Seeing the Intensity of a Sound-producing Event Modulates the Amplitude of the Initial Auditory Evoked Response

Sol Libesman, Damien J. Mannion, and Thomas J. Whitford

Abstract

■ An auditory event is often accompanied by characteristic visual information. For example, the sound level produced by a vigorous handclap may be related to the speed of hands as they move toward collision. Here, we tested the hypothesis that visual information about the intensity of auditory signals are capable of altering the subsequent neurophysiological response to auditory stimulation. To do this, we used EEG to measure the response of the human brain (n = 28) to the audiovisual delivery of handclaps. Depictions of a weak handclap were accompanied by auditory handclaps at low (65 dB) and intermediate (72.5 dB) sound levels, whereas depictions of a vigorous handclap were accompanied by auditory handclaps at intermediate (72.5 dB) and high (80 dB) sound levels. The dependent variable was the amplitude of the initial negative component (N1) of the auditory evoked potential. We find that identical clap sounds (intermediate level; 72.5 dB) elicited significantly lower N1 amplitudes when paired with a video of a weak clap, compared with when paired with a video of a vigorous clap. These results demonstrate that intensity predictions can affect the neural responses to auditory stimulation at very early stages (<100 msec) in sensory processing. Furthermore, the established sound-level dependence of auditory N1 amplitude suggests that such effects may serve the functional role of altering auditory responses in accordance with visual inferences. Thus, this study provides evidence that the neurally evoked response to an auditory event results from a combination of a person's beliefs with incoming auditory input. ■

INTRODUCTION

Hermann von Helmholtz, a pioneer in audition research, once described the experience of watching a plucked guitar string by writing, "When we strike a string, its vibrations are at first sufficiently large for us to see them, and its corresponding tone is loudest. The visible vibrations become smaller and smaller, and at the same time the loudness diminishes." (Helmholtz, 1877, p. 10). Here, Helmholtz presented a clear demonstration that visual cues have the capacity to provide information about auditory intensity. Everywhere around us, we see the relationship between the characteristics of a physical action and the nature of the auditory consequence. For example, when someone is whispering, they will narrowly open their lips, whereas a shout will involve a wide open mouth. Similarly, a soft clap will involve a slow, weak movement, whereas a loud clap will often involve a fast, powerful movement. Although it is clear that visual cues can provide information about the expected intensity of auditory events, it is unclear whether visual cues can modulate the responsiveness of the primary auditory cortex to auditory events. Addressing this question is the aim of this study.

Using ERPs acquired through EEG, we investigated how predictions regarding the expected intensity of sounds (specifically, handclaps) affected the evoked neurophysiological responses (specifically, the amplitude of the N1 component of the auditory-evoked potential). The N1 component is the negative peak that appears approximately 100 msec following the onset of a brief auditory stimulus. It has been found to have dominant origins in the auditory cortex (Pantev et al., 1995). An important feature of the N1 is that its amplitude is known to be intensity dependent; sounds of higher intensity elicit larger N1 amplitudes than sounds of lower intensity (Mulert et al., 2005; Brocke, Beauducel, John, Debener, & Heilemann, 2000; Dierks et al., 1999; Hegerl, Gallinat, & Mrowinski, 1994; Rapin, Schimmel, Tourk, Krasnegor, & Pollak, 1966).

This study investigated the interaction between visual information and auditory intensity predictions by pairing auditory signals (sounds of hands clapping) with videos of a person performing handclaps. Two different videos were presented: a video of an actor producing a weak handclap and a video of an actor producing vigorous handclap. These visuals were suggestive of generating either low or high auditory intensity, respectively. Handclap sounds of varying intensities were paired with these videos. Our hypothesis was that the degree of activation of

University of New South Wales, Sydney

https://doi.org/10.1162/jocn_a_01486

the primary auditory cortex in response to an auditory event is the combination of (1) the intensity of the auditory signal at the ear and (2) the concomitant visual stream information, depicting the generation of the auditory signal at a particular intensity. Specifically, we propose that the N1 amplitude to a handclap at a given intensity would be greater when paired with a video depicting a vigorous handclap than when paired with a video depicting a weak handclap. Here, the amplitude of the N1 component generated from the received signal shifts toward the response that would be generated from the expected (visual) signal.

METHODS

Participants

A total of 36 participants were recruited from a pool of students enrolled in an introductory psychology course at UNSW Sydney. Four participants did not produce data because of technical problems with data acquisition, and four participants were excluded for failing to identify a (predefined) requisite number of "catch" trials (see Procedure section). Of the remaining 28 participants, 13 were women, and 25 were right-handed. The majority of participants (20 of 28) were between 17 and 20 years old. Participants received course credit for their involvement and gave informed and written consent in accordance with the experiment protocols approved by the Human Research Ethics Advisory Panel in the School of Psychology, UNSW Sydney (#2968). All participants were naïve to the purposes of the experiment.

Apparatus

Auditory stimuli were presented via an AudioFile device (Cambridge Research Systems) and over-ear headphones (Beyerdynamic; Model DT990 Pro). The sound level produced by the headphones was determined using an artificial ear, microphone, and analyzer (Brűel & Kjæer, Nærum, Denmark; Models 4152, 4144, and 2250, respectively). All subsequently reported sound levels are in units of dB SPL as determined by this calibration method.

Visual stimuli were presented from a XLT2420T BenQ computer monitor (60 Hz, 1920 × 1080 resolution). Participants viewed the monitor in a well-lit room at a distance of 82 cm for a visual angle of 37.0° × 20.8°. The experiment was controlled using PsychoPy (Peirce, 2007, 2008).

