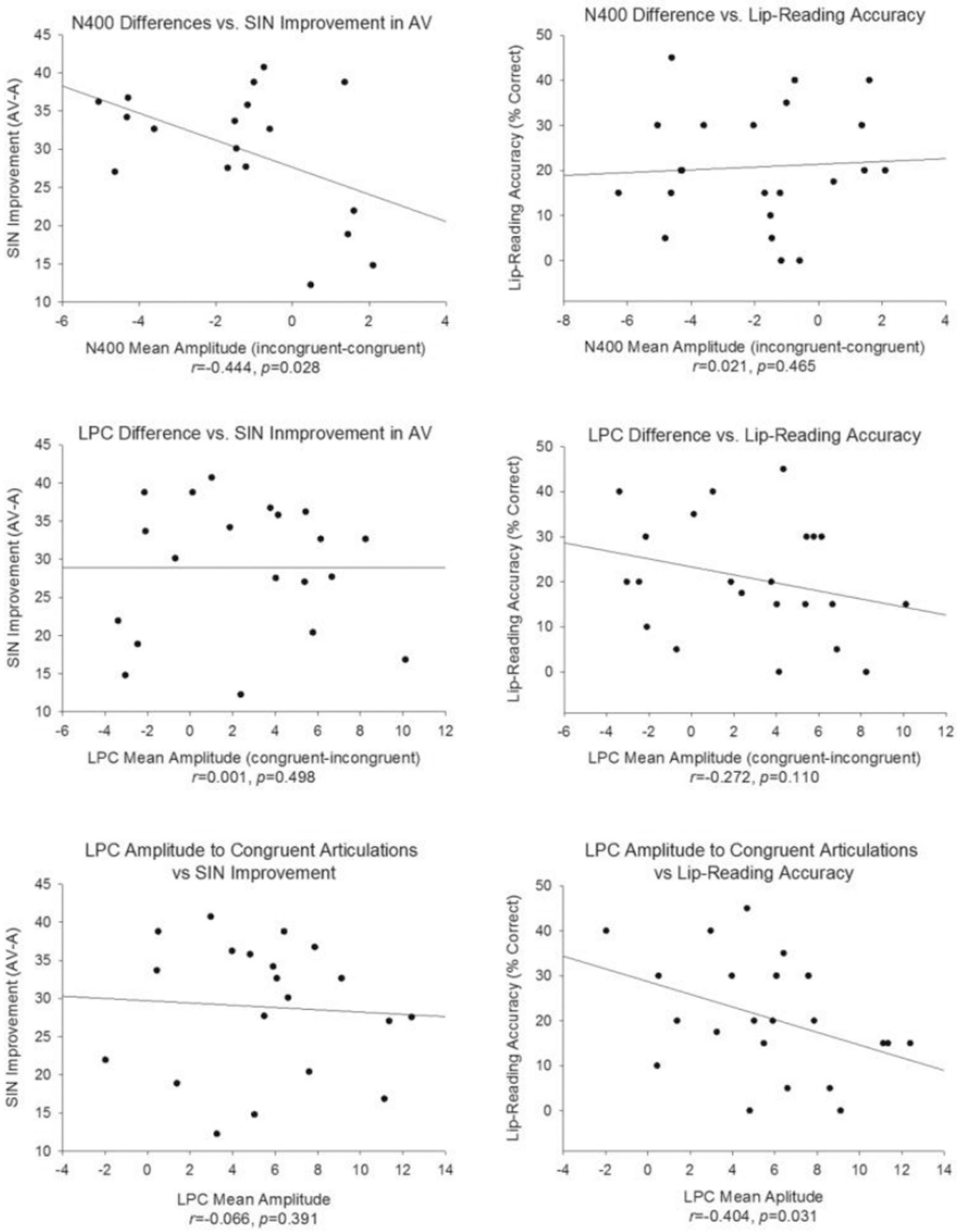


**Figure 4. Correlations**

Relationships between ERP measures and SIN are shown on the left, and relationships between ERP measures and lip-reading skills are shown on the right. Note that the N400 difference measure (displayed in the top row) was calculated by subtracting the N400 mean amplitude to congruent articulations from the N400 mean amplitude to incongruent articulations. The LPC difference measure (displayed in the second row), on the other hand, was calculated by subtracting the LPC mean amplitude to incongruent articulations from the LPC mean amplitude to congruent articulations. The SIN improvement measure is the

percent correct improvement in the AV as compared to the A condition. Lastly, the lip-reading accuracy measure is participants' performance on the lip-reading component of the Matching task and is the percent of correctly identified words that were presented in the visual only modality.

**Table 1**

Words presented in the lip-reading component of the Matching task

| Bilabial/labiodental onset | | Alveolar onset | |
|---|---|---|---|
| **one-syllable words** | **two-syllable words** | **one-syllable words** | **two-syllable words** |
| boy | pumpkin | dog | necklace |
| mop | mailman | tree | donkey |
| farm | window | lamb | sweater |
| beach | balloon | knife | zipper |
| woods | flower | scarf | teacher |