Published in final edited form as:

Multisens Res. 2020 February 28; 33(3): 277–294. doi:10.1163/22134808-20191423.

Visual Enhancement of Relevant Speech in a ‘Cocktail Party’

Niti Jaha[1], **Stanley Shen**[1], **Jess R. Kerlin**[1], **Antoine J. Shahin**[1,2,*]

[1]Center for Mind and Brain, University of California, Davis, 95618, USA

[2]Department of Cognitive and Information Sciences, University of California, Merced, CA 95343, USA

Abstract

Lip-reading improves intelligibility in noisy acoustical environments. We hypothesized that watching mouth movements benefits speech comprehension in a ‘cocktail party’ by strengthening the encoding of the neural representations of the visually paired speech stream. In an audiovisual (AV) task, EEG was recorded as participants watched and listened to videos of a speaker uttering a sentence while also hearing a concurrent sentence by a speaker of the opposite gender. A key manipulation was that each audio sentence had a 200-ms segment replaced by white noise. To assess comprehension, subjects were tasked with transcribing the *AV-attended* sentence on randomly selected trials. In the auditory-only trials, subjects listened to the same sentences and completed the same task while watching a static picture of a speaker of either gender. Subjects directed their listening to the voice of the gender of the speaker in the video. We found that the N1 auditory-evoked potential (AEP) time-locked to white noise onsets was significantly more inhibited for the *AV-attended* sentences than for those of the auditorily-attended (*A-attended*) and *AV-unattended* sentences. N1 inhibition to noise onsets has been shown to index restoration of phonemic representations of degraded speech. These results underscore that attention and congruency in the AV setting help streamline the complex auditory scene, partly by reinforcing the neural representations of the visually attended stream, heightening the perception of continuity and comprehension.

Keywords

Audiovisual integration; auditory-evoked potentials; ‘cocktail party’; phonemic restoration

1. Introduction

Conversing in a noisy background, such as in a ‘cocktail party’ (Cherry, 1953), is one of the most common everyday situations. In these situations, listeners constantly adjust their attention and perceptual strategies, taking advantage of the unfolding sensory cues, to

[*]To whom correspondence should be addressed. ashahin@ucmerced.edu.

Conflict of Interest
The authors declare no conflicts of interest, financial or otherwise.

Data Accessibility
The original behavioral and EEG data can be found at: https://doi.org/10.6084/m9.figshare.c.4510976.v1.

segregate the relevant speech from the interfering speech. This process is known as auditory stream segregation (Bregman, 1990). Some of these cues include the speaker's fundamental frequency (F0), or its perceptual analog, the pitch (Alho *et al.*, 1987; Grimault *et al.*, 2000; Oxenham, 2008; Woods *et al.*, 2001), the spatial origins of the sound sources (Ihlefeld and Shinn-Cunningham, 2008), the speech's amplitude variation, or speech envelope (Drullman *et al.*, 1994; Zeng *et al.*, 1999), and the speaker's mouth movements (Grant and Seitz, 2000; Shahin and Miller, 2009; Sumby and Pollack, 1954). The behavioral advantages gained from these cues are well-established (see references above), but less is known about the neurophysiological underpinning of sensory cue utilization in cluttered acoustical environments. The focus of this study is on the neurophysiological influence of visual cues in discriminating relevant from irrelevant speech representations in a 'cocktail party'.

One mechanism thought to underlie enhanced comprehension in a 'cocktail party' is *via* reinforcing the auditory cortex's ability to track the envelope of the attended speech. In a discrimination task of two concurrent speech sentences with spatially distinct sources, Kerlin *et al.* (2010) revealed that theta band neural activity (4–7 Hz) was greater for the attended than the ignored concurrent speech stream. Because the theta frequency band also reflects the amplitude variations of the speech envelope, they concluded that auditory spatial-selective attention heightens the tracking of the attended speech envelope in the auditory cortex. In support, Zion Golumbic *et al.* (2013) examined the phase-tracking of the auditory cortex to attended and ignored speech streams in a 'cocktail party' scenario and found that the phase of low-frequency delta and theta activity (1–7 Hz) of auditory neurons tracked the attended speech stream with a greater fidelity than the ignored one. A similar mechanism has been reported during visual enhancement of speech in a 'cocktail party' — visual strengthening of the auditory cortex's tracking of the speech envelope (Crosse *et al.*, 2015; O'Sullivan *et al.*, 2013; Zion Golumbic *et al.*, 2013) of the audiovisually attended (*AV-attended*) sentence relative to the *AV-unattended* sentence. This effect is less robust when visual cues are absent (auditory-only task).

The purpose of this study was to further understand the AV neurophysiology mediating comprehension in a 'cocktail party'. EEG was acquired as participants watched videos of a human speaker while they listened to two sentences, one of which was congruent with the mouth movements of the speaker in the video. The two acoustic sentences were played from one loudspeaker of a sound bar and were always spoken by people of opposite genders. We also incorporated an auditory-only task, in which subjects listened to the same sentences as in the AV task while watching a static picture of a speaker of either gender. Subjects attended to the sentences containing the voice of the gender of the speaker in the picture. A key manipulation was that each acoustic sentence had a 200-ms segment replaced by white noise. The purpose of the noise-replaced segment was to demonstrate that visually-mediated speech comprehension enhancement is also attributable to filling-in of missing phonetic representations, e.g., formant dynamics, and is not restricted to strengthening the encoding of the speech envelope (Crosse *et al.*, 2015; O'Sullivan *et al.*, 2013; Zion Golumbic *et al.*, 2013). This filling-in process follows from the classical phonemic restoration or continuity illusion design (Samuel, 1981; Warren, 1970). In phonemic restoration, speech with a noise-replaced segment is often perceived continuous through the noise. Neurophysiologically, this process has been shown to be facilitated by the suppression of the auditory cortex's response to onsets and offsets of the noise segments (Riecke *et al.*, 2009; Shahin *et al.*, 2012). The reasoning is that filling-in of missing representations heightens the perception of continuity, thus reducing the perception of interruption, resulting in a suppressed auditory response to interruptions. The suppressed auditory response includes the N1 AEP (Shahin *et al.*, 2012) and theta activity (4–7 Hz) (Riecke *et al.*, 2009; Shahin *et al.*, 2012). This continuity perception is further amplified by visual context (Bhat *et al.*, 2014).

Based on the above scientific premise, we hypothesized that if visual context (congruency and visual attention) supports the auditory cortex’s tracking of the speech envelope and filling-in of phonetic representations of the congruent speech, then we should expect greater N1 suppression to noise onsets for the *AV-attended* sentences *versus* the *AV-unattended* sentences, with a smaller or non-significant effect observed in the auditory-only task (*A-attended versus A-unattended*). Our findings indeed show this effect, providing tangible evidence that lip-reading reinforces the encoding of the relevant speech stream’s neural representations at the auditory cortex, and in turn enhances comprehension in a ‘cocktail party’.