The EEG was recorded using a BioSemi ActiveTwo system with 64 Ag–AgCl active electrodes placed per the extended 10–20 system, sampling at a rate of 2048 Hz. The Ag/Ag–Cl electrodes were connected to all 64 cap channels, with additional electrodes attached to the mastoids and nose and placed 1 cm from the outer canthi of both eyes and 1 cm under the left eye to monitor horizontal and vertical eye movements. Online referencing was to sensors located in the parietal region of the cap (Common Mode Sense active electrode, Driven Right Leg passive electrode). (The continuous EEG record for each participant is available at https://doi.org/10.6084/m9.figshare.c.4286837.v3.) EEG data were processed using BrainVision Analyser (Version 2.1).

Stimuli

Auditory claps were produced by convolving an anechoic recording of a clap with a room impulse response. The impulse response was obtained from the Salford-BBC Spatially-sampled Binaural Room Impulse Responses database (Satongar, Lam, & Pike, 2014) and characterized a frontally positioned source in an enclosed room at a distance of 1 m. The clap (obtained from https://freesound.org/people/Anton/sounds/345) was downsampled to the sampling rate of the impulse response (48 kHz) before convolution, and the resulting waveform was again downsampled to the sampling rate of the presentation device (44.1 kHz). Manipulations of clap level were produced by multiplications of this waveform. The sounds for each of the three audible levels are available at https://doi.org/10.6084/m9.figshare.c.4286837.v3.

Visual depictions of claps were produced by recording videos of the first author (with visible hands, arms, and torso) producing a handclap with either "weak" or "strong" levels of force. Compared with the "weak" clap, the top hand in the "strong" clap moved from a higher position, traveled more rapidly to the bottom hand, and produced greater vibration on collision. The recordings were made at a spatial resolution of 1920 × 1080 pixels using a Sony Cybershot RX100 digital camera. Videos were converted to grayscale, resampled to 960 × 540 pixels (giving a viewing angle of 18.5° × 10.4° on the presentation monitor), and temporally cropped such that the sequence had a duration of 54 frames (900 msec). The contact of the hands occurred at frame 27 (thus providing 433 msec of anticipatory motion). The pixel intensities in the resulting sequences were then each normalized to have a mean of 0.0 and a standard deviation of 0.45 before being clipped to be within a $[-1, +1]$ interval. This was performed to enforce that the two videos were similar in their overall distributions of pixel intensity. Example frames are shown in Figure 1, and videos are available at https://doi.org/10.6084/m9.figshare.c.4286837.v3.

Design

We used a within-subject design with a total of six cells, with three conditions each for "weak" and "strong" visual depictions. The three conditions for the "weak" clap visual depiction involved the clap sounds being delivered at three levels (silent, 65 dB, and 72.5 dB), and the three conditions for the "strong" clap visual depiction involved the clap sounds being delivered at three levels (silent, 72.5 dB, and 80 dB). We chose this design over a fully crossed alternative to retain the ecological

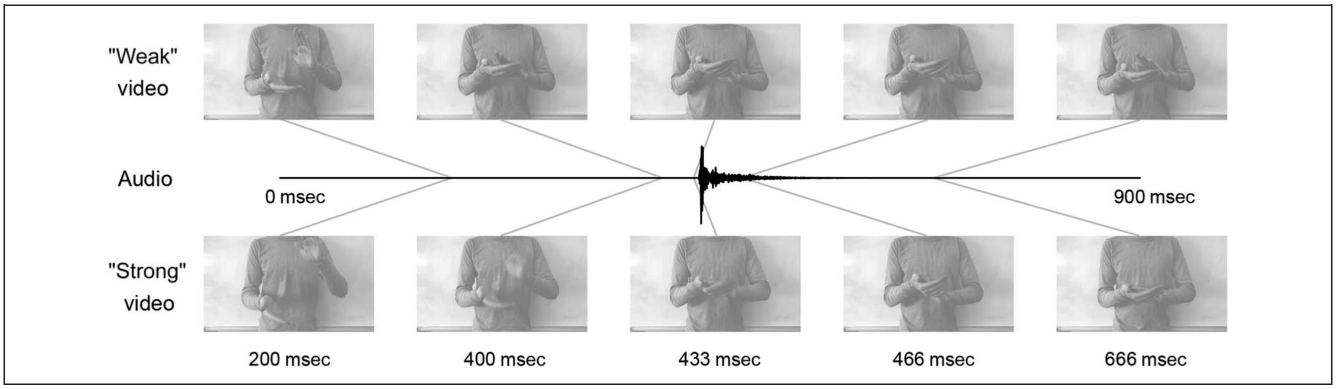

Figure 1. The time course of the two videos administered in this experiment.

audiovisual association; over the course of the experiment, a "weak" visual clap was more likely to be paired with an auditory clap of a lower sound level than a "strong" visual clap.

Procedure

The experimental task was conducted in a single session, which lasted approximately one and a half hours. After being fitted with an EEG cap and electrodes, participants had their EEG continuously recorded as they completed the experimental protocol. This protocol was approximately 50 min in duration and consisted of 12 experimental runs. Each run contained 39 trials. Each trial began with the onset of the visual depiction of a clap. At the frame at which the hands collided, a parallel port signal was emitted, which simultaneously triggered the onset of auditory delivery and marked the EEG record. Following conclusion of the video sequence, there was an interval of random duration (uniformly sampled from between 3 and 5 sec) before the trial ended during which the screen was uniformly gray.

The first eight runs of the experiment consisted of trials of each of the four audiovisual conditions (weak-video, 65 dB sound; weak-video, 72.5 dB sound; strong-video, 72.5 dB sound; strong-video, 80 dB sound). The trials were presented in random order. Over the course of these eight runs, 72 trials of each of the four audiovisual conditions were presented.

