Open Access

Neural Dynamics of Audiovisual Integration for Speech and Non-Speech Stimuli: A Psychophysical Study

Nicholas Altieri*

Idaho State University, Dept. of Communication Sciences and Disorders, 921 S. 8th Ave., Pocatello, ID 83209, USA

Abstract: This study investigated the extent to which audiovisual speech integration is special by comparing behavioral and neural measures using both speech and non-speech stimuli. An audiovisual recognition experiment presenting listeners with auditory, visual, and audiovisual stimuli was implemented. The auditory component consisted of sine wave speech, and the visual component consisted of point light displays, which include point-light dots that highlight a talker's points of articulation. In the first phase, listeners engaged in a discrimination task where they were unaware of the linguistic nature of the auditory and visual stimuli. In the second phase, they were informed that the auditory and visual stimuli were spoken utterances of /be/ ("bay") and /de/ ("day"), and they engaged in the same task. The neural dynamics of audiovisual integration was investigated by utilizing EEG, including mean Global Field Power and current density reconstruction (CDR). As predicted, support for divergent regions of multisensory integration between the speech and non-speech stimuli was obtained, namely greater posterior parietal activation in the non-speech condition. Conversely, reaction-time measures indicated qualitatively similar multisensory integration across experimental conditions.

Keywords: Audio-visual speech, integration, capacity, global field power.

1. INTRODUCTION

This study investigated the spatio-temporal dynamics of audiovisual integration in speech recognition, compared to non-speech stimuli that employed identical temporal dynamics as well as acoustic and visual characteristics. While speech recognition in normal-hearing individuals relies heavily on auditory processing, research on the multimodal aspects of language perception has long established the influence of the visual signal on recognition, and the level of enhancement that results from the presence of the signal [1] [2]. One of the most studied phenomena in multimodal perception relates to *audiovisual enhancement*. In the context of word, syllable, or sentence recognition, *enhancement* denotes the level of gain in accuracy experienced when the listener is able to see the talker's face [2-4]. Audiovisual enhancement is most noticeable under poor listening conditions and low auditory signal-to-noise ratios (S/N) (e.g., [2, 5]) and often in individuals with hearing loss [6]. Evidence also indicates that visual speech information does provide some benefit under relatively good listening conditions [2] (cf. [7]).

Another example of the influence of visual speech information on auditory processing is the *McGurk effect* [1]. The McGurk effect and other related perceptual fusions occur when the conflicting visual signal influences the semantic content of the auditory percept. The McGurk effect occurs, for example, when listeners are presented with a talker articulating an auditory /ba/ dubbed over a visually articulated "ga", which leads to the fusion of "da" or "tha". This effect has been attributed to the integration of phonetic information, and the cross-modal inhibitory influence that visual information may have on conflicting auditory processes [8] [9]. Likewise, other research has reported cases in which the listener may perceive the visual stimulus when the auditory and visual components of the signal mismatch [10]. Taken together, audiovisual enhancement and perceptual fusions may be thought of as *audiovisual integration*.

The purpose of this study was to examine the extent to which audiovisual speech integration differs from integration for non-speech stimuli. First, audiovisual enhancement will be measured behaviorally by comparing capacity (a reaction-time measure) across "speech" versus "non-speech" conditions. Capacity assesses "efficiency" or energy expenditure and has recently been used to assess the extent to which individuals efficiently combine speech information from different modalities [7, 11, 12]. By hypothesis, qualitatively similar levels of energy expenditure (capacity) are predicted for speech and non-speech stimuli [7, 13]. Specifically, multisensory integration may be inefficient, particularly in high accuracy settings (see [14] for discussion of "inverse effectiveness").

Speech and non-speech stimuli may elicit qualitatively similar capacity levels, although different brain circuits may be involved in processing different stimulus types. Audiovisual speech and non-speech integration will be examined neurologically using EEG current density reconstruction (CDR) to uncover potential differences in underlying patterns of neural activation. The prediction here is that audiovisual activation patterns should be observed across a distributed network of brain regions extending beyond STS. Predicted brain areas involved in multisensory integration include: inferior/posterior frontal, temporal and parietal ar-

*Address correspondence to this author at the Idaho State University, Dept. of Communication Sciences and Disorders, 921 S. 8th Ave., Pocatello, ID 83209, USA; Tel: 208-282-2741; Email: altinich@isu.edu

eas, and the supramarginal gyrus (SMG) [15]. A subsidiary hypothesis, based on recent fMRI findings comparing BOLD activation for audiovisual speech to non-speech stimuli (e.g., tools such as a paper-cutter or a hammer [13]), is that non-speech will activate different brain circuits than speech stimuli, including broader posterior regions of the temporal and parietal lobe in the early stages of processing (Fig. **5** from the discussion). Capacity and CDRs will be discussed in conjunction with mean Global field power (mGFP).