2. Materials and Methods

2.1. Subjects

Nineteen adult subjects (2 male, 17 female, 17 right-handed) participated in this study. They had a mean ± SD age of 23.21 ± 4.92 years, with one subject’s age not reported. Subjects self-reported being native English speakers and having normal hearing, normal or corrected vision, and no history of language deficits or neurological disorders. Being a native English speaker in this study entailed having experience with English before age 5. All subjects provided written informed consent in accordance with the guidelines of the University of California, Davis Institutional Review Board, and they were monetarily compensated for their participation.

2.2. Stimuli

The visual and acoustic stimuli were extracted from the TCD-TIMIT database of English (Harte and Gillen, 2015) AV sentences spoken by male and female speakers with Irish accents. We limited our stimuli to one male speaker and one female speaker. We selected 29 video clips of each speaker uttering a sentence, for a total of 58 unique sentences. Each video clip lasted 4 to 6 s and began and ended with a still face (no mouth movements) and silence. For each sentence, early- and late-noise-replaced versions were created. In Adobe Audition, loudness-equalized, randomly sampled segments of white noise replaced 200 ms of the spoken segment beginning either 25% (early-noise-replaced) or 75% (late-noise-replaced) into the sentence time course. That is, the placement of white noise was not fixed to a specific latency, but rather it was relative to the duration of the whole sentence. Early-replaced white noise onset occurred around 0.5–1.25 s after sentence onset, while late replaced white noise onset occurred around 1.5–3.75 s after sentence onset.

Figure 1 shows the stimulus presentation design. Each trial contained two concurrent audio sentences, with an early-noise-replaced sentence paired with a late-noise-replaced sentence spoken by a speaker of the opposite sex. Trials were counterbalanced across the early and late replaced sentences and attended and unattended sentences. This resulted in 116 trials (232 concurrent sentences, since each trial had two sentences) for the auditory-only task and 116 for the AV task, with each sentence being presented four times within each task (attended early-noise-replaced, attended late-noise-replaced, unattended early-noise-replaced, and unattended late-noise-replaced). Acoustic sentences were paired according to similar lengths for simultaneous presentation. To achieve this, the acoustic waveforms were combined in Adobe Audition such that the sound onsets were aligned in each pair. For the AV task, acoustic sentence pairs were played simultaneously with a male or female speaker's video, in which their mouth movements were congruent with one of the acoustic sentences (*AV-attended* condition); the other acoustic sentence made up the *AV-unattended* condition. For the auditory-only task, a static video of the first frame of each AV video was created in Adobe Premiere by compiling the image at a presentation rate of 30 frames/s for the same duration as the original video. These static videos were paired with audio pairs in the same way as in the *AV* task to create the *A-attended* (audio sentence with a speaker-matched static video) and *A-unattended* (the other audio sentence) conditions.

2.3. Procedure

Subjects sat about 85 cm in front of a 24-inch Dell monitor. EEG and behavioral responses were acquired while subjects watched and listened to the AV and auditory-only videos and made judgments on what they heard. EEG was recorded with a 64-channel cap (BioSemi ActiveTwo system [BioSemi, Amsterdam, The Netherlands], 10–20 Ag-AgCl electrode system, with Common Mode Sense and Driven Right Leg passive electrodes serving as grounds, A/D rate 1024 Hz). The stimuli were presented using Presentation Software (version 18.1, Neurobehavioral Systems [NBS], Berkeley, CA, USA). The sound was played through one loudspeaker of a sound bar (model S2920W-C0, Vizio, Irvine, CA, USA) situated below the monitor, at an intensity level of around 70 dBA sound pressure level. We should note that sound intensity varied among human speakers, from word to word, and from early to late portions of sentences. Notwithstanding, in the EEG analysis, data from all stimuli were counterbalanced across conditions, minimizing acoustic factors on the results. To ensure accurate timing for the EEG analyses, the white noise onset triggers were embedded within the wave file metadata. The experiment consisted of eight blocks that lasted just over 5 min each. Each block consisted of 29 trials (29 pairs of sentences) presented in an event-related mixed design and randomized among the auditory-only and AV tasks. The same sentence pairs were presented for each condition — acoustic sentence pairings did not change over the course of the experiment. However, acoustic-to-video sentence pairing and white noise placement did change. Trial duration was approximately between 9.5 and 12 s. Trials began and ended with silence and a still image. The final frame remained on the screen for 5.5–6 s until the end of trial. In the auditory-only task, subjects listened to an audio sentence while watching a static image of the corresponding speaker (*A-attended* condition). They simultaneously heard another sentence by a speaker of the opposite sex (*A-unattended* condition). In the AV task, subjects watched and listened to a sentence with a congruent video (*AV-attended* condition); they simultaneously heard a competing sentence uttered by a speaker of the opposite sex (*AV-unattended* condition). For all trials where the female speaker was displayed on the screen, subjects attended to the female voice and ignored the male speaker's sentence. Likewise, for all trials where the male speaker was displayed on the screen, subjects attended to the male voice and ignored the sentence spoken by the female speaker. Subjects did not have to make a response or transcribe the attended sentence for every trial; however, throughout the experiment, subjects were asked to transcribe the sentence they attended to in the previous trial for a total of 20 randomly selected trials across the AV and auditory-only conditions. Subjects had no knowledge of which trials they would be asked to transcribe, so these 'catch-trials' served to verify that the subjects remained on task, and to reduce contamination of the stimulus trial EEG with movement or motor activity. Subjects had unlimited time to respond on the 'catch-trials', and presentation of stimuli resumed after they transcribed their responses. Responses of 'catch-trials' were typed on a keyboard placed on a foam pad on the subject's lap. Because the 20 'catch-trials' were randomly selected from the 232 total trials, subjects were presented with roughly equal numbers of auditory-only and AV 'catch-trials' sentences (on average 10.3 ± 2.2 auditory-only sentences and 9.5 ± 1.9 AV sentences). There were a few misses in some subjects, which explains why the cumulative group average is not exactly 20 'catch-trials'.

2.4. Data Analysis

2.4.1. Behavior—First, 'catch-trials' were extracted from the NBS Presentation log files generated from the experiment. Responses of the *AV-attended* and *A-attended* trials were graded on two scales, which were averaged to form the composite score 'Transcription Accuracy' for each sentence: (1) Gist: a score of 0 or 1, where a subject earned a 1 for correctly transcribing the 'gist' of the sentence. The 'gist' was defined as preserving the overall semantic meaning of the sentence. This score rewarded subjects for perceiving the correct words at the time of perception, regardless of whether they retained the exact words in their working memory before they were asked to type the response; (2) Correctness: a maximum of 1 point, calculated according to the fraction of words correct over the total number of words in the original sentence. Articles ('a', 'an', and 'the') were excluded from being scored.

