Exp Brain Res (2005)
DOI 10.1007/s00221-005-0071-5

RESEARCH ARTICLE

Gabriella Musacchia · Mikko Sams · Trent Nicol
Nina Kraus

Seeing speech affects acoustic information processing in the human brainstem

Received: 15 November 2004 / Accepted: 18 May 2005

Abstract Afferent auditory processing in the human brainstem is often assumed to be determined by acoustic stimulus features alone and immune to stimulation by other senses or cognitive factors. In contrast, we show that lipreading during speech perception influences early acoustic processing. Event-related brainstem potentials were recorded from ten healthy adults to concordant (acoustic-visual match), conflicting (acoustic-visual mismatch) and unimodal stimuli. Audiovisual (AV) interactions occurred as early as ~11 ms post-acoustic stimulation and persisted for the first 30 ms of the response. Furthermore, the magnitude of interaction depended on AV pairings. These findings indicate considerable plasticity in early auditory processing.

Keywords Auditory · Visual · Brainstem · Multisensory · Speech

G. Musacchia (✉) · T. Nicol · N. Kraus
Auditory Neuroscience Laboratory,
Department of Communication Sciences,
Northwestern University, 2240 Campus Dr.,
Evanston, IL, 60208 USA
E-mail: g-musacchia@northwestern.edu
Tel.: +1-847-4912465
Fax: +1-847-4912523
E-mail: tgn@northwestern.edu
E-mail: nkraus@northwestern.edu

M. Sams
Laboratory of Computational Engineering,
Helsinki University of Technology,
Tekniikantie 14, Espoo (Innopoli II),
P.O.Box 9203, 02015 Hut, Finland
E-mail: Mikko.Sams@hut.fi

N. Kraus
Auditory Neuroscience Laboratory,
Department of Communication Sciences;
Neurobiology and Physiology; Otolaryngology,
Northwestern University, 2240 Campus Dr.,
Evanston, IL, 60208 USA

Introduction

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. One of the most ubiquitous and well-studied examples of AV integration in humans is seeing and hearing speech. Although acoustic and visual information are seamlessly combined without conscious control (Marks 2004), seeing articulation greatly aids speech acquisition (Kent 1984) and perception (Green 1987; Grant and Seitz 2000), especially in noisy environments (Sumby and Pollack 1954; MacLeod and Summerfield 1987). In addition, seeing articulation that does not match acoustic speech can drastically change what people "hear" (MacDonald and McGurk 1978; Sekiyama and Tohkura 1991). A prevalent model of how AV integration is accomplished posits that information from different modalities is processed along unisensory streams, which converge in cortical structures (Summerfield 1987; Massaro 1998). The combined representation is then processed in a feed-forward fashion that does not affect early, subcortical processing. While this hypothesis has proven to account for copious multisensory phenomena, evidence of AV interaction in subcortical structures encourages modification of the model. These observations prompted the current study, which investigates the timing of seen and heard speech interactions in the human brainstem.

Neuroimaging data have consistently identified cortical sites that show AV effects to speech, and evoked potential data show that effects in these areas happen as early as ~100 ms post-acoustic onset. Speech processing areas, such as primary auditory cortex, posterior superior temporal cortex (Lu et al. 1991; Calvert et al. 1997, 2000; Binder et al. 2000; Callan et al. 2003), Broca's area (Burton et al. 2000) and pre-motor cortex (Scott and Johnsrude 2003; Watkins and Paus 2004) have also shown activity during observation of visual articulatory movements (Calvert et al. 1997, 1999; Campbell et al. 2001; Nishitani and Hari 2002). In these studies, AV stimuli elicited response enhancement, relative to the sum of the unimodal responses, in multisensory cortices. Sensory-specific cortices, on the other hand, demonstrate response decrements due to AV interaction (Bushara et al. 2003; Klucharev and Sams 2004; Saito et al. 2005). Activity in sensory-specific and superior temporal cortices was affected by visual articulatory information as early as ~100 ms post-acoustic onset in electroencephalogram (EEG) and magnetoencephalogram (MEG) studies (Lu et al. 1991; Sams et al. 1991; Linkenkaer-Hansen et al. 1998; Möttönen et al. 2002). Non-speech stimuli have been shown to elicit AV interactions at earlier latencies (~90 ms) over primary auditory areas (Giard and Peronnet 1999).

At the subcortical level, neurons of the superior colliculus (SC) have been shown to receive convergent auditory and visual inputs, as well as exhibit AV response properties (Wallace et al. 1993, 1998; Stein 1998). This compelling line of research has revealed a predominance of supra-additive responses to convergent AV stimuli (from the same time or location) with sub-addition, or suppression, observed less often. Orientation accuracy and AV response properties of neurons in the SC neurons are severely degraded when the ecto-sylvian cortex is deactivated (Stein et al. 2002; Jiang et al. 2002; Jiang and Stein 2003; Perrault et al. 2003). These data suggest that cortical activity is necessary for AV responses to occur in subcortical structures. However, lesions of the SC also disrupt orientation to AV stimuli (Burnett et al. 2004) and there are some AV areas of the SC that do not receive descending projections from the cortex (Wallace et al. 2004). Because the time course of afferent and efferent AV response properties is not known, we cannot tell when interaction first occurs or the time course of corticofugal modulation.