This study shall assess integration when listeners perceive auditory and visual stimuli as acoustic-phonetic gestures [16], and also when they perceive the same exact events as non-speech. Speech and non-speech mode can be evoked by sine wave speech replicas of natural speech [17]. *Sine wave speech* uses sinusoidal pulses to trace the formants of speech, although it lacks cues relevant to natural speech such the broadband formant structure. In a study investigating differences between audiovisual speech and non-speech integration, Toumainen and colleagues paired sine wave and natural speech tokens with congruent and incongruent visual articulations of the utterances [18]. They observed that integration (the perception of the visual portion of the stimulus) typically occurred when listeners perceived the auditory stimulus as speech. This contributed to the interpretation that multisensory speech perception is special because it fundamentally differs in measureable ways, both behaviorally and/or in terms of neural activation patterns, from other cognitive processes.

Using sine wave speech as non-speech controls is vital for this study because it contains the same physical properties as the speech stimuli. The difference between the speech and non-speech conditions therefore hinges on the participants level of awareness; only in the speech condition are phonetic and language representations engaged to perform the task. Previous studies comparing differences in behavioral responses and neural activation patterns for speech and non-speech stimuli have utilized non-speech controls that have vastly different bottom-up properties than linguistic stimuli [13].

Capacity: Measuring Audiovisual Integration using Reaction Time

The ability to combine cues from different modalities can be assessed by using a measure of *capacity. Capacity* constitutes a cumulative measure of work competed in which audiovisual processing speed can be compared to the processing speed in the auditory and visual only conditions [7] [11, 28]. Capacity is assessed using the coefficient *C(t)* [7]. This constitutes a probabilistically defined RT measure, in which parallel independent processing establishes benchmark for multisensory enhancement. As we shall see, capacity compares the RT distribution from trials where auditory and visual cues are presented to the RT distributions from trials where either auditory-only or visual-only information is present; these latter "unisensory" cases constitute the parallel independent predictions. The capacity function uses the entire distribution of responses at the level of the integrated hazard function. The integrated hazard function is defined as:

$$H\left(t^*\right)=\int_0^{t^*} h(t)dt$$

where

$$h(t)=\frac{f(t)}{S(t)}$$

The term *f(t)* denotes the probability density function of RTs, and *S(t)* represents the survivor function—representing the probability that a response has not yet occurred by a certain time (*S(t)* = 1 – *F(t)*). Finally, *h(t)* gives the probability of a response in the next instant, given that a response has not yet occurred [28].

The use of capacity is advantageous compared to mean RTs or mean accuracy for several important reasons, making it an optimal measure to describe audiovisual integration. First, it uses integrated hazard functions. Hazard functions provide several advantages over means [12], and also better capture the notion of efficiency, and ultimately integration. Hazard function can be interpreted in terms of the instantaneous amount of work completed, and integrated (i.e., cumulative) hazard function used in the capacity measure indicate the total amount of work completed by a certain time. One exemplary motivation for using capacity in conjunction with neural-based measures has emerged in data showing that capacity is a superior predictor of the performance of neural circuits underlying memory (compared to mean accuracy) [46]. Wenger and colleagues carried out a study using capacity (RTs) to assess performance in an episodic cued memory. The authors observed that the performance of a computational model of a hippocampal circuit, as a function of different levels of degradation, was a superior predictor of capacity in normal aging individuals versus those with mild cognitive impairment or dementia of the Alzheimer's type.