2.4.2. Auditory-Evoked Potentials—EEG data were processed using EEGLAB (Delorme and Makeig, 2004), ERPLAB (Lopez-Calderon and Luck, 2014), and an in-house MATLAB script. Each subject's EEG files, containing all blocks, were down-sampled to 512 Hz, merged into one file, and epoched from 100 to 5000 ms around the beginning of the trial. Recall that a trial began-with silence and still frames. Then, the activity within each epoch was baselined to the mean potential of the entire epoch (mean potential was removed) prior to conducting Independent Component Analysis (ICA). ICA was then performed, with bad channels excluded. ICA components consistent with ocular artifacts were rejected (mean 2 per subject). Subsequently, bad channels (maximum of 3 per subject) were interpolated using the spherical interpolation method implemented in EEGLAB. Individual data were then average-referenced and filtered between 0.1 and 30 Hz using a zero-phase (fourth order) bandpass Butterworth filter. Individual EEG data were then re-epoched from −100 ms to 500 ms around noise onsets and baselined to the 100 ms pre-noise stimulus period and linearly detrended. Epochs with amplitude shifts greater than ± 150 mV at any channel were excluded from the data. Finally, trials for each subject were averaged in the time domain to produce separate AEPs for each condition (*AV-attended*, *A-attended*, *AV-unattended*, *and A-unattended*).

Because we collapsed across early-noise-replaced and late-noise-replaced for each condition, each condition contained 116 trials. However, the trial numbers (mean and SD) for each condition following artifact correction were as follows: *AV-attended*, 107 ± 16 trials; *AV-unattended*, 107 ± 14 trials; *A-attended*, 107 ± 15 trials; *A-unattended*, 107 ± 16 trials.

2.5. Statistical Analyses

2.5.1. Behavior—Accuracy of transcribing the *AV-attended* and *A-attended* sentences was assessed using a paired *t*-test of the transcription accuracy for the two conditions.

2.5.2. Auditory-Evoked Potentials—We analyzed the EEG data *via* cluster-based permutation tests (CBPTs) implemented in the FieldTrip toolbox (Maris and Oostenveld, 2007; Oostenveld *et al.*, 2011). Because our hypothesis was limited to the N1 AEP, the CBPT was confined to the 50–200 ms post-white-noise-onset period of the AEP waveforms. Using the FieldTrip functions, for each contrast between two conditions (*AV-attended vs. AV-unattended*; *A-attended vs. A-unattended*; *AV-attended vs. A-attended*; *AV-unattended vs. A-unattended*), we executed the CBPT to determine if, and in which channels, significant N1 amplitude differences occurred. Initially, two-tailed paired-sample *t*-tests were conducted on the amplitude values of samples of two conditions for each channel to assess univariate effects at the sample level. Data samples with *t*-values exceeding an alpha level of 0.05 (two-tailed) were selected for cluster formation, such that neighboring time points and channels with a univariate *p*-value equal to or smaller than 0.05 were grouped together. Clustering of neighboring channels was based on FieldTrip's triangulation method. Cluster-level statistics were calculated as the sum of all the *t*-values within each time-channel cluster. Significance of these cluster-level statistics was assessed *via* a non-parametric null distribution using a Monte Carlo approximation. This was created by repeating the abovementioned steps for each of the 5000 permutations of the data, whereby the data labels of the conditions were randomly shuffled. The maximum of the cluster-level test statistics was logged for each permutation to form the null distribution. Significance was assessed by contrasting the real cluster-level test statistics to the null distribution of maximum cluster-level statistics. Cluster-based differences were considered significant if the cluster's *p*-value was less than 0.025 for each contrast. In the Results section, we report the exact *p*-values.

3. Results

3.1. Behavior

Figure 2 shows the individual transcription accuracy for *AV-attended* and *A-attended* sentences. Seventeen out of 19 subjects were more accurate in transcribing the AV than auditory-only sentences. A paired *t*-test showed that this effect was significant ($t_{(18)} = 3.9$; $p = 0.001$). These results validate previous accounts demonstrating greater speech comprehension with lip-reading (Banks *et al.*, 2015; Grant and Seitz, 2000; Jesse and Janse, 2012; Sumby and Pollack, 1954).

3.2. Auditory-Evoked Potentials

We compared AEPs to noise onsets of the attended and unattended speech sentences of the AV and auditory-only tasks using the cluster-based permutation test (CBPT). We reasoned that an N1 AEP suppression to noise onsets is an indication of a more robust auditory encoding of the unfolding speech. Thus, we posited that the audio sentence supported by visual context (*AV-attended*) should show greater N1 suppression to noise onsets compared to the unattended sentence (*AV-unattended*) or compared to an attended sentence without visual context (*A-attended*). The first comparison between the *AV-attended* and *AV-unattended* conditions should reveal the combined influence of selective attention and visual context. The second comparison between the *AV-attended* and *A-attended* conditions should factor out selective attention and isolate the influence of visual context. Together, the two contrasts would signify that visual context reinforces tracking and filling-in by the auditory cortex of the relevant speech representations.

3.2.1. *Attended* versus *Unattended*—Figure 3(A) shows the AEP waveforms temporally-locked to the noise onset at channel Cz for the attended *versus* unattended conditions for the AV (left panel) and auditory-only (right panel) tasks. The CBPT revealed a fronto-central negative AEP cluster distinguishing the *AV-attended* from the *AV-unattended* AEP in the period of 133–197 ms ($p = 0.014$). The amplitude AEP values in this period were significantly less negative (smaller) for the *AV-attended* than the *AV-unattended* AEPs. Note, this window represents the cluster of cumulative time points reaching significance. However, the window of significance varied from channel to channel. For example, at channel Cz, the window of significance was confined to 133–185 ms. This period begins in the later part of the N1 wave and appears to be also partly due to a shift in the N1 latency — earlier for the *AV-attended* condition than the *AV-unattended* condition. There were no differences between the AEP waveforms of the *A-attended* and *A-unattended* conditions (negative cluster, $p = 0.55$). Figure 3(B) shows the topographies of the AV contrast AEP waveforms (left panels) and auditory-only contrast AEP waveforms (right panels) for the period of 133–197 ms. The rightmost topography of each contrast also shows the *t*-value topography for the 133–197 ms significant window and the cluster of channels (bold dots) that exhibited this effect. The *t*-value topography is revealing, because it shows that the observed difference most likely represents auditory sources, with maximum negativity occurring fronto-centrally (blue), with reversals (positivity) around posterior-temporal sites (red-orange). Figure 3(C) shows the boxplot of the AEP amplitude for the 133–185 ms significant window for all conditions at channels Cz.

