Published in final edited form as:
Exp Brain Res. 2006 January ; 168(1-2): 1–10. doi:10.1007/s00221-005-0071-5.

Seeing speech affects acoustic information processing in the human brainstem

Gabriella A. E. Musacchia[1,*], **Mikko Sams**[4], **Trent G. Nicol**[1], and **Nina Kraus**[1,2,3]

[1] *Auditory Neuroscience Laboratory, Departments of Communication Sciences, Northwestern University, Evanston, IL, USA* [2] *Departments of Neurobiology and Physiology, Northwestern University, Evanston, IL, USA* [3] *Department of Otolaryngology, Northwestern University, Evanston, IL, USA* [4] *Laboratory of Computational Engineering, Helsinki University of Technology, Finland*

Abstract

Afferent auditory processing in the human brainstem is generally assumed to be determined by acoustic stimulus features and immune to stimulation by other senses or cognitive factors. In contrast, we show that lipreading during speech perception influences acoustic processing astonishingly early. Event-related brainstem potentials were recorded from 10 healthy adults to concordant (acoustic-visual match), conflicting (acoustic-visual mismatch) and unimodal stimuli. Audiovisual interactions occurred around 11ms post-stimulation and persisted for the first 30ms of the response. Furthermore, response timing and magnitude depended on audiovisual pairings. These findings indicate that early auditory processing is more plastic than previously thought.

Keywords

auditory; brainstem; multisensory; visual; speech

Natural perceptions are rich with sensations from the auditory and visual modalities (Marks, 1982). As a friend says hello, we are cheered by their friendly tone and the sight of their smile. At a concert, we are amazed at the sight and sound of a trumpet player's technique. Real-world audiovisual experiences do not evoke distinct "multisensory" sensations. Although the combination of acoustic and visual information goes seamlessly unnoticed by the perceiver, it has a strikingly potent effect on perception (Marks, 2004). One of the most prevalent models of audiovisual integration posits that information from different modalities is processed in a hierarchical fashion along unisensory streams, which converge in higher-order structures (e.g., Massaro, 1998). The combined representation is then processed in a feed-forward fashion that does not affect doenstream processing. While this hypothesis has proven to account for copious multisensory phenomena, evidence of audiovisual interaction in lower-order structures encourages modification of the model. These observations have prompted a growing number of scientists to investigate the underlying neural mechanisms of audiovisual integration, or how, where and when in the brain we bind one sensation to another.

*Corresponding author Gabriella A.E. Musacchia, Email: g-musacchia@northwestern.edu, Address: Department of Communication Sciences; Northwestern University, 2240 Campus Dr., Evanston, IL 60208, USA, Tel: 847.491.2465, Fax: 847.491.2523, Mikko Sams, Email: Mikko.Sams@hut.fi, Address: Laboratory of Computational Engineering, Helsinki University of Technology, Tekniikantie 14, Espoo (Innopoli II), P.O. Box 9203, FIN-02015 HUT, Finland, Tel: +358 9 451 4848, Fax: +358 9 451 4830, Trent G. Nicol, Email: tgn@northwestern.edu, Address: Department of Communication Sciences; Northwestern University, 2240 Campus Dr., Evanston, IL 60208, USA, Tel: 847.467.1227, Fax: 847.491.2523, Nina Kraus, Email: nkraus@northwestern.edu, Address: Departments of Communication Sciences, Neurobiology and Physiology, and Otolaryngology; Northwestern University, 2240 Campus Dr., Evanston, IL 60208, USA, Tel: 847.491.3181, Fax: 847.491.2523.

Many phenomena have shown that visual information has a remarkable influence on acoustic perception (Spence and McDonald, 2004). When acoustic and visual cues occur in close temporal and spatial proximity, their combined information enhances orientation (Zambarbieri, 2002), detection (McDonald et al., 2000), classification (Ben Artzi and Marks, 1995), and reaction time (McDonald and Ward, 2000). Circumstances in which the acoustic and visual stimuli are not matched can cause a variety of illusions including ventriloquism (Howard and Templeton, 1966), and fused perceptions that do not correspond to either unimodal stimulus (McGurk and MacDonald, 1976;Massaro, 1998).

One of the most ubiquitous and well-studied examples of audiovisual integration in humans is seeing and hearing speech. Both the motor theory of speech (Liberman and Mattingly, 1985) and more recent investigations (Rizzolatti and Craighero, 2004) encourage the idea that the acoustic properties of speech are intrinsically linked with the articulation pattern of the visuofacial movement. For infants, visuofacial information is commandeered to aid speech acquisition (Kent, 1984). For both children and adults, watching concordant articulatory gestures improves the perception of artificially degraded speech (e.g., Grant, 2001) and speech in noise (MacLeod and Summerfield, 1987;Green, 1987;Middelweerd and Plomp, 1987). Observing facial movements that do not match the acoustic speech can drastically change what people “hear”, even when the acoustic signal is clear. For example, when subjects see a talker say /ga/ and are simultaneously presented with an acoustic /ba/, they typically hear /da/ (MacDonald and McGurk, 1978). Although these striking perceptual illusions were once thought to add yet another example to the pool of data supporting the “special” nature of speech, analogous illusions have been reported with musical stimuli (Saldana and Rosenblum, 1993) and non-native speech (Sekiyama and Tohkura, 1991;Sekiyama et al., 2003).

A growing body of data shows that integration of acoustic speech and facial movements share cortical and subcortical sites of convergence with nonspeech stimuli (for review, see Calvert, 2001). These areas include portions of the superior temporal sulcus (Giard and Peronnet, 1999), intraparietal (Callan et al., 2001;Calvert et al., 2000), prefrontal and premotor (Bushara et al., 2001) cortices. Audiovisual centers in the superior colliculus (SC), a structure that engages mechanisms of gaze control and attention, have been shown to integrate spatial and temporal information from nonspeech acoustic and visual cues (Rauschecker and Harris, 1989;Stein et al., 2002), though human data in this area are sparse.

Studies in the monkey and cat model showed that the response patterns of audiovisual neurons in the SC determine SC-mediated orienting behavior (for review, see Stein et al., 2002). The probability of a gaze shift is increased when spatially or temporally concordant stimuli produce response enhancement, relative to the unimodal responses, in the SC. Conversely, incompatible stimuli produce response depression and a decreased probability of behavioral response. The human SC has also been shown to exhibit increased activity, compared to the sum of the unimodal responses, to temporally and spatially concordant audiovisual stimuli (Calvert et al., 2000). A different type of response pattern was evident, however, for spatial proximity in the barn owl homologue of the SC (Hyde and Knudsen, 2002). In the barn owl, spatially concordant audiovisual stimuli elicited a smaller amplitude of response than that to a stimulus whose sight and sound location did not match. In both mammal and owl species, separate lines of evidence have concluded that audiovisual integration in the SC is governed by cortical activity (Jiang and Stein, 2003;Gutfreund et al., 2002).