The principal aim of this investigation was to test whether viewing articulatory gestures influenced the subcortical response to acoustic speech. Our approach was to record event-related responses to seen and heard speech using well-established methodology for recording the unimodal auditory speech-evoked brainstem response (Cunningham et al. 2001; King et al. 2002; Wible et al. 2004, 2005; Russo et al. 2004, 2005; Kraus and Nicol 2005). The speech-evoked response has been shown to be similar in precision to the click-evoked brainstem response, whose reliability and replicability have enabled its widespread clinical use. Peak-latency differences to click stimuli as small as a few milliseconds can be diagnostically significant in individuals with audiological or neurological abnormalities (Møller 1999; Jacobson 1991). Similarly, small delays in brainstem timing can distinguish normal and language-learning impaired groups using speech (Cunningham et al. 2001; King et al. 2002; Hayes et al. 2003; Wible et al. 2004, 2005; Russo et al. 2005; Johnson et al. in press).

The hypothesis for the current study was that acoustic and visual speech generates AV interactions in human subcortical structures. The time course of the interaction, recorded by evoked potentials, could help inform the extent to which AV mechanisms operate—early or late in the processing stream. To investigate this, an acoustic speech syllable was paired with either concordant or conflicting visual articulatory gestures (Klucharev et al. 2003). Brainstem responses were recorded when unimodal stimuli were presented separately and together. This presentation paradigm enabled two complementary data analysis strategies. Modulation effects, or, how the unimodal acoustic (UA) response is changed by the addition of visual stimuli, could be identified by differences between the AV response and responses to the UA stimulus. In addition, AV response features that deviated from the mathematical combination of the UA and unimodal visual (UV) responses could be considered evidence of true, non-linear, AV interaction mechanisms.

Materials and methods

Subjects

Ten adults (five females and five males; ages 18–35, mean age 25) participated in this experiment after giving informed, written consent. This experiment was carried out in accordance with the ethical principles laid down in the 1964 Declaration of Helsinki. All subjects performed visual and auditory tests to confirm normal or corrected 20/20 vision (Logarithmic Visual Acuity Chart "2000", Precision Vision) and hearing thresholds at or below 20 dB HL for octaves from 500 to 4000 Hz. The testing session was conducted in a sound-attenuated booth with a background sound level of 34 dB SPL. Subjects were seated in a comfortable chair, facing a 0.20 × 94 cm projection screen, 2.3 m away.

Stimuli and presentation sequence

The acoustic stimulus consisted of a five-formant synthetic 100 ms speech syllable, /da/, created with a DH Klatt synthesizer. Following a 10 ms consonant burst, a 30 ms formant transition was followed by a 60 ms steady-state vowel with a fundamental frequency of 100 Hz. Additional stimulus details can be found in previous studies (Bradlow et al. 1999). The consonant burst was amplified by an additional 3 dB (CoolEdit Pro 2000, Syntrillium), in order to elicit robust responses to acoustic onset. 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 that began and ended with the same neutral resting position (FinalCut Pro 4, Apple Software and MorphMan 4.0, Stoik Imaging). Each frame was presented for 33.15 ms (SD = 1.2), which brought the total visual stimulus duration to 630 ms. The release of the consonant was edited to occur at frame 11 for all three visual tokens. When presented together, acoustic speech onset occurred synchronously with presentation of the 11th frame (Fig. 1).

Stimulus sequences were delivered with Presentation software (Neurobehavioral Systems Inc. 2001) and presented in separate blocks of UA, AV and UV stimuli. The rate of presentation for all three-stimulus conditions was 1.59 s^{-1}. In the UA stimulus sequence, short blocks of 200 acoustic stimuli were presented at 84 dB SPL binaurally through ear inserts (ER-3, Etymotic research). Both stimulus polarities (condensation and rarefaction) were presented equally to ensure that the cochlear microphonic did not affect the brainstem response. To control for attention, subjects were asked to count how many sets of 50 /da/ tokens they heard. In the AV stimulus sequence, the synthesized speech syllable was paired with randomly presented /da/ ($AV_{Concordant}$, 40%), /fu/ ($AV_{Conflicting}$, 40%) and /du/ (target, 20%) visual utterances. The UV stimulus sequence consisted of randomly presented visual tokens (/da/ 40%, /fu/ 40%, /du/ 20%). To control for attention in the AV and UV conditions, subjects were asked to watch the video and count the number of /du/ tokens presented in each block.

Recording parameters

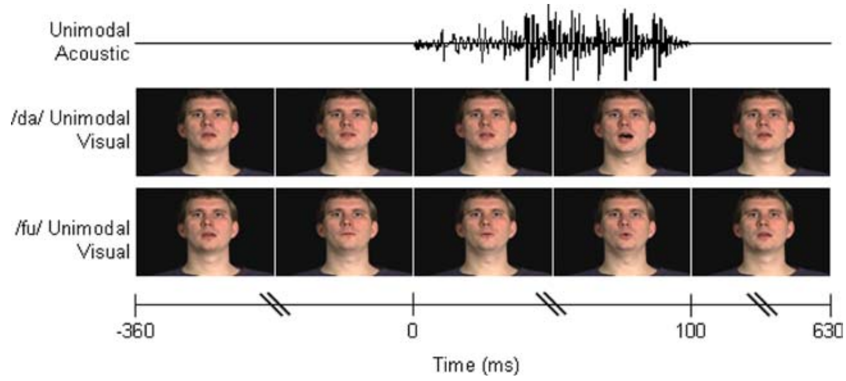
Continuous EEG was acquired with Neuroscan 4.3 (Compumedics, El Paso, TX, USA) from Cz (impedance < 5 kΩ), referenced to the nose, band pass filtered from 0.05 to 3000 Hz and digitized at 20,000 Hz. Simultaneously, online averaged evoked potentials were collected with an artifact criterion of > ± 65 μV to ensure that at least 1000 good repetitions per condition were collected. These averages were not used for data analysis. Instead, the continuous EEG was processed offline to create the epoched averages for each condition. The continuous file was band pass filtered from 75 to 2000 Hz to select the brainstem response frequencies (Hall 1992). The EEG was then divided into epochs (−20–120 ms post-acoustic onset). An artifact criterion of > ± 65 μV was applied to the epochs created from the continuous files in order to reject epochs that contained myogenic and eye blink artifacts. The remaining epochs were then separately averaged, according to stimulus type, and contained between 1000 and 1100 sweeps per non-target type. In order to correct for DC drift, the mean amplitude of the 20 ms epoch immediately preceding acoustic onset was subtracted from the response.