A caveat of the above results is that significance in one contrast and nonsignificance in the second contrast does not indicate an interaction between modality (auditory-only *versus* AV) and attentional state (attended *versus* unattended) (Gelman and Stern, 2006; Nieuwenhuis *et al.*, 2011). To properly address this issue, we conducted a *post-hoc* Analysis of Variance (ANOVA) to test for interaction. The mean individual N1 amplitudes for the significant period (133–185 ms) at channel Cz were obtained for all conditions and contrasted using an ANOVA, with the variables modality and attentional state. The ANOVA revealed a main effect of attentional state [$F(1, 18) = 7.2$, $p = 0.015$] and an interaction between the variables [$F(1, 18) = 4.5$, $p = 0.049$]. The interaction was attributed to (1) smaller N1 amplitudes occurring for *AV-attended* than *AV-unattended* (*p* = 0.036) but not for *A-attended* than *A-unattended* (*p* = 1); (2) smaller N1 amplitudes occurring for *AV-attended* than *A-attended* (*p* = 0.038) but not for *AV-unattended versus A-unattended* (*p* = 0.99).

3.2.2. *AV* versus *Auditory-Only*

3.2.2. *AV* versus *Auditory-Only*—Figure 4(A) shows the AV *versus* auditory-only AEP waveforms temporally-locked to the noise onset at channel FCz for the attended (left panel) and unattended (right panel) streams. The CBPT revealed a fronto-central negative AEP cluster distinguishing the *AV-attended* from the *A-attended* in the period of 131–174 ms (*p* = 0.018). The amplitude AEP values in this period were less negative (smaller) for the *AV-attended* AEP than the *A-attended* AEPs. At channel FCz, the window of significance was confined to 133–174 ms. This period begins in the later part of the N1 and appears to also be partly due to a shift in the N1 latency — earlier for the *AV-attended* than the *A-attended*. There were no differences between the AEP waveforms of the *AV-unattended* and *A-unattended* waveforms (*p* = 0.55). Figure 4(B) shows the topographies of the attended contrast AEP waveforms (left panels) and unattended contrast AEP waveforms (right panels) for the period 133–174 ms. The rightmost topography of each contrast also shows the *t*-value topography for this period and the cluster of channels (bold dots) that exhibit this effect. The *t*-value topography shows that the observed differences between AV and auditory-only conditions most likely represent auditory sources, with maximum negativity occurring fronto-centrally (blue), with reversals (positivity) around posterior-temporal sites (red-orange). Figure 4(C) shows the boxplot of the AEP amplitude for the 133–174 ms window for all conditions at channel FCz.

Similar to the previous section, to test for interaction, we conducted a *post-hoc* ANOVA. The mean individual N1 amplitudes for the significant period (133–174 ms) at channel FCz was obtained for all conditions, and then contrasted using an ANOVA, with the variables modality and attentional state. The ANOVA revealed a main effect of modality [*F*(1, 18) = 10.2, *p* = 0.005] and an interaction approaching significance between the variables [*F*(1, 18) = 3.98, *p* = 0.061]. The interaction was attributed to smaller N1 amplitudes occurring for *AV-attended* than *A-attended* (*p* = 0.002) but not for *AV-unattended versus A-unattended* (*p* = 0.43).

4. Discussion

We sought to understand the neurophysiology supporting visual enhancement of speech comprehension in a ‘cocktail party’. In the AV task, we presented individuals with two concurrent audio sentences spoken by individuals of different genders and a video of a human speaker uttering one of the sentences. In the control auditory-only condition, the video was a static picture of the human speaker. Individuals attended to the sentence belonging to the voice of the gender of the person in the video. Our experimental design incorporated a noise segment in place of a speech segment along each of the concurrent speech streams. The inhibition of the N1 AEP response to noise onsets signified how well the relevant speech stream was encoded in the auditory cortex. Smaller N1 amplitude indicated that listeners perceived the speech more coherently (less interrupted by noise). We found more robust N1 inhibition for the *AV-attended* than the *AV-unattended* sentences and for the *AV-attended versus* the *A-attended* sentences. There were no differences between the attended and unattended sentences of the auditory-only task or between the two modalities for the unattended sentences. These findings demonstrate that visual enhancement of speech stream selection is not merely an attentional process. Indeed, these findings suggest that comprehension and stream segregation of the relevant from irrelevant speech stream are optimized by audiovisual congruency, in addition to attention. This concurs with Bhat *et al.*'s (2014) findings. They presented individuals with AV-congruent and -incongruent words that had a speech segment replaced by noise. They found that the N1 of the noise onsets/offsets was significantly inhibited when individuals perceived the AV-congruent words as continuous *versus* interrupted. No N1 differences were observed for AV-incongruent words. While attention is an important element for speech segregation, the saliency of a bottom-up signal (e.g., the acoustics) restricts attentional selectivity (Shinn-Cunningham, 2008; Talsma *et al.*, 2010). Thus, attention alone may not be sufficient to restore noise-replaced segments. However, congruent visual speech information may strengthen the formation of auditory object representations (e.g., phonemes) of the relevant speech stream, and in turn optimizes restoration of the noise-replaced speech segments, leading to stronger comprehension and stream segregation.