The excellent temporal resolution of event-related potentials (ERP) has been utilized to investigate whether sites of audiovisual convergence in humans are activated early or late in the processing stream. Because integration is by nature a complex phenomenon, differences in results due to various stimulus features, tasks, and experimental conditions have precluded comprehensive understanding of when audiovisual interaction first occurs. Evidence for

"early", commonly accepted to be <200ms, and "late" audiovisual integration has been shown in both sensory-specific (Sams et al., 1991a;Mottonen et al., 2002) and higher-order cortices (Sams et al., 1991b). In studies using nonspeech stimuli, ERPs to audiovisual stimuli were found to be distinct from those to unimodal acoustic and visual tokens as early as 90ms over primary auditory cortex (Giard and Peronnet, 1999) and at 40 to 50ms over the nonprimary visual areas (Molholm et al., 2002). Sams and colleagues showed that visual speech modulated information processing in the auditory cortex at 180ms post-acoustic stimulation (Sams et al., 1991a). An emerging hypothesis from the ERP data of speech and nonspeech studies is that early audiovisual integration in primary auditory cortices is related to the interaction of non-phonetic, or "what", features that are shared by the stimuli (i.e., coincidence in time) and later interaction in heteromodal areas is more related to phonetic, or "where/how", feature discrimination (Klucharev et al., 2003).

Despite considerable data, the debate over whether audiovisual stimuli are processed in a strictly feed-forward manner that begins in higher-order structures has not been resolved (for review, see Giard and Fort, 2004). Because the precise time course of interaction in subcortical structures is not known we cannot tell whether audiovisual interaction occurs before (i.e., in early afferent processing) or after cortical processing and is a result of corticofugal modulation.

The robust features of the human brainstem response make it an excellent tool to investigate neural timing differences between unimodal and audiovisual stimuli in lower-order structures. The auditory brainstem response to an acoustic click has been used to diagnose sensorineural hearing impairment and neurologic disease since the seventies. The use of this measure is particularly effective because the brainstem response is unaffected by cognitive state changes and can be recorded reliably during sleep (for review, see Jacobson, 1991). Over the past five years, the brainstem response to speech has been used to describe the neural encoding of complex sounds in humans (King et al., 2002;Wible et al., 2004;Russo et al., 2004), and animal models (Wible et al., In Press). Responses elicited from both types of stimuli consist of a series of event-related peaks of activity. The successive peaks have been associated with the hierarchical activation of specific brainstem nuclei, such that earlier peaks represent the activity of more peripheral structures (e.g., the superior olivary complex) and later peaks represent the interaction of peripheral and more central populations (e.g., the lateral lemiscus and inferior coliculi) (Hall, 1992).

The complex characteristics of the speech-evoked brainstem response comprise precise transient and periodic features that can be quantified in terms of timing and magnitude measures. Peak-latency measures give information about when the response culminates in time, and the degree of neural synchrony. The overall strength of the neural response, as well as spectral encoding, can be quantified with amplitude measures and fast Fourier transform analysis. The precise and replicable nature of the brainstem response enables effective comparisons between individual responses and normative values (Russo et al., 2004). Peak-latency differences to click stimuli as small as tenths of a millisecond can be diagnostically significant in individuals with audiologic or neurologic abnormalities (Moller, 1999;Moller, 2000) and can distinguish normal and language-learning impaired groups using speech (Cunningham et al., 2001;King et al., 2002;Hayes et al., 2003).

Cross-modal interactions in the brainstem can be identified using the dominant paradigm employed in cortical studies. In typical audiovisual experiments, the cortical response to each of the two unimodal stimuli presented alone is compared with the response to the audiovisual presentation (e.g., Schroger and Widmann, 1998;Fort et al., 2002). This paradigm enables two data analysis strategies. Interaction effects can be identified by differences between the audiovisual (AV) and responses to unimodal acoustic (UA) and visual (UV) cues presented in isolation (i.e., AV≠UA and AV≠UV). In addition, audiovisual response features that deviate

from the mathematical combination of the unimodal responses (i.e., AV≠UA+UV) may be considered as evidence of true, nonlinear, audiovisual interaction mechanisms.

The principal aim of this study was to test whether visual articulatory gestures, or lip-reading influences the subcortical response to acoustic speech. The working hypothesis was that acoustic and visual speech would interact in low-order structures, by mechanisms of interaction commonly observed in the animal SC. If the hypothesis was true, AV interaction would be evident in speech-evoked brainstem event-related potentials, the timing of which could help resolve the extent to which AV mechanisms operate early or late in the processing stream.

Materials and Methods

Event-related brainstem responses to audiovisual (AV), Unimodal Visual (UV) and Unimodal Acoustic (UA) speech stimuli were recorded from ten adults aged 18–35 with hearing thresholds better than 20dB HL and 20/20 corrected vision (Logarithmic Visual Acuity Chart "2000", Precision Vision).

Visual stimuli were created from a digital recording of a male speaker articulating /da/, /du/ and /fu/ utterances. All three articulations were edited to 19 frames (FinalCut Pro 4) that began and ended at the same neutral resting position (MorphMan 4.0). The release of the consonant was edited to occur at frame 11 for all three visual tokens. A 100ms synthetic speech phoneme, /da/, was created with a DH Klatt synthesizer. When presented together, speech onset occurred with place of articulation at frame 11 for all AV tokens (Figure 1).

Stimulus sequences were delivered with Presentation software (Neurobehavioral Systems, Inc., 2001). The testing session consisted of three stimulus conditions. Subjects were presented with sequences of UA, AV and UA tokens separately. A five-minute break was given between each condition. In all three conditions, five blocks of 200 tokens were presented with a two-minute break between blocks. In the UV condition, each block consisted of randomly ordered /da/ (40%), /fu/ (40%) and /du/ (20%) visual utterances. To maintain visual fixation and attention, subjects were asked to attend to the video and silently count the number of /du/ tokens. Visual tokens were projected onto a 38" × 48" screen with subjects seated 84" from the screen. Each frame was presented for 33.2 ms +/− 1.2ms. In the UA stimulus sequence, the projector was turned off and acoustic stimuli were presented at 84dB binaurally through comfortable ear inserts. Subjects were asked to count how many blocks of 50 /da/ tokens they heard. In the AV stimulus sequence, the synthesized speech syllable /da/ was paired with the /da/ ($AV_{Concordant,}$ 40%), /fu/ ($AV_{Conflicting}$, 40%) and /du/ (20%) visual utterances. As in the UV condition, AV tokens were presented randomly in the five blocks and subjects counted the number of /du/ tokens.