Response measurements

Signal-to-noise ratios (SNR) were calculated by comparing the pre- (−20–0 ms) stimulus and post- (0–100 ms) stimulus periods. The timing of the brainstem response was quantified by peak-latency and cross-correlation measures. The peaks of Waves V, γ,ε and κ (Fig. 2a) were chosen by visual inspection for all subjects, in all conditions, by two investigators. Cross-correlation measures (Pearson's r) were performed over a latency range that included Wave γ and the completion of its negative trough (8–20 ms). This analysis technique shifts one waveform in time to obtain a maximal correlation value. The lag at which this maximum correlation is attained is an indication of a response timing difference. Peak latency measures and cross-correlations provide information about when the response culminates in time and the degree of neural synchrony.

To assess the effects of visual speech on the size of the acoustic response, rectified mean amplitude (RMA) of the periodic and onset portions of the response was calculated. Individuals' latencies for Waves V, ε and κ were used to delineate the per-subject time ranges for RMA calculations. Onset RMAs were calculated between V and ε; RMAs, spectral analysis and cross-

Fig. 1 Acoustic and visual stimuli. Compressed timelines of two visual stimuli and the uncompressed acoustic stimulus are 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 differed over the length of the utterance. The release of the consonant was edited to occur at frame 11 for all three visual tokens. For AV presentation, the speech stimulus was paired with each visuofacial movement and acoustic onset occurred at time 0



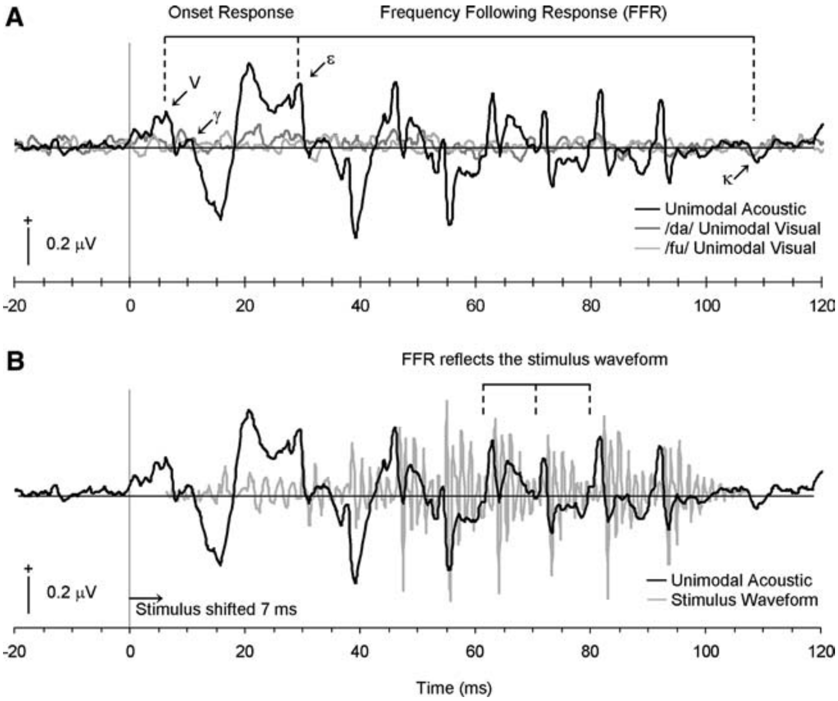


Fig. 2 Stimulus waveform and unimodal grand average responses. Time 0 = acoustic stimulus onset. **a** prominent peaks of the UA response (*black*) to speech onset include Wave V followed by a positive deflection called Wave γ. The periodic portion of the response, the frequency following response, beginning at Wave ε and ending at Wave κ, is the region in which time between peaks reflects phase locking to the stimulus waveform. Replicable waves were not observed in the UV /da/ (*dark gray*) or /fu/ (*light gray*) conditions. **b** the grand average UA response is overlayed on the stimulus waveform. The onset of the stimulus has been shifted in time to correspond to response onset. Peaks of the periodic portion of the stimulus waveform can be seen to correspond to peaks of the frequency following response

correlation of the frequency following response (FFR) were calculated between ε and κ.

Data analysis

Modulation effects or changes in the acoustic response due to the addition of visual stimuli, were investigated using a repeated-measure ANOVA with three levels as within-subject factors (UA, $AV_{Concordant}$ and $AV_{Conflicting}$). Interaction effects, or the difference between the AV responses and the summed unimodal responses were explored using a repeated-measure ANOVA with four levels as within-subjects factor ($AV_{Concordant}$, $AV_{Conflicting}$ and their summed unimodal counterparts). Greenhouse–Geisser corrections were applied if applicable. Protected paired *t*-test were performed subsequent to significant ANOVAs. Correlation values and lags were subjected to single-sample *t*-test to determine if they differed from zero.

Results I: description of responses

The grand average responses of all subjects to the three-unimodal stimuli (UA /da/, UV /da/ and UV /fu/) are shown in Fig. 2a. The onset of the acoustic stimulus elicited a series of transient, biphasic peaks. Figure 2b shows that the vowel portion of the stimulus evoked an FFR, which reflects phase locking to the waveform of the stimulus (Marsh et al. 1975; Galbraith et al. 1995).