Townsend and Nozawa [11] derived the benchmark capacity coefficient for tasks in which observers were presented with 0, 1, or 2 target stimuli and have to respond if either 1 or 2 stimuli are present. For present purposes, if we let $H_{AV}(t)$ denote the integrated hazard function obtained from audiovisual trials, and let $H_A(t)$ and $H_V(t)$ signify the integrated hazard functions obtained from the auditory-only and visual-only trials, respectively. The capacity coefficient, *C(t),* is defined in equation 1:

$$C(t)=\frac{H_{AV}(t)}{H_A(t)+H_V(t)} \qquad (1)$$

Yet another significant advantage of *C(t)* lies in the ability to distinguish certain types of processing. The term in the denominator corresponds to the predictions of a parallel independent race model [19]. First, deviations from *C(t)* = 1 indicate that independence has been falsified—this could be due to limitations in processing resources, cross-channel interactions [21] or co-activation [11, 19]. However, for co-activation to be diagnosed capacity must be much greater than a value of "1" [28]. The *C(t)* function therefore provides a useful non-parametric measure of integration efficiency in a variety of settings, with three possible outcomes and model-based interpretations.

1. *C(t)* can be greater than 1 at time *t*, indicating faster RT and thus more work completed in the audiovisual condition compared to the auditory- and visual-only conditions. This points to efficient integration since RTs in the

audiovisual condition are faster than would be predicted by independent race models.

2. *C(t)* can of course be less than "1" at a certain time point, pointing to slower RTs in the audiovisual condition compared to the unimodal conditions, and therefore inefficient audiovisual integration.
3. Third, *C(t)* can equal to 1 at time *t*, indicating that audiovisual recognition is neither faster nor slower than parallel independent model predictions.

1.1. Neural Assessments of Integration

1.1.1. Mean Global Field Power

The EEG measures will include analyses of mean global field power (mGFP) for each individual participant. mGFP represents the standard deviation across electrodes at a given time point [47]. In general, larger mGFP values at a given time point indicates the presence of a specific underlying neural component. An important aim of this study is to quantitatively and qualitatively relate mGFP to capacity. As one example, if *C(t)* >1, greater mGFP in the AV stimulus compared to the maximum of the A (or V-only) may be observed. Alternatively, capacity and mGFP may be inversely related.

1.1.2. Source Reconstruction

Cortical source reconstruction was carried out using Curry® 6.0 Neuroscan (TX) software in order to examine the extent to which auditory, visual, and audiovisual activation patterns conform to predictions regarding the spatial distribution of brain regions involved in processing of audiovisual stimuli. A major advantage of CDR analysis over most dipole reconstruction schemes is that it provides a more accurate description of the extent of activation (although cf. [15] [24] for discussion concerning shortcomings of using CDR analyses). The approach described here was designed to replicate the procedures implemented by Bernstein and colleagues for CDR of ERP data [15]. This was done specifically to identify neural activation patterns for audiovisual integration. For consistency between studies, a similar volume conductor specification (three shell spherical head model) was implemented.

The CDR activation patterns, shown later, display the results of a weighted group average of the ERP signals. CDR analysis was carried out by utilizing the "minimum norm least squares algorithm" to stabilize the solution. This particular criteria explains the field of activity through the source configuration that minimizes power. The minimum norm least squares algorithm allows for the detection of regularization parameters (inverse of the S/N ratios observed in the data, see e.g., [48]). The algorithm was applied to the ERP data across the interval of 52-220 ms because this time interval spanned the major activation peaks in the mean global field power that appeared in each of the stimulus types. Results were computed at 8ms time steps. The volume conductor specification included a three shell spherical head model with an outer radius of 9 cm. Electrode positions on the spherical head model were estimated using the 3-dimensional coordinates of a GSNHydrocel 128 electrode array which were input into Curry. Results and analyses of activation patterns were confined to a segmented brain provided by Curry (see also [15, 49]). Finally, solutions were obtained using a common threshold across each condition.[1]

2. MATERIALS AND METHODOLOGY

The following study shall explore integration more thoroughly by taking into account statistically motivated behavioral measures in conjunction with EEG data. This experiment utilized an identification task employing sine wave and point light stimuli that one can perceive as either speech or non-speech auditory and visual events. This study will allow for the direct comparison of multisensory processing for speech, to integration for controlled non-speech stimuli. Behavioral analyses of integration (*C(t)*), combined with source localization methods (current density reconstruction, CDR, from ERP data) similar to Bernstein *et al.* shall be carried out [15]. CDR and EEG data will provide a valuable time based neural measure that may serve as a covariate of behavioral measures such as *C(t)*.

2.1. Participants

Data were obtained from four right-handed college-aged participants (2 female) with a mean age of 22 were recruited from The University of Oklahoma, Norman campus. Participants were presented with eight blocks of auditory-only, visual-only, and audiovisual trials over the course of four separate days within a one week period. Each experimental session (two blocks per day) lasted approximately 45-50 minutes. All of the participants reported having normal-hearing and normal or corrected vision. Participants were native speakers of American English. This study was approved by the University of Oklahoma Institutional Review Board (IRB).