Continuous ectroencephalographic (EEG) activity was acquired with Neuroscan 4.3 from Cz (impedance < 5 kΩ), referenced to the nose, band pass filtered from 0.05 to 3000Hz and digitized at 20000Hz. The first frame of each visual token triggered a mark in the EEG file. After collection, the EEG was further band pass filtered from 75 to 2000 Hz to select the brainstem response frequencies (Hall, 1992). The continuous EEG was divided into epochs (20ms pre- to 120ms post-acoustic onset), and baseline corrected over the pre-stimulus interval. An artifact criterion of > +/− 65mV was used to reject epochs that contained eye movement or myogenic artifacts. The epoched files were combined into 1000-sweep averages in the $AV_{Concordant}$, $AV_{Conflicting}$ and Unimodal conditions. The rare /du/ responses in the UV and AV conditions were not analyzed. Waves V, γ, ε, and κ (Figure 2) were picked by visual inspection for all subjects, in all conditions, and were reviewed by a second investigator.

Results

The grand average neural responses to the three unimodal conditions (UA /da/, UV /da/ and UV /fu/) are illustrated in Figure 2. Visual inspection of the individual and grand average UA waveforms showed similar morphology across subjects. The onset of the acoustic stimulus elicited transient, biphasic peaks. The vowel portion of the stimulus evoked the periodic portion of response, called the frequency following response (FFR), in which time between peaks reflects the wavelengths of the frequencies present in the stimulus (Marsh et al., 1975;Galbraith et al., 1995). The first prominent peak of the acoustic response, Wave V (mean latency 7.01ms, s.d. 0.33), had waveform morphology, latency, and across-subject reliability that was comparable to the well-established peaks in the click-evoked brainstem response and the onset response to acoustic speech reported in previous studies (Russo et al., 2004). In all subjects, and evident in the average, Wave V was followed by a negative trough and a positive peak that will be referred to as Wave γ (mean latency 10.38, s.d. 0.63). The periodic portion of the response (FFR) began with a positive peak, Wave ε (mean latency 29.53ms, s.d. 0.42) and ended at negative peak, Wave κ (mean latency 109.62ms, s.d. 2.60). Neither the /da/ or /fu/ UV responses elicited replicable peaks across subjects, indicating that the visual stimulus alone elicited little evoked activity with the recording parameters and electrode placement reported here.

Results I: Visuofacial movements delay the brainstem response to speech onset

The addition of either visual stimulus prolonged the latency of Wave γ, relative to the UA brainstem response to speech (Figure 3A). Repeated measures ANOVAs with three levels as within-subjects factor (UA, $AV_{Concordant}$ and $AV_{Conflicting}$) were conducted for Waves V, γ, ε, and κ. Latency differences across conditions were evident only for Wave γ ($F=9.36$, $p=0.002$). Subsequent protected paired t-tests (2-tailed) showed that Wave γ in both the $AV_{Concordant}$ (mean latency 11.70ms; $t=3.88$, $p=0.001$) and $AV_{Conflicting}$ (mean latency 11.61ms; $t=3.60$, $p=0.002$) conditions were delayed relative to the UA response (Figure 3B). Wave γ latencies in the two AV conditions did not significantly differ.

Inter-peak intervals between Wave V and Wave γ ($\gamma_{latency}$-$V_{latency}$) were computed to determine if the delay in the AV condition occurred subsequent to Wave V. An ANOVA ($F=9.86$, $p=0.001$) and post-hoc tests as described above showed the inter-peak interval to be prolonged in the $AV_{Concordant}$ ($t=3.98$, $p=0.001$) and $AV_{Conflicting}$ ($t=3.69$, $p=0.002$) conditions when compared to the UA values. This finding, combined with the null result for Wave V latencies across conditions, confirmed that neural generators associated with Wave V (e.g., the lateral lemniscus and inferior colliculi) were impervious to visual influence and did not contribute to the audiovisual delay of Wave γ.

To further examine the audiovisual delay of Wave γ, cross-correlation measures were performed over a latency range that included Wave γ and its negative trough (8 to 20ms). This analysis technique shifts one waveform in time to obtain a maximal correlation value (Pearson's r). Correlation values and lags were subjected to single sample tests. A maximal correlation between UA and $AV_{Concordant}$ responses occurred with a lag of 0.69ms ($t=2.66$, $p=0.026$). The lag (0.36ms) between UA and $AV_{Conflicting}$ responses was not significantly different from zero.

The difference between the AV conditions and their computed UA+UV counterparts evinced a true nonlinear audiovisual interaction at Wave γ. A significant ANOVA ($F=9.29$, $p<0.001$) with four levels as within-subjects factor ($AV_{Concordant}$, $AV_{Conflicting}$, $UA+UV_{Concordant}$, $UA+UV_{Conflicting}$) and protected paired t-tests (2-tailed) revealed that both the $AV_{Concordant}$ ($t=3.64$, $p= 0.001$) and $AV_{Conflicting}$ ($t=3.79$, $p<0.001$) responses were delayed when compared to their respective unimodal sums. Concordant and conflicting Wave γ latencies were not

statistically different when compared in either the AV or UA+UV condition. Repeated measures ($F=9.28$, $p<0.001$) and protected t-tests showed that the delay at Wave γ persisted in both the $AV_{Concordant}$ ($t=3.64$, $p=0.001$) and $AV_{Conflicting}$ ($t=3.79$, $p<0.001$) compared to UA +UV conditions when the latencies were normalized to their UA value ($AV_{latency}$-$UA_{latency}$). The high degree of replicability across subjects that contributes to the diagnostic strength of the brainstem response can be brought to bear by the evaluation of individual latencies. Therefore, it is important to note that our data reflect a distribution of the extent of Wave γ delay across individuals (Table 1, Figure 4). Normalized UA+UV latencies are clustered around the UA latency (shown as a dashed line at 0), while AV values are visibly later in most of the subjects. The perceptual or subject characteristic correlates of the degree of early audiovisual interaction were not pursued in this study, but would be an intriguing direction of investigation.

A repeated measures ANOVA of Wave V to Wave γ inter-peak intervals with the same four levels described above ($F=9.29$, $p<0.001$) and protected paired t-tests (2-tailed) showed prolonged intervals in the $AV_{Concordant}$ ($t=3.64$, $p=0.001$) and $AV_{Conflicting}$ ($t=3.79$, $p<0.001$) conditions compared to their respective unimodal sums. Again, no differences were observed between concordant and conflicting conditions in either AV or UA+UV conditions. Careful investigation of the individual and grand average waveforms over the FFR period revealed no indication of timing differences in this region.

Results II: Two types of visual stimuli modulate the size of the acoustic brainstem response to speech differently

To assess the effects of visual speech on the amplitude of the acoustic response, rectified mean amplitude (RMA) was calculated. Individual subject latencies for Waves V, ε, and κ were used to describe the per-subject time ranges for RMA calculations. Onset RMAs were calculated between V and ε; FFR RMAs were calculated between ε and κ. There were notable AV effects in the onset RMA.