In all subjects, and evident in the average, the first prominent peak, Wave V (UA mean latency 6.16 ms, SD = 0.34), was followed by a negative trough, previously reported as Wave A (Russo et al. 2004). Wave V mean latency and standard deviation was similar to the normative values reported in previous studies. A positive peak that was not observed in previous studies followed Waves V and A. Some differences in response morphology were expected due to differences between the current and previous stimuli. To avoid confusion between the present and previously reported peaks, the Greek alphabet was used to describe peaks following Wave A. The positive peak following Wave A was referred to as Wave γ. The periodic portion of the FFR began with a positive peak, Wave ε, and ended at a negative peak, Wave κ. Neither the /da/ nor /fu/ UV responses elicited replicable peaks across subjects and exhibited low SNRs (0.94 and 1.32, respectively), indicating that the visual stimulus alone elicited little evoked activity with the recording parameters and electrode placement reported here.

Results could not be explained by differences across conditions in SNR or overall electrical activity, as measured by the RMA over pre-stimulus periods. The SNR values demonstrated that the signal measurements were distinguishable from noise in the UA and AV conditions ($SNR_{UA} = 5.23$, SD = 1.02, $SNR_{Concordant} = 5.68$, SD = 1.20, $SNR_{Conflicting} = 4.55$, SD = 1.47). The SNR values were not significantly different across conditions ($F(2,18) = 0.96$; $p = 0.44$, ε = 0.96). The overall electrical activity generated by electrical noise and non-stimulus related EEG activity, measured by the RMA over −20–0 ms, was not significantly different across UA, AV and UV conditions ($F(2,18) = 0.495$; $p = 0.63$).

Results II: lipreading delays the brainstem response to speech onset

The presentation of either visual stimulus modulated the timing of the brainstem response to speech at Wave γ (Fig. 3, Table 1). There were no differences in Wave V, ε and κ latencies across conditions. Latency differences at Wave γ were evident across conditions ($F(2,18) = 6.77$; $p < 0.05$, $\varepsilon = 0.51$) and prolonged in both $\mathrm{AV}_{\mathrm{Concordant}}$ and $\mathrm{AV}_{\mathrm{Conflicting}}$ responses, relative to the UA response ($p_{\mathrm{Concordant}} < 0.01$, $t = 3.26$; $p_{\mathrm{Conflicting}} < 0.01$, $t = 3.11$). Wave γ latencies in the concordant condition were prolonged in nine out of ten subjects and in seven out of ten in the conflicting condition. Wave γ latencies did not differ significantly between the two AV conditions.

Inter-peak intervals between Wave V and Wave γ ($\gamma_{\mathrm{latency}}$-$\mathrm{V}_{\mathrm{latency}}$) were computed to confirm that the modulation delay occurred subsequent to Wave V. The inter-peak interval difference was evident across UA and AV conditions ($F(2,18) = 4.88$; $p < 0.05$, $\varepsilon = 0.56$) and was prolonged in both the $\mathrm{AV}_{\mathrm{Concordant}}$ ($p = 0.02$, $t = 2.53$) and $\mathrm{AV}_{\mathrm{Conflicting}}$ ($p = 0.01$, $t = 2.85$) conditions when compared to the UA condition. A prolonged inter-peak interval was evident in nine out of the ten individuals in the $\mathrm{AV}_{\mathrm{Concordant}}$ condition and in seven subjects in the $\mathrm{AV}_{\mathrm{Conflicting}}$. This finding, combined with the null result for Wave V latencies across conditions ($F(2,18) = 0.87$; $p = 0.44$, $\varepsilon = 0.84$), indicated that modulation of the unimodal response did not begin previous to Wave γ.

A maximal correlation between UA and $\mathrm{AV}_{\mathrm{Concordant}}$ onset responses occurred with a lag of 0.69 ms ($p < 0.05$, $t = 2.66$). The maximum correlation between UA and $\mathrm{AV}_{\mathrm{Conflicting}}$ responses (0.36 ms lag) was not significantly different from zero.

The difference between the two AV conditions and their computed UA + UV counterparts revealed a true non-linear AV interaction at Wave γ. Wave γ latencies were different across conditions ($F(3,27) = 6.21$; $p < 0.05$, $\varepsilon = 0.38$) with delays evident in both the $\mathrm{AV}_{\mathrm{Concordant}}$ ($p < 0.01$, $t = 2.91$) and $\mathrm{AV}_{\mathrm{Conflicting}}$ ($p < 0.01$, $t = 3.17$) responses when compared to their respective unimodal sums. Nine out of ten individuals exhibited this latency interaction in the $\mathrm{AV}_{\mathrm{Concordant}}$ and eight out of ten in the $\mathrm{AV}_{\mathrm{Conflicting}}$ condition.