Neurophysiologically, suppression of the N1 AEP, as well as of the P1 and P2 AEPs, has been associated with AV integration (Baart, 2016; Baart *et al.*, 2014; Besle *et al.*, 2004; Pilling, 2009; Shatzer *et al.*, 2018; Stekelenburg and Vroomen, 2007, 2012; van Wassenhove *et al.*, 2005). Van Wassenhove *et al.* showed that the N1 was smaller and occurred earlier for the AV *versus* auditory-only percepts. We found the same pattern in the current study. Van Wassenhove *et al.* explained their findings in terms of the predictive coding model, whereby visual speech predicts the unfolding speech cues and renders certain auditory activity redundant, hence the suppressed N1 AEP (Besle *et al.*, 2004; van Wassenhove *et al.*, 2005). In the context of the current design, visual speech may predict the unfolding amplitude variations of the speech envelope, thus visually enhancing the fidelity of the speech envelope representations, and in turn reducing perceptual sensitivity to interruptions in the envelope. This account is supported by the findings of Zion Golumbic *et al.* (2013). However, visual enhancement of speech comprehension cannot be limited to the speech envelope. Previous reports (Shahin *et al.*, 2017; Shatzer *et al.*, 2018) showed that synchrony judgment of asynchronous auditory and visual stimuli (mouth movements and corresponding acoustic speech) is significantly influenced by the spectral fidelity of the speech, even when the speech envelope is held constant. This demonstrates that visual networks interact with formant structures — the building blocks of phonemes. Also, the McGurk illusion (McGurk and Macdonald, 1976), indexing visually-mediated alteration of auditory perception, offers very convincing evidence of visual interaction with formant dynamics. Furthermore, Abbott and Shahin (2018) recently revealed that the McGurk illusion and visually-mediated phonemic restoration are byproducts of the same underlying AV mechanism. The Dynamic Reweighting Model (DRM; Bhat *et al.*, 2015), which coincides with the predictive coding model, putatively outlines how visual context interacts with the spectral profile of speech, i.e., phonetic representations. It posits that visual enhancement of spoken language processing is attributed to a visually-directed shift of processing along the auditory cortex. As meaningfulness of visual speech increases (mouth movements clearly conveying phonemes), the visual system directs processing along the auditory cortex by inhibiting low-level auditory networks while exciting high-level auditory networks. This shift allows visual networks to engage phonetic representations (e.g., formant dynamics) at the non-primary auditory cortex, while simultaneously inhibiting processing of simple features in sounds, such as acoustic onsets. This is especially useful in noisy situations where phonetic encoding at the auditory cortex can be reinforced by visual cues while onset encoding of interfering sounds is suppressed. In the context of the current design, the visually-mediated upward shift results in filling-in of missing speech representations, leading to more salient phonetic representations, while interfering noise onsets are suppressed due to inhibition of low-level auditory networks.

While visual context is key to the visually-mediated N1 suppression account, Stekelenburg and Vroomen (2007) offer a modified interpretation. They showed that N1 suppression occurs regardless of whether the preceding visual stimulus is contextually meaningful for the incoming auditory percept; rather it is tied to whether the visual percept anticipates the timing of the auditory percept. This does not fit with our interpretation. In our design, it was unfeasible for vision to anticipate the timing of white noise onset, since speech, not white noise, is predicted by visual speech; in addition, noise occurrence did not have a fixed time with respect to sentence onsets — it occurred either 25% or 75% from the beginning of the sentence. The divergence of our and van Wassenhove *et al.*'s (2005) N1 characterization and that of Stekelenburg and Vroomen (2007) may be explained by the differing analytic approaches employed. In Stekelenburg and Vroomen (2007), and many other studies, including our own (to cite a few, Baart and Samuel, 2015; Shahin *et al.*, 2018; Teder-Sälejärvi *et al.*, 2002), they subtracted the visual-only evoked potentials from the AV-evoked potentials to assess auditory effects. We did not do that here. Previously, Shahin *et al.* (2018) argued against using the subtraction method, since the visual-only waveforms may also contain auditory activity. Previous findings revealed that the auditory cortex is activated during silent lip-reading (Abbott and Shahin, 2018; Calvert *et al.*, 1997; Pekkola *et al.*, 2005). Thus, subtraction of this condition may negate this activity. Shahin *et al.* (2018) resorted to other analyses to rule out visual-evoked potentials' contamination of the AEPs. For example, Shahin *et al.* (2018), used Independent Component Analysis (ICA) to remove components that are consistent with visual activity, as an alternative to waveform subtraction, to draw conclusions on visually-mediated auditory effects.

In conclusion, the current results provide evidence that visual networks support comprehension in a 'cocktail party' *via* suppression of the auditory cortex's response to acoustic onsets of irrelevant sounds. However, acoustic onset representation suppression may also be driven by visual reinforcement of the relevant speech cues (e.g., speech envelope, Zion Golumbic *et al.*, 2013), or filling-in of missing phonetic information (Abbott and Shahin, 2018; Bhat *et al.*, 2014; Shahin and Miller, 2009). Thus, visual 'net enforcement' of speech may be due to neural reinforcement of the relevant speech cues (e.g., speech envelope, phonetic information), the weakening of neural responses of irrelevant speech cues, or both. Furthermore, in our design, mouth movement was an additional cue that distinguished the AV from auditory-only conditions. In the auditory-only condition, the two sentences varied in speech envelope and pitch. In the AV condition, the two streams varied in speech envelope, pitch and mouth movements. Hence, the current study does not inform on how perceivers weigh each cue and how one cue influences the robustness of the other cues. Future undertaking would benefit from examining cross-modal cue interaction and its benefit to speech segregation.

Funding

This work was funded by the National Institute on Deafness and Other Communications Disorders, Grant R01 DC-013543 (AJS).

References

Abbott NT and Shahin AJ (2018). Cross-modal phonetic encoding facilitates the McGurk illusion and phonemic restoration, J. Neurophysiol 120, 2988–3000. [PubMed: 30303762]

Alho K, Töttölä K, Reinikainen K, Sams M and Näätänen R (1987). Brain mechanism of selective listening reflected by event-related potentials, Electroencephalogr. Clin. Neurophysiol 68, 458–470. [PubMed: 2444425]

Baart M (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays, Psychophysiology 53, 1295–1306. [PubMed: 27295181]

Baart M and Samuel AG (2015). Turning a blind eye to the lexicon: ERPs show no crosstalk between lip-read and lexical context during speech sound processing, J. Mem. Lang 85, 42–59.

Baart M, Stekelenburg JJ and Vroomen J (2014). Electrophysiological evidence for speech-specific audiovisual integration, Neuropsychologia 53, 115–121. [PubMed: 24291340]

Banks B, Gowen E, Munro KJ and Adank P (2015). Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation, Front. Hum. Neurosci 9, 422 DOI:10.3389/fnhum.2015.00422. [PubMed: 26283946]

Besle J, Fort A, Delpuech C and Giard M-H (2004). Bimodal speech: early suppressive visual effects in human auditory cortex, Eur. J. Neurosci 20, 2225–2234. [PubMed: 15450102]

Bhat J, Pitt MA and Shahin AJ (2014). Visual context due to speech-reading suppresses the auditory response to acoustic interruptions in speech, Front. Neurosci 8, 173 DOI:10.3389/fnins.2014.00173. [PubMed: 25053937]

Bhat J, Miller LM, Pitt MA and Shahin AJ (2015). Putative mechanisms mediating tolerance for audiovisual stimulus onset asynchrony, J. Neurophysiol 113, 1437–1450. [PubMed: 25505102]

Bregman AS (1990). Auditory Scene Analysis: the Perceptual Organization of Sound. MIT Press, Cambridge, MA, USA.

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SCR, McGuire PK, Woodruff PWR, Iversen SD and David AS (1997). Activation of auditory cortex during silent lipreading, Science 276, 593–596. [PubMed: 9110978]

Cherry EC (1953). Some experiments on the recognition of speech, with one and with two ears, J. Acoust. Soc. Am 25, 975–979.