The final four runs of the experiment consisted of trials of the two video-only conditions (weak-video, silent; strong-video, silent). Over the course of these four runs, 72 trials of each of the two video-only conditions were presented.

Each run also included three randomly interspersed "catch" trials, intended to allow the identification of participants who were not reliably attending to the clap events. These trials were identical to a randomly selected condition, with the exception that a small green cross was briefly presented (67 msec) following the collision of the hands in the video. Participants were required to demonstrate they had detected the green cross by means of a keypress. These catch trials were not included in the EEG analysis.

Analysis

For each participant, the EEG data were first rereferenced offline to an average of the mastoid electrodes. The continuous EEG was then band-pass filtered from 0.1 to 30 Hz (eighth-order zero-phase Butterworth IIR). ERPs were then extracted, where each ERP was 600 msec in duration and encompassed the 200 msec prior and 400 msec following the onset of the initiation of the clap sound. These ERPs were then baseline-corrected by subtracting the average of the 200 msec preonset signal, separately for each condition. Eyeblink artifacts were then corrected with the method of Gratton, Coles, and Donchin (1983) using an algorithm that involves the subtraction of ocular artifacts by creating a propagation factor that captures the relationship between ocular activity monitored with an electrooculogram (created with external electrodes) and EEG traces at each electrode. Electrodes PO7, P8, Oz, POz, P6, and O2 were leading to more than 75% of trials to be rejected for three participants, and so topographic interpolation was conducted on these electrodes for these participants. For each electrode, ERPs containing a voltage change between adjacent 200-msec intervals in excess of 200 μV or a maximum gradient greater than 50 μV were then excluded. On average, the "weak" visual (65 dB) retained 88% of trials (*SD* = 20%), the "weak" visual (72.5 dB) retained 87% of trials (*SD* = 21%), the "strong" visual (72.5 dB) retained 89% of trials (*SD* = 18%), and the "strong" visual (80 dB) retained 85% of trials (*SD* = 24%). The remaining ERPs were then averaged across trials, separately for each condition. Finally, the ERPs from the silent conditions were subtracted from the audiovisual conditions with the corresponding visual clap ("weak" or "strong"); this was designed to reduce the influence of any purely visual contributions to the ERPs and is a typical procedure in multisensory studies (Stekelenburg & Vroomen, 2007; Guthrie & Buchwald, 1991). The resulting ERPs are used in all subsequent analyses and are shown (averaged across participants) in Figure 2A.

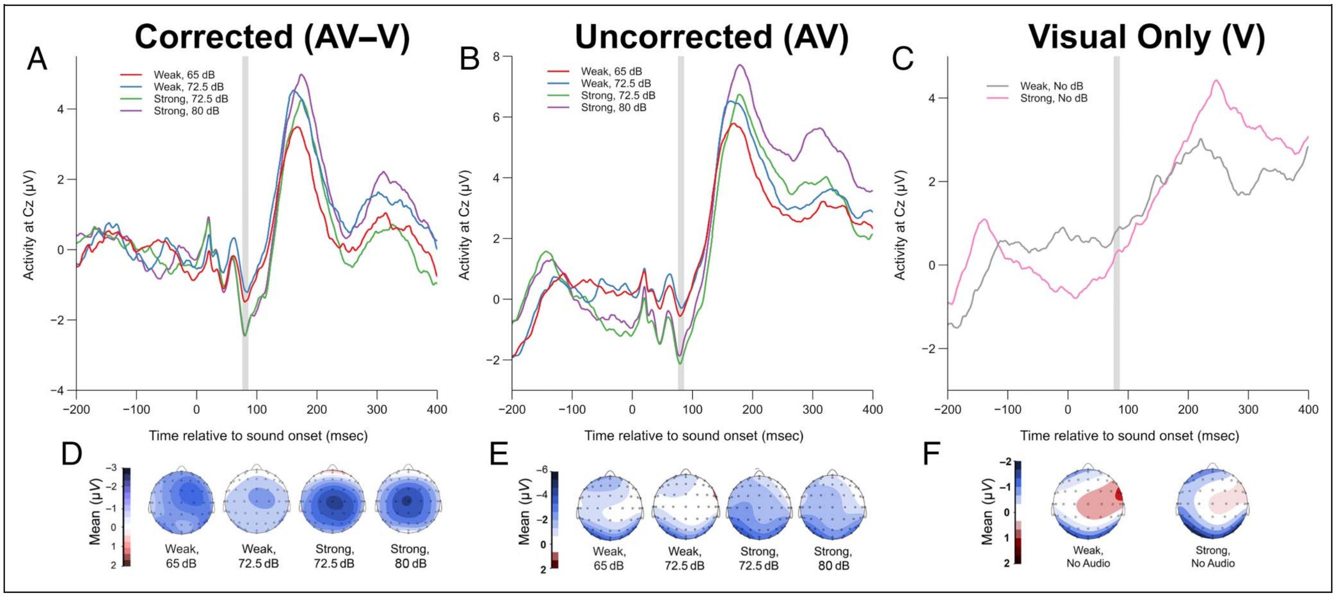

Figure 2. Differences between the AV–V, AV, and V-only waveforms in N1 amplitude. (A) Grand-averaged waveforms for the corrected (AV–V) condition. (B) Grand-averaged waveforms for the uncorrected (AV) condition. (C) Grand-averaged waveforms for the V-only condition. The 0 msec point on the *x* axis represents the onset of the auditory stimulus (which was silent in the V-only condition). All waveforms recorded from electrode Cz, baselined to −200 to 0 msec prestimulus. (D), (E), and (F) represent the scalp topographies for the AV–V, AV, and V-only conditions, respectively. The time window used for the scalp topographies and statistical analysis was 76–86 msec.