Inter-peak intervals between Wave V and γ also demonstrated an interaction ($F(3,27) = 4.46$; $p = 0.011$, $\varepsilon = 0.39$). Prolonged intervals were evident in both $\mathrm{AV}_{\mathrm{Concordant}}$ ($p < 0.05$, $t = 2.06$) and $\mathrm{AV}_{\mathrm{Conflicting}}$ ($p = 0.011$, $t = 2.71$) conditions compared to their respective unimodal sums. Again, no differences in

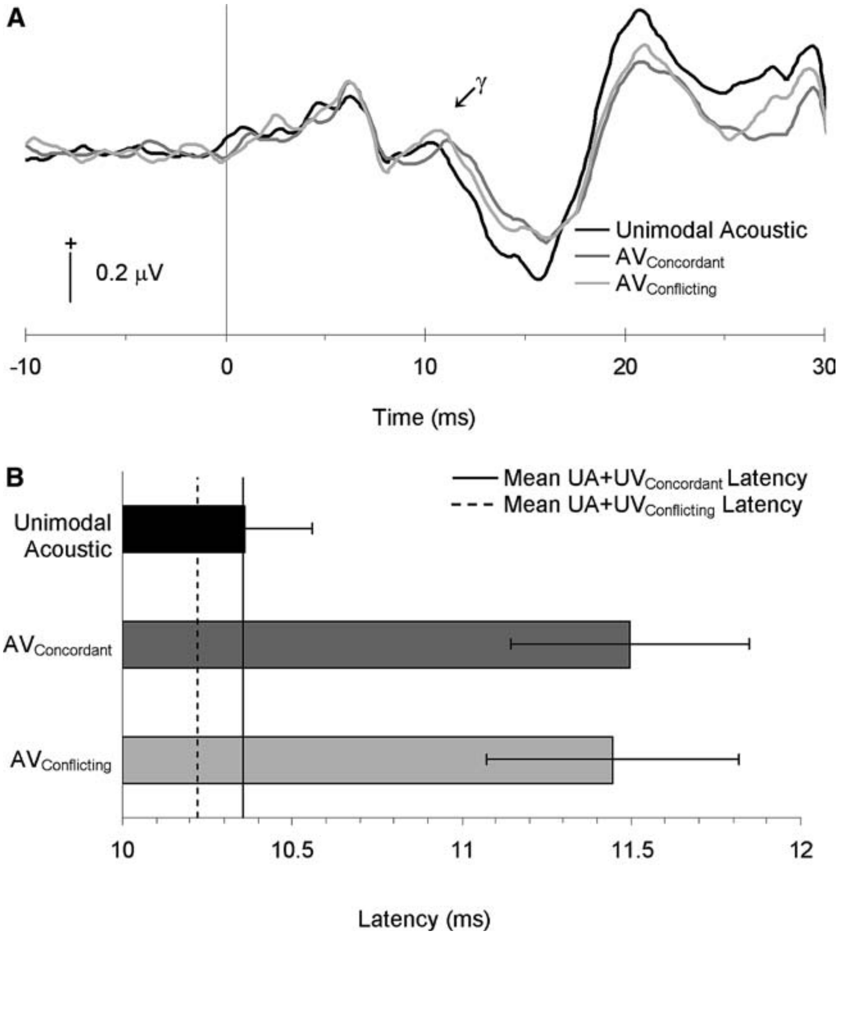


Fig. 3 Onset responses in UA and the two AV conditions. **a** grand average onset responses to UA (*black*), $\mathrm{AV}_{\mathrm{Concordant}}$ (*dark gray*) and $\mathrm{AV}_{\mathrm{Conflicting}}$ (*light gray*) are shown. The size of both AV responses is noticeably smaller than that of the UA response from approximately 10–30 ms. Wave γ latency was prolonged, relative to the UA latency in both $\mathrm{AV}_{\mathrm{Concordant}}$ and $\mathrm{AV}_{\mathrm{Conflicting}}$ conditions. **b** mean Wave γ latencies are shown for UA and the two AV responses. Error bars show the standard error

Table 1 Individual Wave γ latencies (ms) in UA, $AV_{Concordant}$, $AV_{Conflicting}$ and the sum of UA + UV responses

Subject	UA	AV Concordant	AV Conflicting	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	10.35	10.45	10.50	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	10.36	11.49	11.44	10.35	10.20
SD	0.62	1.10	1.17	0.82	0.83

interaction effects were observed between concordant and conflicting conditions. It is important to note that our data reflect some variance in Wave γ delay across individuals. The perceptual or subject characteristics that may have contributed to this variance were not pursued in this study, but are an intriguing direction of future research.

Results III: two types of visual stimuli modulate the size of the acoustic brainstem response to speech differently

The two types of visual stimuli modulated the size of the acoustic brainstem response differently. The RMA values, as measured between waves V and ε, were different across UA (Mean RMA 0.26 μV, SD = 0.11), $AV_{Concordant}$ (Mean RMA 0.19 μV, SD = 0.05) and $AV_{Conflicting}$ (Mean RMA 0.21 μV, SD = 0.06) conditions ($F(2,18) = 5.82$; $p < 0.01$, $\varepsilon = 0.59$) and were diminished in both the $AV_{Concordant}$ ($p < 0.01$, $t = 3.31$) and $AV_{Conflicting}$ ($p < 0.05$, $t = 2.37$) responses compared to the 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, Table 2 and Fig. 4 show a greater suppression in the $AV_{Concordant}$ response than the $AV_{Conflicting}$ response ($p < 0.05$, $t = 2.47$).

Table 2 RMA (μV) of individual onset responses in UA, $AV_{Concordant}$ and $AV_{Conflicting}$ conditions

Subject	UA	AV Concordant	AV Conflicting
1	0.21	0.20	0.21
2	0.22	0.17	0.18
3	0.51	0.30	0.34
4	0.35	0.20	0.19
5	0.18	0.17	0.22
6	0.20	0.10	0.10
7	0.12	0.16	0.21
8	0.26	0.17	0.20
9	0.23	0.24	0.21
10	0.32	0.15	0.21
Mean	0.26	0.19	0.21
SD	0.11	0.05	0.06