A repeated measures ANOVA ($F=5.82$, $p=0.011$) with three levels of within-subjects factor (UA, $AV_{Concordant}$ and $AV_{Conflicting}$) and subsequent protected t-tests showed the RMA of the onset response to be diminished in both the $AV_{Concordant}$ ($t=3.31$, $p=0.004$) and $AV_{Conflicting}$ ($t=2.37$, $p=0.029$) conditions compared to UA. In contrast to the onset timing finding in which both $AV_{Concordant}$ and $AV_{Conflicting}$ Wave γ latencies were delayed to the same degree, the size of the $AV_{Concordant}$ response (Mean RMA 0.186, s.d. 0.053) was diminished more than the $AV_{Conflicting}$ (Mean RMA 0.207, s.d. 0.060) compared to the UA response (Mean RMA 0.259, s.d. 0.109) (Figure 5). Protected paired t-tests between $AV_{Concordant}$ and $AV_{Conflicting}$ were significant for both RMA values ($p=0.026$) and RMA values normalized to the UA response size (AV_{RMA}-UA_{RMA}) ($p=0.023$).

Suppression was not observed in the UA+UV conditions. A three-level repeated measures ANOVA ($F=8.75$, $p=0.002$) and subsequent post-hoc protected paired t-tests showed that the UA+UV RMA in the concordant condition was larger than that of the UA ($t=4.11$, $p<0.001$). There was no difference between UA and UA+UV RMA values in the conflicting condition. Because the RMAs of the UA+UV responses were either the same or larger than the UA value, it was not surprising that the pattern of AV suppression, relative to the UA response, was also observed when AV RMAs were compared to the UA+UV responses. A repeated measures ANOVA with four levels of within-subject factor ($F=11.26$, $p<0.001$) and post-hoc paired t-tests confirmed that the AV onset RMAs were smaller than those in the UA+UV conditions ($t_{Concordant}=4.97$, $p<0.001$; $t_{Conflicting}=3.01$, $p=0.006$). The extent of the AV suppression over the onset response was not correlated with the length of the Wave γ delay. No evidence of AV interaction was observed over the FFR region of the responses.

Discussion

The current study demonstrates that seeing speech delayed the human brainstem response to speech as early as 11ms post-acoustic stimulation. Observation of both /da/ and /fu/ facial movements while listening to /da/ delayed the Wave γ latency by about 1.3ms relative to the UA response. A latency shift of this magnitude is striking in light of clinical criteria, which defines abnormal brainstem timing in tenths of milliseconds (Jacobson, 1991;Hall, 1992). The observed delay in the audiovisual conditions cannot be attributed to activity elicited by the visual stimuli alone, because no delay was observed with respect to the UA response when the acoustic and visual unimodal responses were simply added together. This early delay occurred when the acoustic stimulus was paired with either concordant or conflicting visual speech. The effect seems to be slightly more robust for matching acoustic and visual stimuli, based on cross-correlation statistics.

Additionally, the amplitude of the brainstem onset response was diminished in AV conditions and the extent of diminution depended on the type of visuofacial movement. Overall, both the $AV_{Concordant}$ and $AV_{Conflicting}$ RMA values were smaller, or suppressed, compared to the UA values. The $AV_{Concordant}$ RMA values were more suppressed than those to the $AV_{Conflicting}$ stimulus. This finding further supports greater AV interaction for the concordant stimulus. Amplitude suppression, relative to the UA condition, in the AV conditions could not be attributed to linear mechanisms of interaction because the computed sum of their respective unimodal responses did not show the same pattern of diminution. The addition of visual stimulation had no effect on the FFR region, which is thought to encode the spectral features of a complex sound.

These data show that early subcortical auditory processing is susceptible to visual influence. The observed differences between the latency of Wave γ elicited by UA and AV stimuli are, to our knowledge, the earliest reported audiovisual speech interaction. These latency differences take place before the earliest reported excitation from the primary auditory cortex, detected in direct intracranial recordings at 12–15ms post stimulation (Celesia, 1968). In light of this, it seems that the observed interaction must be taking place in afferent brainstem structures. All observed differences between AV, UA+UV and UA conditions occurred after ~7ms and up to ~30ms of the brainstem response. This indicates that neural encoding of the acoustic onset, in this case the consonant, was affected both in latency and amplitude by visual stimulation whereas the periodic portion of the speech stimulus, or vowel, was unaffected. Our findings corroborate a large body of data, which provide the premise for the existence of subcortical interaction mechanisms, the timing and extent of which have remained elusive until now.

A decrease in population synchrony and the alignment of audiovisual spatial maps in afferent brainstem structures could explain the observed Wave γ delay and response suppression, relative to the UA values, in the AV conditions. A fundamental property of event-related potentials is that a decrease in synchrony of firing, for example due to aggregate neural populations firing at slightly different times, results in longer peak latencies (Hall, 1992). Visual or audiovisual nuclei in the brainstem that do not fire in concert with those involved in UA processing could produce the observed delay. The excitation of different brainstem nuclei with opposite dipoles could also produce the observed cancellation, or suppression, of total electrical activity recorded from the surface of scalp. Although human data from the superior colliculus has been limited to nonspeech stimuli, acoustic and visual cues that coincide in time and space have been shown to produce enhancement, rather than the suppression seen here. It is possible that acoustic stimuli (presented with ear inserts) were encoded as spatially disparate from the visual tokens (projected in front of the subject). However, the observed difference between the RMA of the $AV_{Concordant}$ and $AV_{Conflicting}$ responses would be unexpected, given

that the spatial disparity would be equal across the two conditions. Response suppression, like that observed in the current study, has previously been shown to spatially concordant acoustic and visual cues in the optic tectum of the barn owl (Hyde and Knudsen, 2002). It is conceivable that the audiovisual response to our primary means of communication, speech, engages interaction mechanisms in humans more akin to those in the specialized structures of the barn owl.

An alternative explanation for the delayed Wave γ latency and reduced RMA in the AV conditions is that cortical attention mechanisms, generated by pre-articulatory visual movements, produced an overall modulation that affected afferent brainstem activity. Because visual movement preceded the acoustic onset by 360ms, there was ample time for corticofugal modulation of brainstem structures. The visual information preceding the sound might have helped focus attention to the onset of the acoustic stimulus, and such attentional influence could have modified the brainstem response. Previous work showing that subcortical processing of nonspeech stimuli can be modulated by visual attention and training supports this interpretation (Hernandez-Peon et al., 1956;Suga and Ma, 2003). Because the /da/ and /fu/ visual stimuli were different prior to acoustic onset, efferent activation of the brainstem could also explain the difference seen in the $AV_{Concordant}$ and $AV_{Conflicting}$ RMA values. However, the putative attentional effect would be limited to corticofugal modulation of more central brainstem structures because effects of visual influence were not found prior to Wave V.