Crosse MJ, Butler JS and Lalor EC (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions, J. Neurosci 35, 14195–14204. [PubMed: 26490860]

Delorme A and Makeig S (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, J. Neurosci. Methods 134, 9–21. [PubMed: 15102499]

Drullman R, Festen JM and Plomp R (1994). Effect of reducing slow temporal modulations on speech reception, J. Acoust. Soc. Am 95, 2670–2680. [PubMed: 8207140]

Gelman A and Stern H (2006). The difference between ‘significant’ and ‘not significant’ is not itself statistically significant, Am. Stat 60, 328–331.

Grant KW 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. [PubMed: 11008820]

Grimault N, Micheyl C, Carlyon RP, Arthaud P and Collet L (2000). Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency, J. Acoust. Soc. Am 108, 263–271. [PubMed: 10923890]

Harte N and Gillen E (2015). TCD-TIMIT: an audio-visual corpus of continuous speech, IEEE Trans. Multimed 17, 603–615.

Ihlefeld A and Shinn-Cunningham B (2008). Disentangling the effects of spatial cues on selection and formation of auditory objects, J. Acoust. Soc. Am 124, 2224–2235. [PubMed: 19062861]

Jesse A and Janse E (2012). Audiovisual benefit for recognition of speech presented with single-talker noise in older listeners, Lang. Cogn. Proc 27, 1167–1191.

Kerlin JR, Shahin AJ and Miller LM (2010). Attentional gain control of ongoing cortical speech representations in a ‘cocktail party’, J. Neurosci 30, 620–628. [PubMed: 20071526]

Lopez-Calderon J and Luck SJ (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials, Front. Hum. Neurosci 8, 213 DOI:10.3389/fnhum.2014.00213. [PubMed: 24782741]

Maris E and Oostenveld R (2007). Nonparametric statistical testing of EEG- and MEG-data, J. Neurosci. Methods 164, 177–190. [PubMed: 17517438]

McGurk H and Macdonald J (1976). Hearing lips and seeing voices, Nature 264, 746–748. [PubMed: 1012311]

Nieuwenhuis S, Forstmann BU and Wagenmakers E-J (2011). Erroneous analyses of interactions in neuroscience: a problem of significance, Nat. Neurosci 14, 1105–1107. [PubMed: 21878926]

Oostenveld R, Fries P, Maris E and Schoffelen J-M (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data, Comput. Intell. Neurosci 2011, 1 DOI:10.1155/2011/156869. [PubMed: 21837235]

O’Sullivan JA, Crosse MJ, Power AJ and Lalor EC (2013). The effects of attention and visual input on the representation of natural speech in EEG, Conf. Proc. IEEE Eng. Med. Biol. Soc 2013, 2800–2803. [PubMed: 24110309]

Oxenham AJ (2008). Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants, Trends Amplif. 12, 316–331. [PubMed: 18974203]

Pekkola J, Ojanen V, Autti T, Jääskeläinen IP, Möttönen R, Tarkiainen A and Sams M (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T, NeuroReport 16, 125–128. [PubMed: 15671860]

Pilling M (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception, J. Speech, Lang. Hear. Res 52, 1073–1081. [PubMed: 19641083]

Riecke L, Esposito F, Bonte M and Formisano E (2009). Hearing illusory sounds in noise: the timing of sensory-perceptual transformations in auditory cortex, Neuron 64, 550–561. [PubMed: 19945396]

Samuel AG (1981). Phonemic restoration: insights from a new methodology, J. Exp. Psychol. Gen 110, 474–494. [PubMed: 6459403]

Shahin AJ and Miller LM (2009). Multisensory integration enhances phonemic restoration, J. Acoust. Soc. Am 125, 1744–1750. [PubMed: 19275331]

Shahin AJ, Kerlin JR, Bhat J and Miller LM (2012). Neural restoration of degraded audiovisual speech, Neuroimage 60, 530–538. [PubMed: 22178454]

Shahin AJ, Shen S and Kerlin JR (2017). Tolerance for audiovisual asynchrony is enhanced by the spectrotemporal fidelity of the speaker’s mouth movements and speech, Lang. Cogn. Neurosci 32, 1102–1118. [PubMed: 28966930]

Shahin AJ, Backer KC, Rosenblum LD and Kerlin JR (2018). Neural mechanisms underlying cross-modal phonetic encoding, J. Neurosci 38, 1835–1849. [PubMed: 29263241]

Shatzer H, Shen S, Kerlin JR, Pitt MA and Shahin AJ (2018). Neurophysiology underlying influence of stimulus reliability on audiovisual integration, Eur. J. Neurosci 48, 2836–2848. [PubMed: 29363844]

Shinn-Cunningham BG (2008). Object-based auditory and visual attention, Trends Cogn. Sci 12, 182–186. [PubMed: 18396091]

Stekelenburg JJ and Vroomen J (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events, J. Cogn. Neurosci 19, 1964–1973. [PubMed: 17892381]

Stekelenburg JJ and Vroomen J (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception, Neuropsychologia 50, 1425–1431. [PubMed: 22410413]

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

Talsma D, Senkowski D, Soto-Faraco S and Woldorff MG (2010). The multifaceted interplay between attention and multisensory integration, Trends Cogn. Sci 14, 400–410. [PubMed: 20675182]

Teder-Sälejärvi WA, McDonald JJ, Di Russo F and Hillyard SA (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings, Cogn. Brain Res 14, 106–114.

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

Warren RM (1970). Perceptual restoration of missing speech sounds, Science 167, 392–393. [PubMed: 5409744]

Woods DL, Alain C, Diaz R, Rhodes D and Ogawa KH (2001). Location and frequency cues in auditory selective attention, J. Exp. Psychol. Hum. Percept. Perform 27, 65–74. [PubMed: 11248941]

Zeng F-G, Oba S, Garde S, Sininger Y and Starr A (1999). Temporal and speech processing deficits in auditory neuropathy, NeuroReport 10, 3429–3435. [PubMed: 10599857]

Zion Golumbic E, Cogan GB, Schroeder CE and Poeppel D (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a ‘cocktail party’, J. Neurosci 33, 1417–1426. [PubMed: 23345218]