The dependent variable was the amplitude of the N1 component of the auditory evoked potential, which is typically elicited by binaural auditory stimulation and has a central topography that is maximal around Cz (Luck, 2005). These characteristics were observed in the current study, as shown in Figure 2D. Hence, all analyses were conducted on electrode Cz (as is common practice; e.g., Oestreich et al., 2015; Vroomen & Stekelenburg, 2010). To determine the time window in which to evaluate the N1, we averaged the ERPs across all conditions and participants to produce a "collapsed localizer" waveform (Luck & Gaspelin, 2017). The time at which such a waveform was at its minimum (81 msec) was used to set an N1 evaluation window of 76–86 msec, common across all participants and conditions. The average voltage in this time window was then extracted from each participant and condition and was used as the measurement of the N1 amplitude. Differences between the N1 amplitude across conditions were evaluated using a paired-sample *t* test.

RESULTS

The aim of this study was to determine whether visual information about the intensity of a sound- generating event influences the neural processing of an associated auditory signal. We paired visual depictions of handclaps with weak or strong force with auditory claps of different sound levels. The key comparison in determining the effect of visual information on auditory processing was between the visual depictions of "weak" and "strong" claps, when paired with the same intensity clap sound (72.5 dB).

As shown in Figure 3, the mean N1 amplitude to 72.5-dB clap audio was larger when paired with the visually strong clap ($M = -2.320$) than when paired with the visually weak clap ($M = -1.107$). This difference ($M = 1.213$, $SEM = 0.585$) was statistically significant (paired sample, $t(27) = 2.073$, $p = .048$, $d = 0.392$), supporting the hypothesis that visual information about the intensity of an auditory event affects the amplitude of the auditory evoked potential.

The N1 amplitudes evoked by the claps at 72.5 dB were similar to the N1 amplitudes evoked by claps at 65 and 80 dB (which were respectively paired with the weak and strong clap visual), as shown in Figure 3A. The mean N1 amplitudes for the weak visual claps were comparable for the 65 dB ($M = -1.429$) and 72.5 dB ($M = -1.107$) auditory intensities, with this difference ($M = -0.322$, $SEM = 0.320$) not being statistically significant, $t(27) = -1.005$, $p = .324$, $d = -0.190$. Similarly, the mean N1 amplitudes for the strong visual claps were comparable for the 72.5 dB ($M = -2.320$) and 80 dB ($M = -2.346$) auditory intensities, with this difference ($M = 0.026$, $SEM = 0.435$) not being statistically significant, $t(27) = 0.059$, $p = .954$, $d = 0.011$.

Examining the conditions used to correct for any purely visual contribution to the ERPs, we find that the voltages evoked by the silent clap videos in the N1 time window were similar across the weak- and strong-video conditions. When sounds were absent, the mean N1 amplitudes were comparable for the weak clap ($M = 0.807$) and strong clap ($M = 0.277$) videos, with this difference ($M = 0.531$, $SEM = 0.526$) not being statistically significant, $t(27) = 1.010$, $p = .321$, $d = 0.191$.

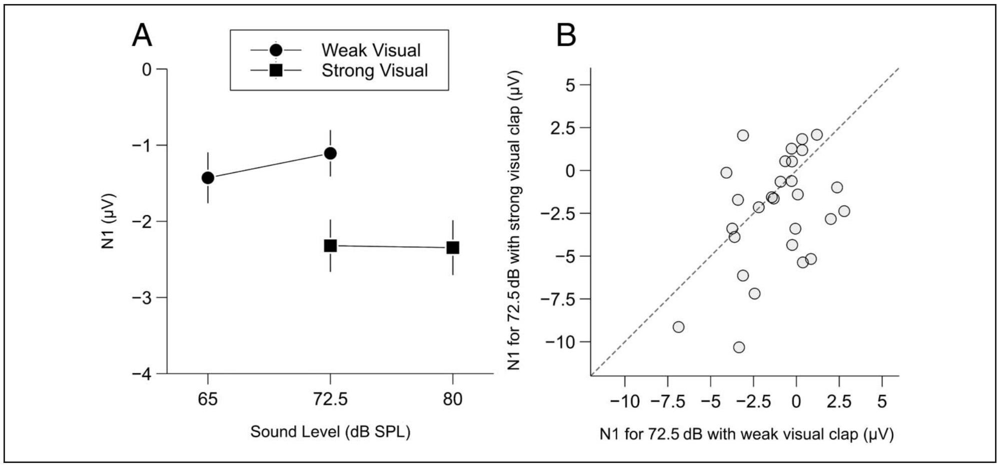

Figure 3. (A) The mean amplitude of participants N1 as a function of sound level, split between the weak-video and strong-video conditions. The *SEM* bars have been corrected to reflect error variance of a within subjects design, as is recommended (Cousineau, 2005). (B) Scatter plot of the mean N1 amplitude for 72.5 dB as a function of the weak and strong videos.