The results of this study cannot clearly differentiate between speech and nonspeech effects because there were no nonspeech controls, however, because the stimuli were in fact speech tokens, we must discuss the implications of our findings in terms of both speech-specific and more generalized audiovisual interaction hypotheses.

One interpretation is that speech and imitation are so closely related by the motor system that articulatory gestures could influence afferent speech processing in a distinctively different way than nonspeech tokens. A long-debated question is whether speech is processed differently than nonspeech sounds (Chomsky, 1985;Hauser et al., 2002). Separate brain mechanisms have been shown to be active for acoustic speech and nonspeech processing (e.g., Tervaniemi and Hugdahl, 2004;Binder et al., 2000) and recent evidence revealed a strong relationship between phoneme perception and motor imitation (Gallese et al., 1996). A related interpretation of the current results is that extensive experience with audiovisual speech results in plasticity of the system such that visual articulatory gestures have unique access to the auditory brainstem. The above hypothesis would support the theory that speech is processed in a qualitatively different way from nonspeech, and posit that precursors of phonetic discrimination operate at the level of the brainstem to discern the degree of audiovisual concordance for later processing.

Alternatively, any visual cue that facilitates attention to acoustic stimulus onset, regardless of linguistic content, may modulate early auditory brainstem activity. Subtle differences in the pre-acoustic visual quality (such as that between /da/ and /fu/ visual facial movements) may have different effects on the response that are independent of their concordance, or lack thereof, to the accompanying sound.

Both interpretations challenge the prevailing view about the human brainstem as a passive receiver/transmitter of modality-specific information. Future investigations on the nature of early audiovisual interactions, and the subject characteristics that contribute to the presence or absence off these effects, will most likely have a great impact on our understanding of sensory processing. The results of the current study are reflections of a new zeitgeist in science today: that our neural system is an active information seeker that incorporates multisensory information at the earliest possible stage in order to discern meaningful objects from the world around it.

Acknowledgements

NIH R01 DC01510 supported this work. The authors wish to thank their collaborators in the Auditory Neuroscience Laboratory at Northwestern University and in the Laboratory of Computational Engineering at the University of Helsinki.

References

Ben Artzi E, Marks LE. Visual-auditory interaction in speeded classification: role of stimulus difference. Percept Psychophys 1995;57:1151–1162. [PubMed: 8539090]

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 2000;10:512–528. [PubMed: 10847601]

Bushara KO, Grafman J, Hallett M. Neural correlates of auditory-visual stimulus onset asynchrony detection. J Neurosci 2001;21:300–304. [PubMed: 11150347]

Callan DE, Callan AM, Kroos C, Vatikiotis-Bateson E. Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study. Brain Res Cogn Brain Res 2001;10:349–353. [PubMed: 11167060]

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

Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 2000;10:649–657. [PubMed: 10837246]

Celesia GG. Auditory evoked responses. Intracranial and extracranial average evoked responses. Arch Neurol 1968;19:430–437. [PubMed: 5677192]

Chomsky, N. The Logical Structure of Linguistic Theory. Chicago, IL: The University of Chicago Press; 1985.

Cunningham J, Nicol T, Zecker SG, Bradlow A, Kraus N. Neurobiologic responses to speech in noise in children with learning problems: deficits and strategies for improvement. Clin Neurophysiol 2001;112:758–767. [PubMed: 11336890]

Fort A, Delpuech C, Pernier J, Giard MH. Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. Cereb Cortex 2002;12:1031–1039. [PubMed: 12217966]

Galbraith GC, Arbagey PW, Branski R, Comerci N, Rector PM. Intelligible speech encoded in the human brain stem frequency-following response. Neuroreport 1995;6:2363–2367. [PubMed: 8747154]

Gallese V, Fadiga L, Fogassi L, Rizzolatti G. Action recognition in the premotor cortex. Brain 1996;119 (Pt 2):593–609. [PubMed: 8800951]

Giard, MH.; Fort, A. Multiple Electrophysiological Mechanisms of Audiovisual Integration in Human Perception. In: Calvert, GA.; Spence, C.; Stein, BE., editors. The Handbook of Multisensory Processes. Cambridge MA: MIT Press; 2004. p. 503-512.

Giard MH, Peronnet F. Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J Cogn Neurosci 1999;11:473–490. [PubMed: 10511637]

Grant KW. The effect of speechreading on masked detection thresholds for filtered speech. J Acoust Soc Am 2001;109:2272–2275. [PubMed: 11386581]

Green KP. The perception of speaking rate using visual information from a talker's face. Percept Psychophys 1987;42:587–593. [PubMed: 3696953]

Gutfreund Y, Zheng W, Knudsen EI. Gated visual input to the central auditory system. Science 2002;297:1556–1559. [PubMed: 12202831]

Hall, JWI. Handbook of Auditory Evoked Responses. Needham Heights, MA: Allyn and Bacon; 1992.

Hauser MD, Chomsky N, Fitch WT. The faculty of language: what is it, who has it, and how did it evolve? Science 2002;298:1569–1579. [PubMed: 12446899]

Hayes EA, Warrier CM, Nicol TG, Zecker SG, Kraus N. Neural plasticity following auditory training in children with learning problems. Clin Neurophysiol 2003;114:673–684. [PubMed: 12686276]

Hernandez-Peon R, Scherrer H, Jouvet M. Modification of electric activity in cochlear nucleus during attention in unanesthetized cats. Science 1956;123:331–332. [PubMed: 13298689]

Howard, I.; Templeton, WB. Human spatial orientation. New York: Wiley; 1966.

Hyde PS, Knudsen EI. The optic tectum controls visually guided adaptive plasticity in the owl's auditory space map. Nature 2002;415:73–76. [PubMed: 11780119]

Jacobson, J. The Auditory Brainstem Response. Prentice Hall; 1991.

Jiang W, Stein BE. Cortex controls multisensory depression in superior colliculus. J Neurophysiol 2003;90:2123–2135. [PubMed: 14534263]

Kent RD. Psychobiology of speech development: coemergence of language and a movement system. Am J Physiol 1984 Jun;246:R888–R894. [PubMed: 6742163]1984

King C, Warrier CM, Hayes E, Kraus N. Deficits in auditory brainstem pathway encoding of speech sounds in children with learning problems. Neurosci Lett 2002;319:111–115. [PubMed: 11825683]

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

Liberman AM, Mattingly IG. The motor theory of speech perception revised. Cognition 1985;21:1–36. [PubMed: 4075760]

MacDonald J, McGurk H. Visual influences on speech perception processes. Percept Psychophys 1978;24:253–257. [PubMed: 704285]

MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. Br J Audiol 1987;21:131–141. [PubMed: 3594015]

Marks LE. Bright sneezes and dark coughs, loud sunlight and soft moonlight. J Exp Psychol Hum Percept Perform 1982;8:177–193. [PubMed: 6461716]

Marks, LE. Cross-Modal Interactions in Speeded Classification. In: Calvert, GA.; Spence, C.; Stein, BE., editors. The Handbook of Mutisensory Processe. Cambridge, MA: MIT Press; 2004. p. 85-106.