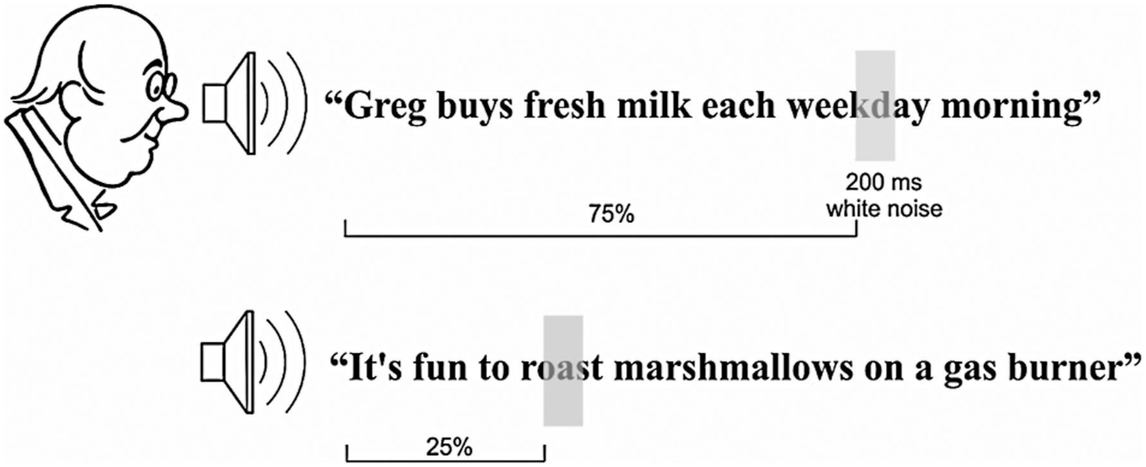
“Greg buys fresh milk each weekday morning”
200 ms
white noise
75%
“It's fun to roast marshmallows on a gas burner”
25%

Figure 1.

Audiovisual task experimental design. Participants watched and listened to a human speaker uttering a sentence while also hearing a concurrent sentence of a speaker (no video) of the opposite gender. A 200 ms segment of each acoustic sentence was replaced by white noise beginning at 25% following sentence sound onset of one sentence and at 75% following sentence sound onset of the other sentence. Audiovisual pairing and noise placements were counterbalanced across trials to rule out stimulus differences. Individuals transcribed what they heard during randomly chosen trials throughout the experiment. A similar task without visual mouth movements (auditory-only task) served as a control condition.

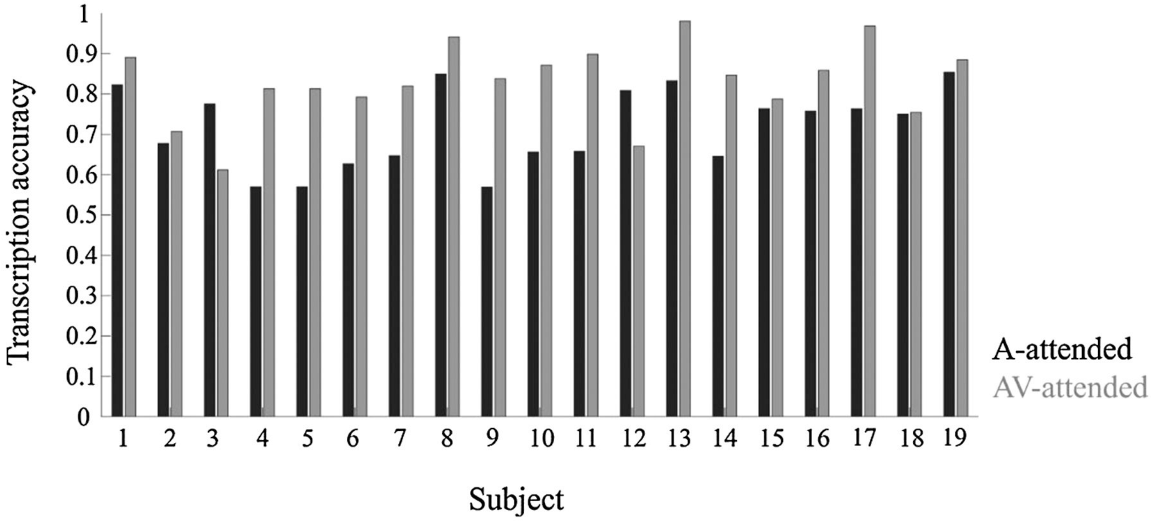

Figure 2.

Individual transcription accuracy for *AV-attended* and *A-attended* speech sentences.