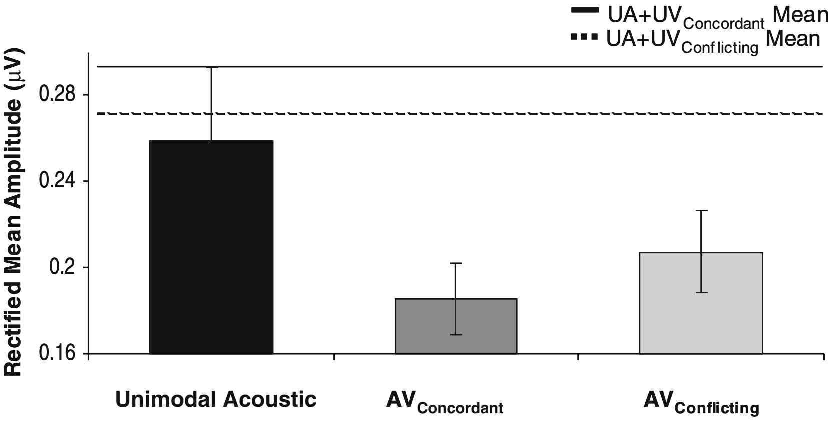
The size of the AV onset responses compared to their summed unimodal counterparts revealed an AV interaction effect. The onset RMA values in both AV conditions were smaller than those in the summed unimodal responses ($F(3,27) = 11.26$; $p < 0.01$, $\varepsilon = 0.40$; $p_{Concordant} < 0.01$, $t = 4.97$; $p_{Conflicting} < 0.01$, $t = 3.01$). The extent of the AV suppression over the onset response was not correlated with the length of the Wave γ delay for either concordant or conflicting stimuli. No statistical evidence of modulation or AV interaction was observed over the FFR region of the responses, using the three methods described in *Response measurements*.

Discussion

The results of the current study demonstrate that seeing facial movements (lipreading) delays and suppresses the amplitude of the human brainstem response to acoustic speech. The effect of AV delay, on average 1.3 ms, was evident in both $AV_{Concordant}$ and $AV_{Conflicting}$ conditions and occurred as early as 11 ms post-acoustic stimulation. Although both the $AV_{Concordant}$ and $AV_{Conflicting}$ RMAs were smaller compared to the UA condition, the extent of diminution depended on the type of facial movement. The $AV_{Concordant}$ response was more suppressed than those to the $AV_{Conflicting}$ response. The observed effects in the AV conditions could not be attributed to activity elicited by the visual stimuli alone, because measures of the summed unimodal responses (UA + UV) did not differ from UA responses.

These results suggest that early 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 AV speech interaction. The time frame of the delay, ~11 ms post-acoustic stimulus, precludes the possibility of AV interaction from simultaneous visual information at acoustic onset, because visual information takes longer to propagate to brainstem structures than acoustic information (Wallace et al. 1998). Therefore, the interaction must be due to the processing of visual information that precedes acoustic stimulation. The authors

Fig. 4 UA and AV onset response magnitude. The rectified mean amplitude (RMA, μV) of the UA response over the onset region (Wave V to ε) was larger than both the $\mathrm{AV_{Concordant}}$ and $\mathrm{AV_{Conflicting}}$ responses. AV RMA values were smaller than their computed counterparts (as indicated by lines) and the $\mathrm{AV_{Concordant}}$ response was smaller than that of the $\mathrm{AV_{Conflicting}}$



suggest two hypotheses as to how this may be accomplished.

One hypothesis is that visual information that precedes acoustic stimulation engages cortical gating or attention mechanisms that directly modulate subcortical acoustic processing. Although early components of the acoustic-evoked response (latency range 2–40 ms) have not generally shown replicable effects of attention (for review, see Picton and Hillyard 1974), some effects have been observed. In AV conditions and in cases of very difficult acoustic target detection, effects of attention have been observed between 20 and 50 ms post-acoustic onset (Woldorff et al. 1987; Hoormann et al. 2000; Teder-Salejarvi et al. 2002; Woldorff and Hillyard 1991). The results of these studies suggest that early auditory processing could be selectively tuned by mechanisms recorded as slow 'anticipatory' evoked responses to stimulus cues. The AV effects described in these studies produced considerably smaller delays than those observed here. Although hypotheses regarding speech versus non-speech stimuli cannot be derived directly from this study, it is possible that lipreading may produce larger differences between unimodal and bimodal stimuli than those observed to non-speech stimuli. The complexity of speech stimuli, relative to flashes and tones for example, or the extensive experience humans have with lipreading may contribute to the difference in effect size.

Converging evidence from animal and human studies also suggests that the corticofugal system has a role in attentional modulation of subcortical auditory nuclei (for review, see Suga and Ma 2003) as low as the cochlear nucleus (Oatman and Anderson 1977). In these studies, activity in the auditory nuclei was reduced when subjects attended to visual stimuli, which parallels the amplitude suppression observed in the current study. Recent investigations have shown that the synthesis of acoustic and visual cues in the cat SC is greatly compromised when areas of the auditory cortex are deactivated (Jiang and Stein 2003), indicating that the cortex plays a functional role in mediating AV integration in the SC. The cortical gating/attentional hypothesis could also explain the range of AV delay across individuals. Target identification scores were used only to ensure 80% correct identification, and statistical analysis of the responses was not performed. Therefore, it is possible that the extent of delay is related to greater attentional focus and higher hit rates.

The alternative hypothesis is that ongoing activity in visual brainstem nuclei, combined with afferent acoustic processing, increases the degree of neural asynchrony, relative to unimodal processing, recorded as total electrical activity from the scalp. A fundamental property of event-related potentials is that a decrease in synchrony of firing, for example, aggregate neural populations firing at slightly different times results in longer peak latencies (Hall 1992). Visual or AV nuclei in the brainstem that do not fire in concert with those involved in UA processing could produce the observed delay. 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 the scalp. Although, AV fMRI data from the human SC have been limited to non-speech stimuli (Calvert 2001), 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 $\mathrm{AV_{Concordant}}$ and $\mathrm{AV_{Conflicting}}$ 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 in the acoustic and visual spatial maps of the barn owl brainstem to spatially concordant cues (Hyde and Knudsen 2001), prompting the theory that concordant stimuli are 'easier' to process. It is conceivable that the AV response to our primary means of communication, speech, engages a similar interaction mechanism.