We also conducted a post hoc exploratory analysis that examined the between-video differences (i.e., weak clap video vs. strong clap video) in anticipatory preclap activity across the corrected (AV–V), the uncorrected audiovisual, and the video-only blocks. As shown in Figure 4B and C, the "strong clap" videos elicited a negative-going deflection ~120 msec preclap with a frontal topography that was not present in the "weak clap" videos. Comparing the uncorrected strong and weak clap videos (pooled across auditory stimuli), the between-video difference (M = 1.78, *SEM* = 0.42) was significant in the anticipatory period immediately preclap (paired sample, $t(27) = 4.24$, $p < .001$, $d = 0.567$). A similar result was observed when comparing the strong and weak clap videos in the V-only conditions: a negative-going deflection with a similar time course and topography was again observed in the strong clap condition (but not the weak clap condition), and the difference in prestimulus activity (M = 2.13, *SEM* = 0.63) was statistically significant, $t(27) = 3.371$, $p = .002$, $d = 0.691$. However, these prestimulus differences were not present in the "corrected" waveforms (which were the primary focus of our analysis), suggesting the subtraction of the visual-only condition was effective at eliminating prestimulus differences between the strong and weak clap conditions ($M = −0.07$, *SEM* = 0.589, $t(27) = −0.123$, $p = .902$, $d = 0.016$; see Figure 4A).

DISCUSSION

The critical finding of this study was that expectations carried in the visual stream regarding the loudness of a physical event (a clap) significantly influenced the

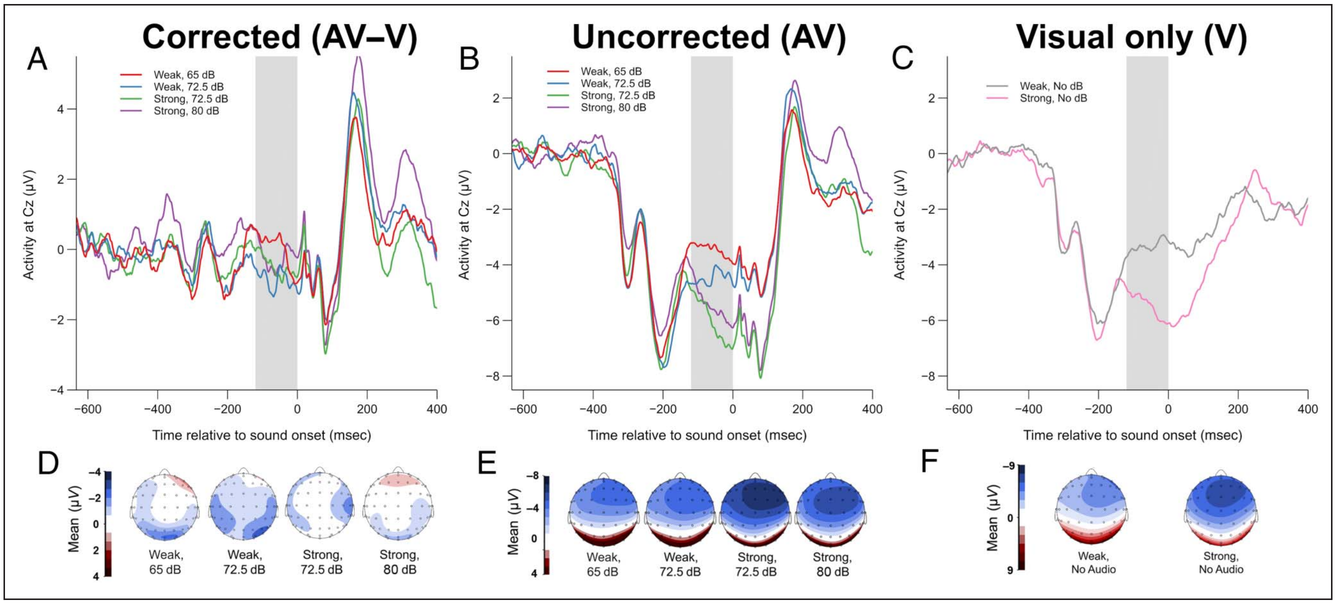

Figure 4. Differences between the AV–V, AV, and V-only waveforms in the anticipatory (prestimulus) period. (A) Grand-averaged waveforms for the corrected (AV–V) condition. (B) Grand-averaged waveforms for the uncorrected (AV) condition. (C) Grand-averaged waveforms for the V-only condition. The 0 msec point on the x axis represents the onset of the auditory stimulus (which was silent in the V-only condition). All waveforms recorded from electrode Cz, baselined to −633 to −433 msec prestimulus. (D), (E), and (F) represent the scalp topographies for the AV–V, AV, and V-only conditions, respectively. The time window used for the scalp topographies and statistical analysis was −120 to 0 msec.

electrophysiological response to the associated auditory stimulus. The key comparison was between the "weak" and "strong" clap video clips when both were paired with the same clap sound level of 72.5 dB. The results revealed that the amplitude of the auditory N1 component was significantly larger when viewing the video depicting a "strong" clap compared with when viewing the video depicting a "weak" clap. This result was observed despite the fact that visual-evoked activity has been subtracted out of the waveforms and the auditory stimulus was physically identical in both conditions (i.e., a 72.5-dB clap). Consistent with additive models of multisensory interactions (Besle, Fort, Delpuech, & Giard, 2004; Besle, Fort, & Giard, 2004; Giard & Peronnet, 1999), this result indicates that the visual characteristics of an auditory event interact with auditory processing. Given the low latency of the N1 component (<100 msec) and the fact the N1 is generated in the primary auditory cortex, this result suggests the multisensory interaction occurs at very early stages and in primary sensory regions. Furthermore, the direction of this effect favors our hypothesis; that the response of the auditory cortex generated from the received auditory signal (as indexed by the amplitude of the N1 component) is shifted toward the response that would be generated from the expected auditory signal.