Marsh JT, Brown WS, Smith JC. Far-field recorded frequency-following responses: correlates of low pitch auditory perception in humans. Electroencephalogr Clin Neurophysiol 1975;38:113–119. [PubMed: 45941]

Massaro, DW. Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, MA: MIT Press; 1998.

McDonald JJ, Teder-Salejarvi WA, Hillyard SA. Involuntary orienting to sound improves visual perception. Nature 2000;407:906–908. [PubMed: 11057669]

McDonald JJ, Ward LM. Involuntary listening aids seeing: evidence from human electrophysiology. Psychol Sci 2000;11:167–171. [PubMed: 11273425]

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

Middelweerd MJ, Plomp R. The effect of speechreading on the speech-reception threshold of sentences in noise. J Acoust Soc Am 1987;82:2145–2147. [PubMed: 3429736]

Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ. Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res Cogn Brain Res 2002;14:115–128. [PubMed: 12063135]

Møller AR. Neural mechanisms of BAEP. Electroencephalogr Clin Neurophysiol Suppl 1999;49:27–35. [PubMed: 10533081]

Møller AR. Diagnosis of acoustic tumors. Am J Otol 2000;21:151–152. [PubMed: 10651452]

Möttönen R, Krause CM, Tiippana K, Sams M. Processing of changes in visual speech in the human auditory cortex. Brain Res Cogn Brain Res 2002;13:417–425. [PubMed: 11919005]

Rauschecker JP, Harris LR. Auditory and visual neurons in the cat's superior colliculus selective for the direction of apparent motion stimuli. Brain Res 1989;490:56–63. [PubMed: 2758330]

Rizzolatti G, Craighero L. The Mirror-Neuron System. Annu Rev Neurosci 2004 Jul;27:169–192. [PubMed: 15217330]2004

Russo N, Nicol T, Musacchia G, Kraus N. Brainstem responses to speech syllables. Clin Neurophysiol 2004;115:2021–2030. [PubMed: 15294204]

Saldana HM, Rosenblum LD. Visual influences on auditory pluck and bow judgments. Percept Psychophys 1993;54:406–416. [PubMed: 8414899]

Sams M, Aulanko R, Hämäläinen M, Hari R, Lounasmaa OV, Lu ST, Simola J. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 1991a; 127:141–145. [PubMed: 1881611]

Sams M, Kaukoranta E, Hämäläinen M, Näätänen R. Cortical activity elicited by changes in auditory stimuli: different sources for the magnetic N100m and mismatch responses. Psychophysiology 1991b;28:21–29. [PubMed: 1886961]

Schroger E, Widmann A. Speeded responses to audiovisual signal changes result from bimodal integration. Psychophysiology 1998;35:755–759. [PubMed: 9844437]

Sekiyama K, Kanno I, Miura S, Sugita Y. Auditory-visual speech perception examined by fMRI and PET. Neurosci Res 2003;47:277–287. [PubMed: 14568109]

Sekiyama K, Tohkura Y. McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. J Acoust Soc Am 1991;90:1797–1805. [PubMed: 1960275]

Spence, C.; McDonald, JJ. The Cross-modal Consequences of the Exogenous Spatial Orienting of Attention. In: Calvert, GA.; Spence, C.; Stein, BE., editors. The Handbook of Multisensory Processes. Cambridge MA: MIT Press; 2004. p. 3-25.

Stein BE, Wallace MW, Stanford TR, Jiang W. Cortex governs multisensory integration in the midbrain. Neuroscientist 2002;8:306–314. [PubMed: 12194499]

Suga N, Ma X. Multiparametric corticofugal modulation and plasticity in the auditory system. Nat Rev Neurosci 2003;4:783–794. [PubMed: 14523378]

Tervaniemi M, Hugdahl K. Lateralization of auditory-cortex functions. Brain Res Brain Res Rev 2003;43:231–246. [PubMed: 14629926]

Wallace MT, Meredith MA, Stein BE. Multisensory integration in the superior colliculus of the alert cat. J Neurophysiol 1998;80:1006–1010. [PubMed: 9705489]

Wible, B.; Nicol, T.; Kraus, N. Encoding of complex sounds in an animal model: Implications for understanding speech perception in humans. Proceedings of the International Conference on Auditory Cortex—Toward a Synthesis of Human and Animal Research; Magdeburg, Germany. September 2003; in press

Wible B, Nicol T, Kraus N. Atypical brainstem representation of onset and formant structure of speech sounds in children with language-based learning problems. Biological Psychology 2004;67:299–317. [PubMed: 15294388]

Zambarbieri D. The latency of saccades toward auditory targets in humans. Prog Brain Res 2002;140:51–59. [PubMed: 12508581]

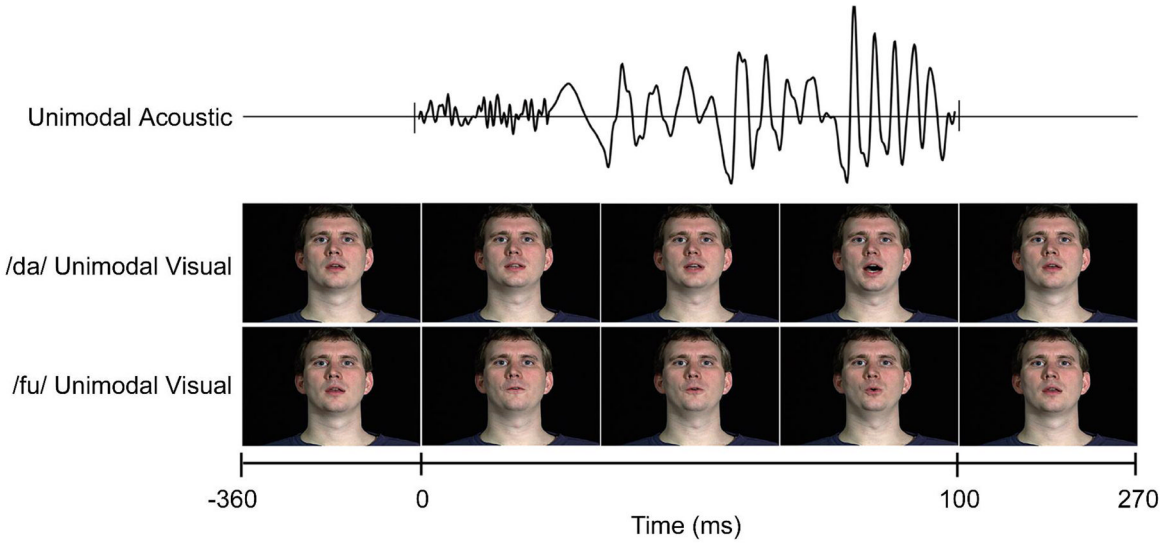

Figure 1.