Although single-channel ERP recording precludes localization, the timing of the AV effects observed in this study is consistent with activation of nuclei before thalamus and cortex. The latency differences between UA and AV responses take place before initial excitation of the human primary auditory cortex, detected in direct intracranial recordings at 12–15 ms post-acoustic stimulation (Celesia 1968). It is important to note that Celesia and colleagues used rapid-onset click stimuli, which elicit earlier latencies than tone or speech stimuli (Hall 1992). Tone stimuli have been shown to elicit a peak of activity at 13.5 ms post-acoustic onset in the human thalamus and at 17 ms in the auditory cortex (Yvert et al. 2002). Because the AV delay observed in the current study occurred at about 11 ms post-acoustic stimulation, i.e., before reported activation of auditory cortex and thalamus, it is reasonable to suggest that the interaction is taking place in the afferent brainstem pathway.

Although Wave V latency was not prolonged in the AV conditions, this does not preclude the contribution of Wave V generators to later peaks. Studies designed to determine the sources of scalp-recorded auditory brainstem response indicate that the inferior colliculus and lateral lemniscus are the primary generators of Wave V (Gardi et al. 1979). However, these studies also consistently demonstrated that the onset discharge of single units in multiple generator sites corresponds in time to the latency of several different (II–V) waves. Detection of where the AV interactions are taking place is furthered by evidence of converging acoustic and visual inputs on neurons in the SC (Meredith and Stein 1986). Despite the localization constraints of ERPs, nuclei of the midbrain emerge as the most likely generators of interaction in the current study.

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

One implication is that speech is processed via a specialized module in which the articulatory gestures could influence afferent speech processing in a way that is unique from non-speech tokens. A long-debated question is whether speech is processed differently than non-speech sounds (Chomsky 1985; Hauser et al. 2002). Separate brain mechanisms have been shown to be active for acoustic speech and non-speech processing (e.g., Tervaniemi and Hugdahl 2003; Binder et al. 2000) and a strong relationship between phoneme perception and motor imitation has been found (Gallese et al. 1996). A related implication is that extensive experience with AV speech results in plasticity of the system such that visual articulatory gestures have unique access to the auditory brainstem. This would suggest that speech is processed in a qualitatively different way from non-speech, and that precursors of phonetic discrimination operate at the level of the brainstem to discern the degree of AV 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) independent of their concordance, or lack thereof, to the accompanying sound, may be responsible for the effect.

These findings challenge the prevailing view about the human brainstem as a passive receiver/transmitter of modality-specific information. Future investigations on the nature of early AV interactions, and the experimental conditions that contribute to the extent of 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 colleagues in the Auditory Neuroscience Laboratory at Northwestern University as well as Dan Zellner and the staff at Northwestern's Digital Media Studio for their film editing expertise.

References

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and non-speech sounds. Cereb Cortex 10:512–528

Bradlow AR, Kraus N, Nicol TG, Mcgee TJ, Cunningham J, Zecker SG, Carrell TD (1999) Effects of lengthened formant transition duration on discrimination and neural representation of synthetic CV syllables by normal and learning-disabled children. J Acoust Soc Am 106:2086–2096

Burnett LR, Stein BE, Chaponis D, Wallace MT (2004) Superior colliculus lesions preferentially disrupt multisensory orientation. Neuroscience 124:535–547

Burton MW, Small SL, Blumstein SE (2000) The role of segmentation in phonological processing: an fMRI investigation. J Cogn Neurosci 12:679–690

Bushara KO, Hanakawa T, Immisch I, Toma K, Kansaku K, Hallett M (2003) Neural correlates of cross-modal binding. Nat Neurosci 6:190–195

Callan DE, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E (2003) Neural processes underlying perceptual enhancement by visual speech gestures. Neuroreport 14:2213–2218

Calvert GA (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb Cortex 11:1110–1123

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. Science 276:593–596

Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS (1999) Response amplification in sensory-specific cortices during crossmodal binding. Neuroreport 10:2619–2623

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

Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, Brammer MJ, David AS (2001) Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts. Cogn Brain Res 12:233–243

Celesia GG (1968) Auditory evoked responses. Intracranial and extracranial average evoked responses. Arch Neurol 19:430–437

Chomsky N (1985) The logical structure of linguistic theory. The University of Chicago Press, Chicago

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

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

Gallese V, Fadiga L, Fogassi L, Rizzolatti G (1996) Action recognition in the premotor cortex. Brain 119:593–609

Gardi J, Merzenich M, McKean C (1979) Origins of the scalp recorded frequency-following response in the cat. Audiology 18:358–381

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

Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 108:1197–1208

Green KP (1987) The perception of speaking rate using visual information from a talker's face. Percept Psychophys 42:587–593

Hall JWI (1992) Handbook of auditory evoked responses. Allyn and Bacon, Needham Heights

Hauser MD, Chomsky N, Fitch WT (2002) The faculty of language: what is it, who has it, and how did it evolve? Science 298:1569–1579

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

Hoormann J, Falkenstein M, Hohnsbein J (2000) Early attention effects in human auditory-evoked potentials. Psychophysiology 37:29–42

Hyde PS, Knudsen EI (2001) A topographic instructive signal guides the adjustment of the auditory space map in the optic tectum. J Neurosci 21:8586–8593