Previously, it has been suggested that predictive cues are a mechanism for the reduction of uncertainty, thereby leading the brain to process a signal more efficiently and attenuating the N1 component (Hughes, Desantis, & Waszak, 2013; Besle, Fort, Delpuech, et al., 2004; Schafer & Marcus, 1973). According to predictive coding theories, the brain is an active inference machine, and making accurate predictions aids the brain in dealing with uncertainty and inferring the most likely cause of a signal (Clark, 2013; Friston, 2005). By this account, if a signal received in a bottom–up cortical area deviates from the top–down prediction, a prediction error will occur, which will increase the evoked neural response (van Laarhoven, Stekelenburg, & Vroomen, 2017; Sanmiguel, Saupe, & Schröger, 2013; Arnal & Giraud, 2012; Friston, 2009). However, in this study, it is unlikely that the observed differences between the "weak" and "strong" clap visuals when paired with a clap sound level of 72.5 dB were due to differences in uncertainty or prediction errors. The reason for this is twofold. First, temporal cues were held constant with both visual cues being edited to have the same amount of temporal predictability (i.e., both videos had 450 msec of anticipatory motion; note, however, that our assumption of equivalent temporal predictability is questionable, as we discuss further below). Second, the degree of sound-level uncertainty did not differ between different conditions: A weak visual cue had a 50% chance of producing either a 65- or 72.5-dB clap sound and a strong visual cue had a 50% chance of producing either a 72.5- or 80-dB clap sound. Factors beyond uncertainty that have also been found to modulate the N1 component are selective attention and variable ISIs (Näätänen & Picton, 1987). However, it is unlikely that differences in attention drove the effect found in this study as both clap visuals had equal task relevancy, and there was no significant difference in the evoked response in the N1 time window between the weak and strong visual control conditions (i.e., when the video was relayed silently). Furthermore, the ISIs did not differ between conditions. Thus, the differences in N1 amplitude we observed between the weak and strong visuals at 72.5 dB were unlikely to have been driven by extraneous factors (temporal predictability, volume uncertainty, ISI) but were instead driven by participants' visually based expectations about the sound influencing their neurophysiological response to the sound.

Hence, our study has provided a unique view on how predictive visual information may bias the neural response to an auditory event. Without employing differences in predictive accuracy, we have shown that the visual characteristics predicting the nature of an auditory event can bias the response of the auditory cortex to that event. Specifically, we have shown that auditory processing may be modulated by "seeing volume." There has been a great deal of work on quantifying the elements that influence loudness judgments. The intensity of an auditory signal is the most dominant determinant of loudness judgments (Stevens, 1955), yet frequency (Melara & Marks, 1990), timbre (Allen, 1971), duration (Miskolczy-Fodor, 1959), and reverberation (Zahorik & Wightman, 2001) have also been shown to influence perceived loudness. It is worth speculating that if the N1 amplitude modulation in this study is correlated with perceived intensity, our perception of loudness may be a synthesis of both auditory and visual information. Support for the notion that the amplitude of the N1 component reflects perceived loudness comes from the fact that the N1 component is extremely sensitive to the intensity of the received auditory signal (Mulert et al., 2005; Brocke et al., 2000; Dierks et al., 1999; Hegerl et al., 1994; Rapin et al., 1966). Furthermore, there is some behavioral evidence that the modulation of the auditory N1 is correlated with loudness estimates (Roussel, Hughes, & Waszak, 2014; Sato, 2008). If this is the case, "seeing volume" may be a new factor in which loudness may be modulated. This would suggest that perception of auditory sound level is not only influenced by the direct properties of the auditory signal but also by the perceived relationship between the signal and visually causative events. This hypothesis is somewhat supported by the McGurk effect (McGurk & MacDonald, 1976) in which the brain fuses information from the auditory and visual streams to resolve similar but ambiguous stimuli. The difference in this study is that the merging of sensory stream information is not in relation to signal identity but rather to signal intensity.

We have argued that our finding that the 72.5-dB clap elicited a smaller N1 amplitude when paired with the "strong-video" (relative to the "weak-video") was due to the "strong-video" generating an expectation of a more intense sound. However, as flagged previously, an alternative "low-level" explanation is that the onset of the clap was more temporally predictable in the "weak-video," as the actors' hands were moving slower. The fact that increasing the temporal predictability of a sound has been shown to decrease the N1 amplitude that it elicits (Oestreich et al., 2015; Lange, 2011) is consistent with this explanation. Although disentangling the factors of temporal predictability and intensity expectations is challenging in the current paradigm without reducing the ecological validity of the stimuli, future studies could test the alternative hypothesis by, for example, replacing the clap videos with a slow-moving versus fast-moving bar. Such research would clarify the cause of the N1 differences observed in the current study.

To our knowledge, there have only been two studies that have investigated the behavioral consequences of visually created intensity expectations on loudness judgments. Using both speech and nonspeech stimuli (clapping), Rosenblum and Fowler (1991) required people to rate the amount of perceived effort put into the generation of a sound and, second, to rate the loudness of a sound when paired with the same visual. When auditory stimuli were paired with a video of a sound emitter that was perceived to be putting in more effort, perceived loudness ratings also increased. Aylor and Marks (1976) required participants to judge the relative loudness of narrowband noise transmitted through different barriers (row of hemlock trees, slat fence, acoustic tile barrier, or no barrier). Here, participants carried out two conditions, one blindfolded and one where the barrier obscuring the sound source was visible. In the blindfolded condition, there were no differences between loudness estimates for any of the barriers. In the condition in which participants had no blindfold, loudness ratings were relatively attenuated when the barriers did not completely visually obstruct the sound source (slat fence, no barrier). On the basis of this finding, it was suggested that when a sound source was occluded by a barrier, participants expected the barrier to diminish the loudness of the auditory stimulus, which in turn raised their loudness estimates. To confirm the relationship between the neural coding of auditory intensity and subjective loudness, future studies are needed to follow up whether psychophysical loudness judgments are influenced by visual cues containing information about the intensity of auditory stimuli.