Unimodal and audiovisual stimuli. The compressed timeline of $AV_{Concordant}$ and $AV_{Conflicting}$ stimuli is shown. Each unimodal visual utterance (/da/, /fu/ and /du/) was digitized from a recording of a male speaker. All three clips began and ended with the same neutral frame, but were different over the length of the utterance. The release of the consonant was edited to occur at frame 11 for all three visual tokens. A 100ms synthesized syllable, /da/, was created to emulate natural speech. For audiovisual presentation, the speech stimulus was paired with each visuofacial movement and acoustic onset occurred at 360ms.

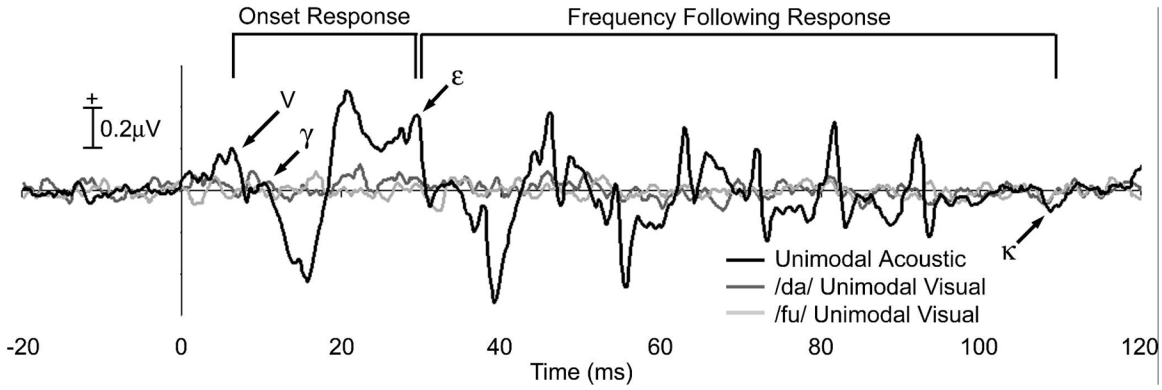

Figure 2.

Grand average responses to unimodal speech stimuli. Time 0 = acoustic stimulus onset. Prominent peaks of the Unimodal Acoustic response (black) to speech onset include Wave V followed by a positive deflection called Wave γ. The periodic portion of the response, called the frequency following response, beginning at Wave ε, and ending at Wave κ, is the region in which time between peaks reflects the wavelengths of the frequencies present in the stimulus. Audiovisual interaction effects were observed after ~7ms of the onset region. Neither replicable waves nor significant peaks were observed in the unimodal /da/ (dark gray) or /fu/ (light gray) conditions.

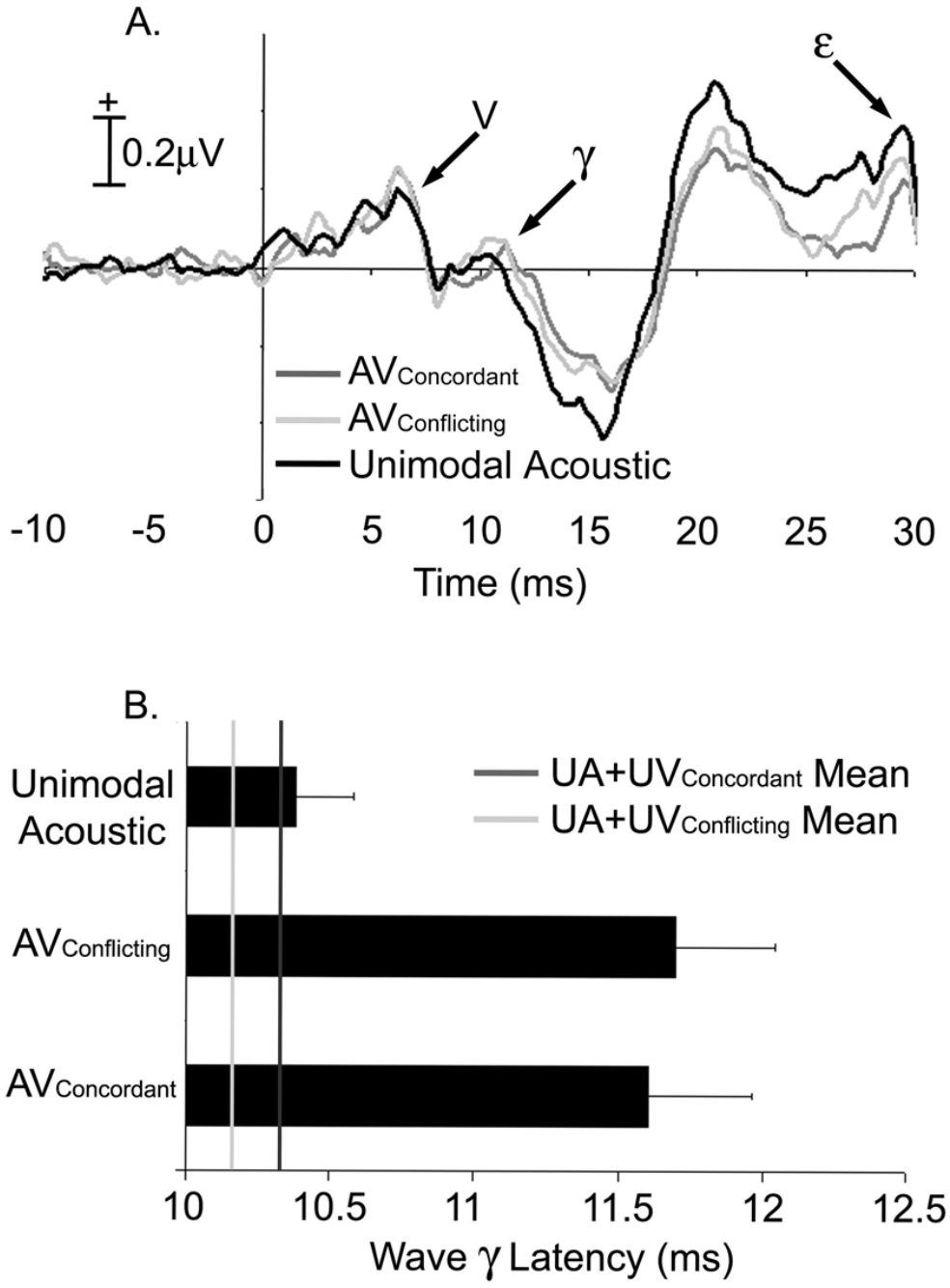

Figure 3.