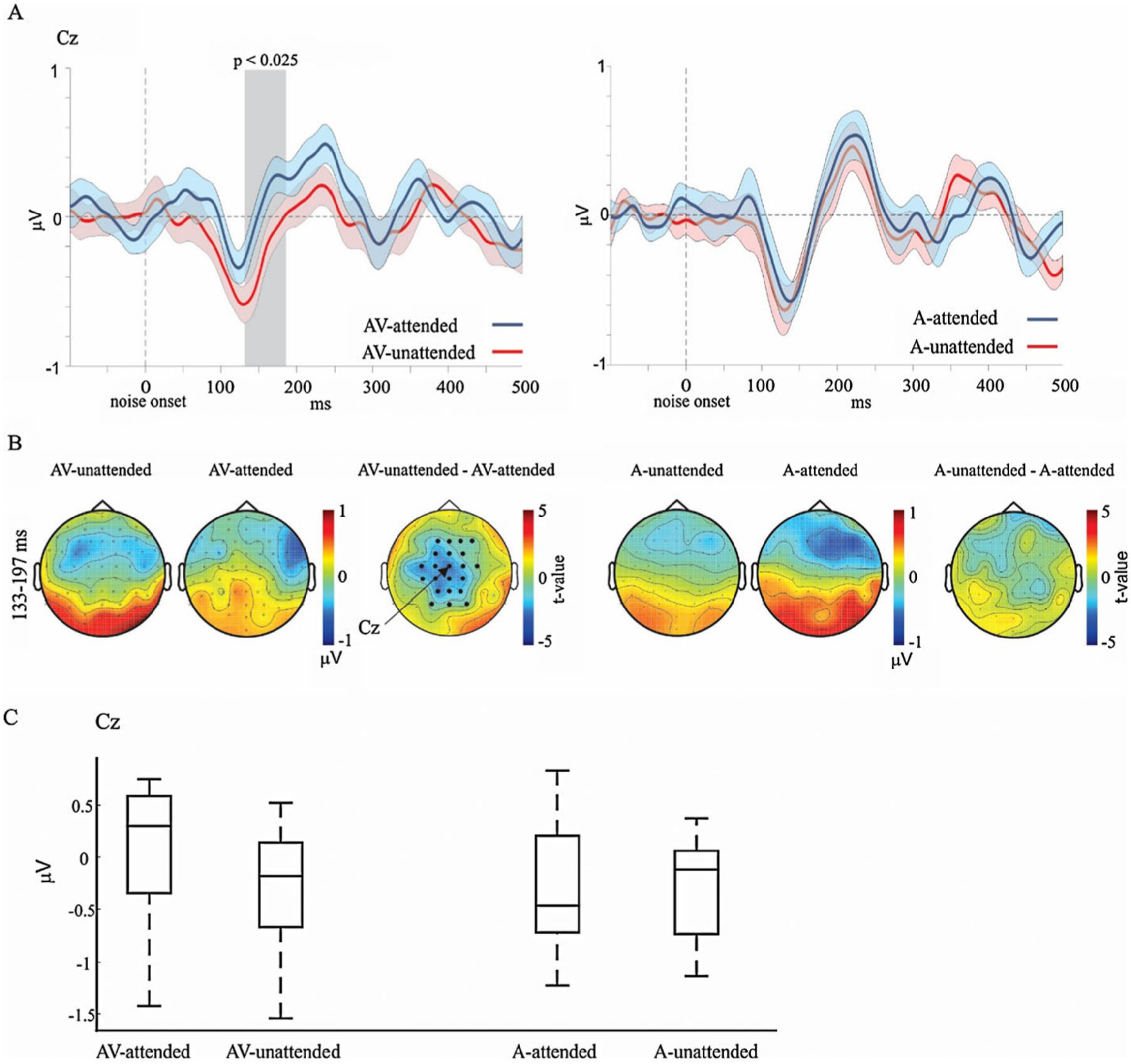

Figure 3.

(A) Auditory evoked potential (AEP) waveforms at channel Cz, time-locked to noise onsets of the *AV-attended* and *AV-unattended* sentences (left panel) and *A-attended* and *A-unattended* sentences (right panel). Gray rectangular area represents the window (133–185 ms) of significance distinguishing the AEP waveforms of *AV-attended* and *AV-unattended* sentences. (B) Left panel: Topographies of the mean AEP activity within the window of significance (13–197 ms) for *AV-unattended*, *AV-attended* conditions and the mean *t*-value topography of the significant window distinguishing the two AEP waveforms. Right panel: similar to left panel, but for *A-unattended* and *A-attended* conditions. (C) Box plots of mean amplitude activity occurring within the window of significance (133–185 ms) at channel Cz for all conditions.

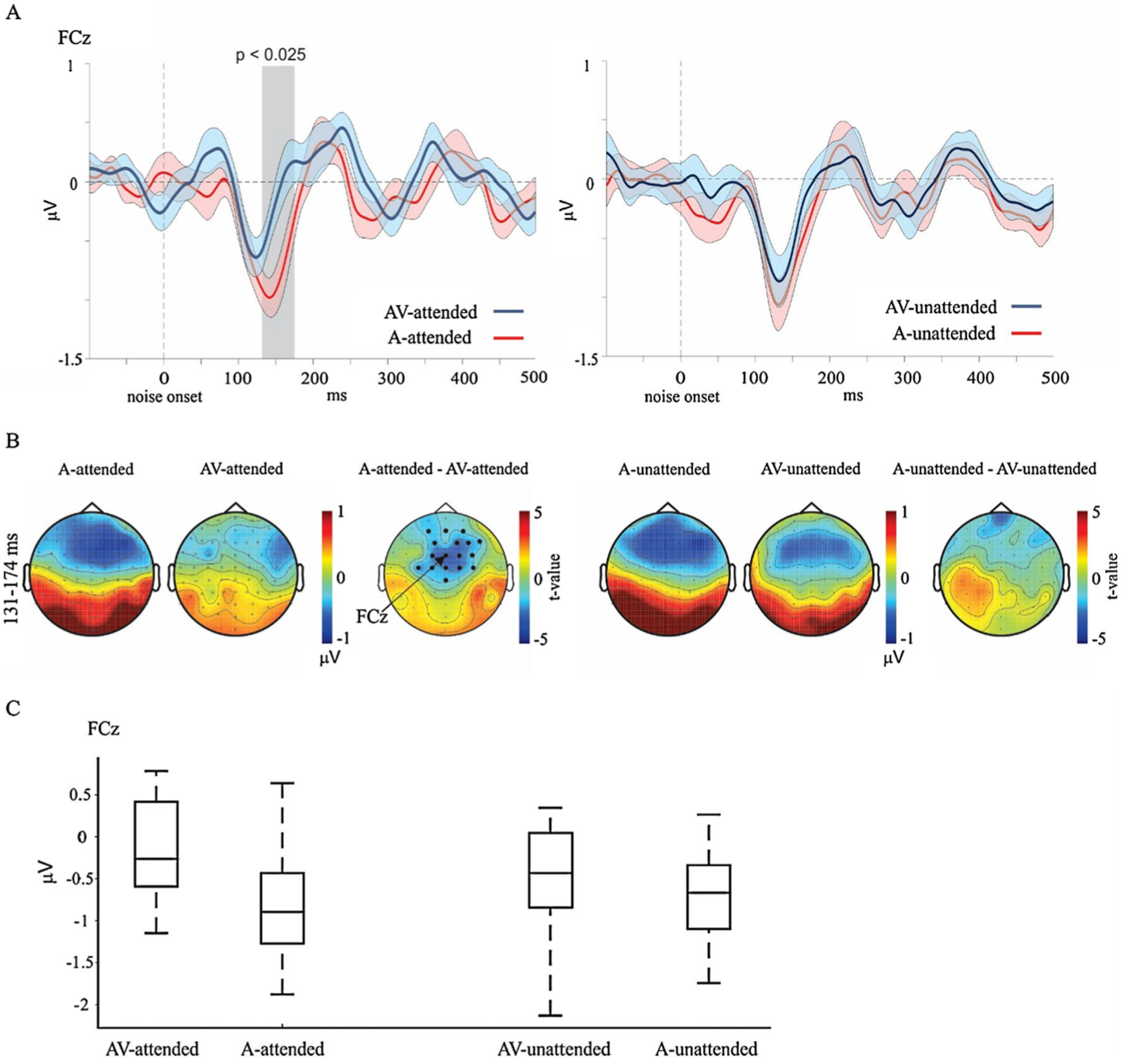

Figure 4.

(A) Auditory evoked potential (AEP) waveforms at channel FCz, time-locked to noise onsets of the *AV-attended* and *A-attended* percepts (left panel) and *AV-unattended* and *A-unattended* conditions (right panel). Gray rectangular area represents the window (133–174 ms) of significance distinguishing the AEP waveforms of *AV-attended* and *A-attended* conditions. (B) Left panel: topographies of the average activity within the window of significance for *A-attended*, *AV-attended* and the mean *t*-value topography of the significant window (131–174 ms) distinguishing the two AEP waveforms. Right panel: similar to left panel but for *A-unattended* and *AV-unattended* conditions. (C) Box plots of mean amplitude activity occurring within the window of significance (133–174 ms) for all conditions.