The finding of this study may have interesting implications for certain populations. First, there is evidence that aging populations undergo hearing loss (Fozard, 1990), aging populations (with and without hearing impairment) demonstrate altered N1 responses compared with younger participants (Tremblay, Piskosz, & Souza, 2003), and furthermore that aging populations demonstrate enhanced multisensory integration (Laurienti, Burdette, Maldjian, & Wallace, 2006). It would be possible that the influence of visual cues on auditory processing may be exaggerated in an aging population who are more reliant on visual cues to support auditory information. Furthermore, the volume dependency of the N1 has been shown to be modulated in part by the serotonergic system, which has implicated its relevance in clinical disorders that are related to serotonin dysregulation (Hegerl, Gallinat, & Juckel, 2001). Low central serotonergic transmission is related to low loudness dependence of the auditory evoked signal, and high serotonergic transmission has been related to high loudness dependence of the auditory evoked signal (Hegerl et al., 2001). It would be interesting to investigate whether serotonin dysregulation extends beyond impacting the "loudness dependence" of the N1 component to impacting the modulation of the N1 based on expected loudness. Finally, people with schizophrenia have been specifically implicated as having deficits in auditory perception (Matthews et al., 2013) and an abnormal auditory N1 response when comparing the difference between audiovisual stimuli (with predictive visual cues) and auditory only stimuli (Stekelenburg, Maes, Van Gool, Sitskoorn, & Vroomen, 2013). As a consequence, people with schizophrenia who exhibit functional differences in the formation and adaption of top–down inferences may demonstrate an N1 response that is less influenced by intensity expectations.

A limitation of the current study is that we could not determine the boundaries of when predictive information about the intensity of a sound may not influence the elicited neural response. This means we cannot determine whether this effect is limited to ecologically valid visual stimuli or whether there are certain generalizable properties of visual collisions that can be distilled to create intensity expectations within abstract cues and finally whether one can learn to associate an irrelevant cue with an intensity expectation. Nonetheless, this study has demonstrated that visually based intensity expectations regarding auditory intensity may bias the amplitude of the N1. Future studies will need to determine the boundaries of this effect.

Finally, in a post hoc analysis, we unexpectedly identified a difference in anticipatory activity between the "strong" and "weak" clap videos: The "strong clap" videos elicited a negative-going potential with a frontal topography from approximately 120 msec preclap. Although we are agnostic as to the underlying cause of the negative going deflection, one possibility is that it reflects a stimulus preceding negativity (SPN; see Brunia, Hackley, van Boxtel, Kotani, & Ohgami, 2011; van Boxtel & Bocker, 2004). The SPN is a slow, negative-going potential that is elicited in anticipation of affective stimuli. Although the most common stimuli used to elicit the SPN are those which generate a significant affective or physiological response (e.g., electric shocks, pictures of opposite-sex nudes), it is possible that the anticipation of a loud (aversive) sound in response to the "strong clap" video could have been sufficient to elicit an SPN. Regardless of the identity of this component, it is important to note that it is unlikely to be responsible for the between-condition differences in N1 amplitude we observed (i.e., the primary result of this study). That is, we subtracted out the activity elicited in the V-only condition from the AV condition to create the corrected-audiovisual waveform (AV–V), as is common in studies of this nature (Stekelenburg & Vroomen, 2007). As illustrated in Figure 2A, there were no systematic differences in prestimulus activity between the (corrected) "strong clap" and "weak clap" videos at 72.5 dB. This result suggests that the observed between-condition differences in N1 amplitude between conditions were not the result of differences in prestimulus activity between the "strong" and "weak" clap videos.

In conclusion, this study has shown that the early evoked neurophysiological response to an auditory stimulus is dependent not only on the intensity of the stimulus but also on one's expectations regarding the intensity of the stimulus.

Acknowledgments

This research was supported in part by grants from the National Health and Medical Research Council of Australia (APP1090507 to T. W.) and the Australian Research Council (DP140104394 and DP170103094 to T. W.). Sol Libesman was supported by an Australian Government Research Training Program Scholarship.

Reprint requests should be sent to Thomas Whitford, School of Psychology Room 138, Mathews Building Entry Via Gate 11 Botany Street RANDWICK, UNSW Sydney, Sydney, NSW 2052, Australia or via e-mail: t.whitford@unsw.edu.au.

REFERENCES

Allen, G. D. (1971). Acoustic level and vocal effort as cues for the loudness of speech. *Journal of the Acoustical Society of America*, *49*, 1831–1841.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*, 390–398.

Aylor, D. E., & Marks, L. E. (1976). Perception of noise transmitted through barriers. *Journal of the Acoustical Society of America*, *59*, 397–400.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*, 2225–2234.

Besle, J., Fort, A., & Giard, M. H. (2004). Interest and validity of the additive model in electrophysiological studies of multisensory interactions. *Cognitive Processing*, *5*, 189–192.

Brocke, B., Beauducel, A., John, R., Debener, S., & Heilemann, H. (2000). Sensation seeking and affective disorders: Characteristics in the intensity dependence of acoustic evoked potentials. *Neuropsychobiology*, *41*, 24–30.

Brunia, C. H., Hackley, S. A., van Boxtel, G. J., Kotani, Y., & Ohgami, Y. (2011). Waiting to perceive: Reward or punishment? *Clinical Neurophysiology*, *122*, 858–868.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to loftus and massons method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45.

Dierks, T., Barta, S., Demisch, L., Schmeck, K., Englert, E., Kewitz, A., et al. (1999). Intensity dependence of auditory evoked potentials (AEPs) as biological marker for cerebral serotonin levels: Effects of tryptophan depletion in healthy subjects. *Psychopharmacology*, *146*, 101–107.