Onset response was delayed and diminished in audiovisual conditions. A. Grand average onset responses to Unimodal Acoustic (black), $AV_{Concordant}$ (dark gray) and $AV_{Conflicting}$ (light gray) are shown. The size of both AV responses is noticeably smaller than that of the Unimodal Acoustic from approximately 10 to 30ms. Wave γ latency was prolonged, relative to the Unimodal Acoustic latency in both $AV_{Concordant}$ ($t=3.31$, $p=0.003$) and $AV_{Conflicting}$ ($t=2.37$, $p=0.002$) conditions. Neither earlier, nor later portions of the response, including Wave V and ε were affected in latency.

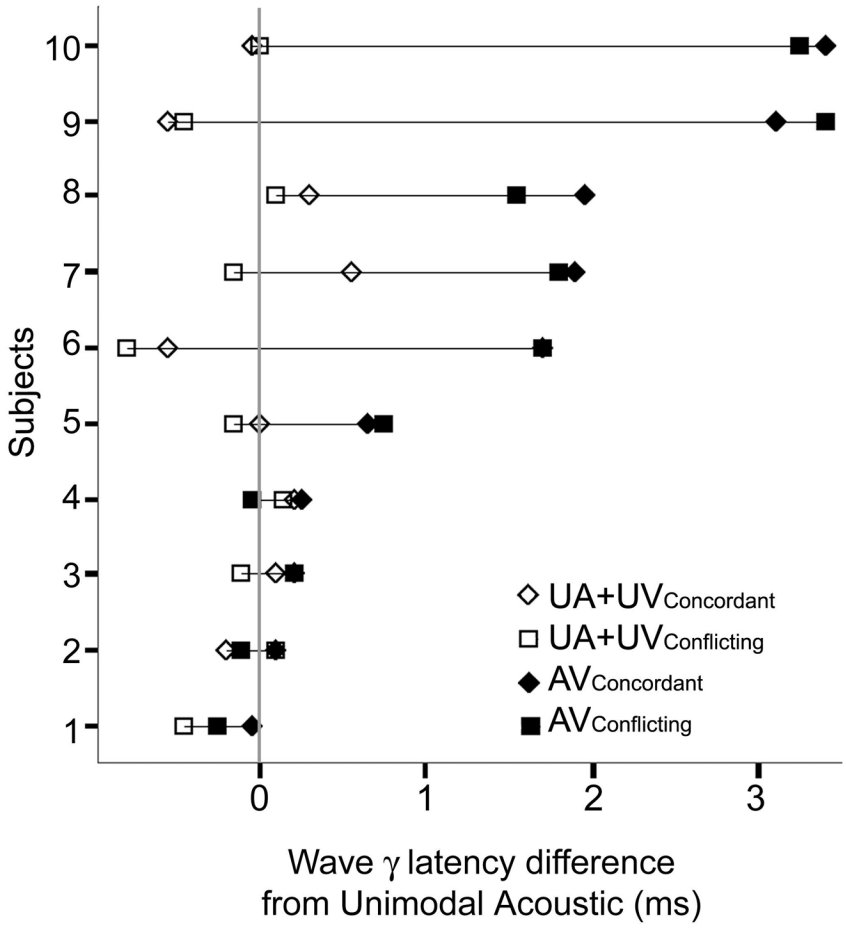

Figure 4.
Wave γ was delayed in audiovisual conditions. Individual subject latencies, normalized to the Unimodal Acoustic latency, are plotted. The Unimodal Acoustic latency is plotted as a gray line at time 0. The audiovisual delay was seen in 60% of the subjects.

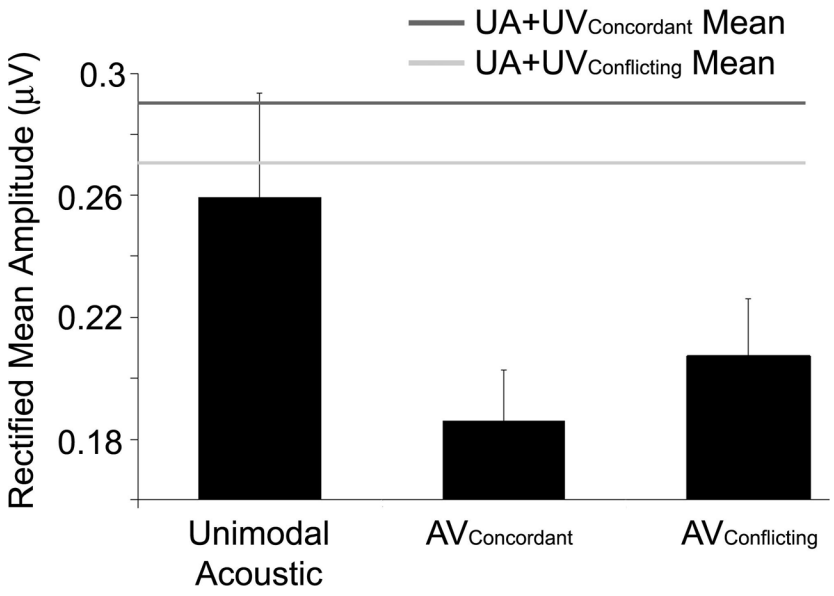

Figure 5.

AV suppression of onset response magnitude. A. The rectified mean amplitude (RMA) of the Unimodal Acoustic response over the onset region (Wave V to ε) was larger than both the $AV_{Concordant}$ and $AV_{Conflicting}$ responses. Audiovisual RMA values were smaller than their computed counterparts (UA+UV) and the $AV_{Concordant}$ response was smaller than that of the $AV_{Conflicting}$.

Table 1

Individual Wave γ Latencies

Subjects	Unimodal Acoustic	$AV_{Concordant}$	$AV_{Concordant}$	$UA+UV_{Concordant}$	$UA+UV_{Conflicting}$
1	10.70	10.65	10.45	10.65	10.25
2	10.10	10.20	10.00	9.90	10.20
3	10.95	11.15	11.15	11.05	10.85
4	11.05	11.30	11.00	11.25	11.20
5	11.25	11.90	12.00	11.25	11.10
6	9.40	11.10	11.10	8.85	8.60
7	9.65	11.55	11.45	10.20	9.50
8	10.55	12.50	12.10	10.85	10.65
9	9.80	12.90	13.20	9.25	9.35
10	10.35	13.75	13.60	10.30	10.35
Mean	10.38	11.70	11.61	10.36	10.21
St. Dev	0.63	1.08	1.14	0.83	0.83