Jacobson J (1991) The auditory brainstem response. Prentice-Hall, Englewood Cliffs

Jiang W, Stein BE (2003) Cortex controls multisensory depression in superior colliculus. J Neurophysiol 90:2123–2135

Jiang W, Jiang H, Stein BE (2002) Two corticotectal areas facilitate multisensory orientation behavior. J Cogn Neurosci Nov 15:1240–1255

Johnson KL,Nicol TG,Kraus N (2005) The brainstem response to speech. A biological marker of auditory processing EAR and HARING (in press)

Kent RD (1984) Psychobiology of speech development: coemergence of language and a movement system. Am J Physiol 246:R888–R894

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

Klucharev V, Sams M (2004) Interaction of gaze direction and facial expressions processing: ERP study. Neuroreport 22:621–625

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

Kraus N, Nicol T (2005) Brainstem origins for cortical 'what' and 'where' pathways in the auditory system. Trends Neurosci 28:176–181

Linkenkaer-Hansen K, Palva JM, Sams M, Hietanen JK, Aronen HJ, Ilmoniemi RJ (1998) Face-selective processing in human extrastriate cortex around 120 ms after stimulus onset revealed by mag. Neurosci Lett 253:147–150

Lu ST, Hämäläinen MS, Hari R, Ilmoniemi RJ, Lounasmaa OV, Sams M, Vilkman V (1991) Seeing faces activates three separate areas outside the occipital visual cortex in man. Neuroscience 43:287–290

Møller AR (1999) Neural mechanisms of BAEP. Electroencephalogr Clin Neurophysiol Suppl 49:27–35

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

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

Marks LE (1982) Bright sneezes and dark coughs, loud sunlight and soft moonlight. J Exp Psychol Hum Percept Perform 8:177–193

Marks LE (2004) Cross-modal interactions in speeded classification. In: Calvert GA, Spence C, Stein BE (eds) The handbook of mutisensory processes. MIT Press, Cambridge, pp 85–106

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

Massaro DW (1998) Perceiving talking faces: from speech perception to a behavioral principle. MIT Press, Cambridge, MA

Meredith MA, Stein BE (1986) Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. J Neurophysiol 56:640–662

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

Nishitani N, Hari R (2002) Viewing lip forms: cortical dynamics. Neuron 19:1211–1220

Oatman LC, Anderson BW (1977) Effects of visual attention on tone burst evoked auditory potentials. Exp Neurol 57:200–211

Perrault TJ, Vaughan JW, Stein BE, Wallace MT (2003) Neuron-specific response characteristics predict the magnitude of multisensory integration. J Neurophysiol 90:4022–4026

Picton TW, Hillyard SA (1974) Human auditory evoked potentials. II. Effects of attention. Electroencephalogr Clin Neurophysiol 36:191–199

Russo N, Nicol T, Musacchia G, Kraus N (2004) Brainstem responses to speech syllables. Clin Neurophysiol 115:2021–2030

Russo NM, Nicol TG, Zecker SG, Hayes EA, Kraus N (2005) Auditory training improves neural timing in the human brainstem. Behav Brain Res 6:95–103

Saito DN, Yoshimura K, Kochiyama T, Okada T, Honda M, Sadato N (2005) Cross-modal binding and activated attentional networks during audio-visual speech integration: a functional MRI study. Cereb Cortex 16 (Epub ahead of print)

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

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends Neurosci 26:100–107

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

Stein BE (1998) Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. Exp Brain Res 123:124–135

Stein BE, Wallace MW, Stanford TR, Jiang W (2002) Cortex governs multisensory integration in the midbrain. Neuroscientist 8:306–314

Suga N, Ma X (2003) Multiparametric corticofugal modulation and plasticity in the auditory system. Nat Rev Neurosci 4:783–794

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

Summerfield Q (1987) Hearing by eye. In Dodd B, Campbell R (eds) Lawrence Erlbaum Associates, Hillsdale, pp 3–51

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

Tervaniemi M, Hugdahl K (2003) Lateralization of auditory-cortex functions. Brain Res Rev 43:231–246

Wallace MT, Meredith MA, Stein BE (1993) Converging influences from visual, auditory, and somatosensory cortices onto output neurons of the superior colliculus. J Neurophysiol 69:1797–1809

Wallace MT, Meredith MA, Stein BE (1998) Multisensory integration in the superior colliculus of the alert cat. J Neurophysiol 80:1006–1010

Wallace MT, Perrault TJ Jr, Hairston WD, Stein BE (2004) Visual experience is necessary for the development of multisensory integration. J Neurosci 27:9580–9584

Watkins K, Paus T (2004) Modulation of motor excitability during speech perception: the role of Broca's area. J Cogn Neurosci 16:978–987

Wible B, Nicol T, Kraus N (2004) Atypical brainstem representation of onset and formant structure of speech sounds in children with language-based learning problems. Biol Psychol 67:299–317

Wible B, Nicol T, Kraus N (2005) Correlation between brainstem and cortical auditory processes in normal and language-impaired children. Brain 128:417–423

Woldorff MG, Hillyard SA (1991) Modulation of early auditory processing during selective listening to rapidly presented tones. Electroencephalogr Clin Neurophysiol 79:170–191

Woldorff M, Hansen JC, Hillyard SA (1987) Evidence for effects of selective attention in the mid-latency range of the human auditory event-related potential. Electroencephalogr Clin Neurophysiol Suppl 40:146–154

Yvert B, Fischer C, Guenot M, Krolak-Salmon P, Isnard J, Pernier J (2002) Simultaneous intracerebral EEG recordings of early auditory thalamic and cortical activity in human. Eur J Neurosci 16:1146–1150