Fozard, J. L. (1990). Vision and hearing in aging. In *Handbook of the psychology of aging* (Vol. 3, pp. 143–156). San Diego, CA: Academic Press Limited.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *360*, 815–836.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in cognitive sciences*, *13*, 293–301.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*, 473–490.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*, 468–484.

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, *28*, 240–244.

Hegerl, U., Gallinat, J., & Juckel, G. (2001). Event-related potentials: Do they reflect central serotonergic neurotransmission and do they predict clinical response to serotonin agonists? *Journal of Affective Disorders*, *62*, 93–100.

Hegerl, U., Gallinat, J., & Mrowinski, D. (1994). Intensity dependence of auditory evoked dipole source activity. *International Journal of Psychophysiology*, *17*, 1–13.

Helmholtz, V. H. (1877). *On the sensations of tone*. In A. J. Ellis (English translation, 1885, 1954). New York: Dover Publications Inc.

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, *139*, 133–151.

Lange, K. (2011). The reduced N1 to self-generated tones: An effect of temporal predictability? *Psychophysiology*, *48*, 1088–1095.

Laurienti, P. J., Burdette, J. H., Maldjian, J. A., & Wallace, M. T. (2006). Enhanced multisensory integration in older adults. *Neurobiology of Aging*, *27*, 1155–1163.

Luck, S. J. (2005). *An introduction to the event-related potential technique* (pp. 45–64). Cambridge, MA: MIT Press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*, 146–157.

Matthews, N., Todd, J., Mannion, D. J., Finnigan, S., Catts, S., & Michie, P. T. (2013). Impaired processing of binaural temporal cues to auditory scene analysis in schizophrenia. *Schizophrenia Research*, *146*, 344–348.

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

Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception & Psychophysics*, *48*, 169–178.

Miskolczy-Fodor, F. (1959). Relation between loudness and duration of tonal pulses. I. Response of normal ears to pure tones longer than click-pitch threshold. *Journal of the Acoustical Society of America*, *31*, 1128–1134.

Mulert, C., Jäger, L., Propp, S., Karch, S., Störmann, S., Pogarell, O., et al. (2005). Sound level dependence of the primary auditory cortex: Simultaneous measurement with 61-channel EEG and fmri. *Neuroimage*, *28*, 49–58.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, *24*, 375–425.

Oestreich, L. K., Mifsud, N. G., Ford, J. M., Roach, B. J., Mathalon, D. H., & Whitford, T. J. (2015). Subnormal sensory attenuation to self-generated speech in schizotypy: Electrophysiological evidence for a continuum of psychosis. *International Journal of Psychophysiology*, *97*, 131–138.

Pantev, C., Bertrand, O., Eulitz, C., Verkindt, C., Hampson, S., Schuierer, G., et al. (1995). Specific tonotopic organizations of different areas of the human auditory cortex revealed by simultaneous magnetic and electric recordings. *Electroencephalography and Clinical Neurophysiology*, *94*, 26–40.

Peirce, J. W. (2007). PsychoPy–Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13.

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10.

Rapin, I., Schimmel, H., Tourk, L. M., Krasnegor, N. A., & Pollak, C. (1966). Evoked responses to clicks and tones of varying intensity in waking adults. *Electroencephalography and Clinical Neurophysiology*, *21*, 335–344.

Rosenblum, L. D., & Fowler, C. A. (1991). Audiovisual investigation of the loudness-effort effect for speech and nonspeech events. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 976–985.

Roussel, C., Hughes, G., & Waszak, F. (2014). Action prediction modulates both neurophysiological and psychophysical indices of sensory attenuation. *Frontiers in Human Neuroscience*, *8*, 115.

Sanmiguel, I., Saupe, K., & Schröger, E. (2013). I know what is missing here: Electrophysiological prediction error signals elicited by omissions of predicted “what” but not “when”. *Frontiers in Human Neuroscience*, *7*, 407.

Sato, A. (2008). Action observation modulates auditory perception of the consequence of others actions. *Consciousness and Cognition*, *17*, 1219–1227.

Satongar, D., Lam, Y., & Pike, C. (2014). The Salford BBC spatially-sampled binaural room impulse response dataset [data file] UK: University of Salford [distributor]. Retrieved from http://usir.salford.ac.uk/id/eprint/30868/.

Schafer, E. W., & Marcus, M. M. (1973). Self-stimulation alters human sensory brain responses. *Science*, *181*, 175–177.

Stekelenburg, J. J., Maes, J. P., Van Gool, A. R., Sitskoorn, M., & Vroomen, J. (2013). Deficient multisensory integration in schizophrenia: An event-related potential study. *Schizophrenia Research*, *147*, 253–261.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*, 1964–1973.

Stevens, S. S. (1955). The measurement of loudness. *The Journal of the Acoustical Society of America*, *27*, 815–829.

Tremblay, K. L., Piskosz, M., & Souza, P. (2003). Effects of age and age-related hearing loss on the neural representation of speech cues. *Clinical Neurophysiology*, *114*, 1332–1343.

van Boxtel, G. J., & Bocker, K. B. E. (2004). Cortical measures of anticipation. *Journal of Psychophysiology*, *187*, 61–76.

van Laarhoven, T., Stekelenburg, J. J., & Vroomen, J. (2017). Temporal and identity prediction in visual-auditory events: Electrophysiological evidence from stimulus omissions. *Brain Research*, *1661*, 79–87.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, *22*, 1583–1596.

Zahorik, P., & Wightman, F. L. (2001). Loudness constancy with varying sound source distance. *Nature Neuroscience*, *4*, 78–83.