2.2. Stimuli

2.2.1. Visual Component

The visual portion of the stimulus consisted of *point light* digitized videos of two female talkers saying the syllables /be/ (pronounced "bay") and /de/ ("day"). Point light displays consist of green fluorescent dots arranged on the talker's articulators (e.g., teeth, jaw, and facial muscles), and appear similar to a pattern of moving dots [22]. Participants who are unfamiliar with point light displays perceive them as a moving pattern of dots, and not a talker saying a word until they are informed about the true content of the stimuli. None of the four participants were aware that the pattern of dots contained linguistic information at the onset of the study.

The videos were recorded at a rate of 30 frames/second with an approximate duration of 450ms. In order to maintain the ecological relation between visual speech and the auditory signal, visual movement began before the onset of sound (approximately 30 ms). The auditory signal lasted for 270 ms for both stimuli. Each of the stimuli was obtained from the Hoosier Multi-Talker Database [23]. The words

[1] Modeling from previous research has indicated that the source generators in the cortex resonsible for the signals of interest were probably less dispersed than the results displayed in the CDR activity [15]. To accommodate this possibility, it has been suggested that results should be described in terms of the regions showing greatest modeled current density.

were chosen because the stimuli were perceptually similar enough under degraded conditions to produce some errors, but distinct enough to yield high overall accuracy (> 95 % correct). The videos for the two types of stimuli were qualitatively similar, but perceptually distinguishable, particularly in the stimulus onset due to difference in place of articulation between “b” and “d”.

2.2.2. Auditory Component

The auditory component of the files was processed as sine wave speech [17]. Sine wave speech uses sinusoids to approximate the amplitude and frequency of natural speech, but removes the acoustic components characteristic of normal speech signals. Before being informed that sine wave speech is speech, listeners typically report hearing “computer beeps”. However, after being informed that the signal is speech, the linguistic content of the utterance can typically be decoded.

2.3. Procedure

Each subject participated in two within-subject conditions: A non-speech condition, and a speech condition. In the non-speech condition presented over the first two days, listeners were informed that they would be presented with two distinct moving dot and sound patterns. Participants were seated 76 cm from a computer monitor with their chin placed comfortably in a chin rest. Each trial began with a fixation cross appearing in the center of the computer monitor for a random duration on a uniform distribution ranging from 400-700 ms followed by an auditory-only, visual-only, or audiovisual stimulus. Listeners participated in two experimental blocks per day with a brief break in between. Each block consisted of 60 auditory-only, visual-only, and audiovisual trials (30 trials from “bay” and “day”) for a total of 180 trials per block, and 360 trials per day. Over the course of the four-day experiment, participants were presented with a total of 1,440 trials with 240 A-only, V-only and AV from the “non-speech” condition (days 1 and 2), and 240 of each trials type from the speech condition (days 3 and 4). The task required listeners to make categorization judgments in trials in which auditory-only, visual-only, and audiovisual information was presented. The configuration corresponding to the syllable “be” was labeled as “A” (the left mouse button), while the syllable “de” was labeled as “B” (right mouse button). Feedback was provided after each trial (“correct” vs. “incorrect” appearing in the center of the computer monitor for a total of 1,500 ms), and participants were presented with 48 practice trials that were not included in the subsequent data analysis at the beginning of each day. Listeners were instructed to make their response by clicking the appropriate button on the mouse as quickly and as accurately as possible.

In the second study phase occurring on days three and four, participants were informed that the stimuli were actually talkers speaking the words/syllables “bay” and “day”. The experimental protocol on days 3 and 4 was identical to days 1 and 2, the only difference was that the participants were informed that the stimuli were speech, and response mapping were relabeled such that the left button became “bay” and the right button became “day”. In the speech condition, participants were required to make a button press response corresponding to the word they thought the talker said based on auditory-only, visual-only, and audiovisual speech information. After the initial practice session, each participant reported perceiving both the auditory sinewaves and visual point lights as the spoken words “bay” and “day”. Listeners were provided a questionnaire asking what they thought they heard during the first two days of the study. None of the participants reported hearing speech sounds on the first two days of the experiment. Instead, each reported hearing a series of “beeps” or “computer” sounds.

The sequence of events, including the stimuli and trial structure, are shown in (Fig. **1**).

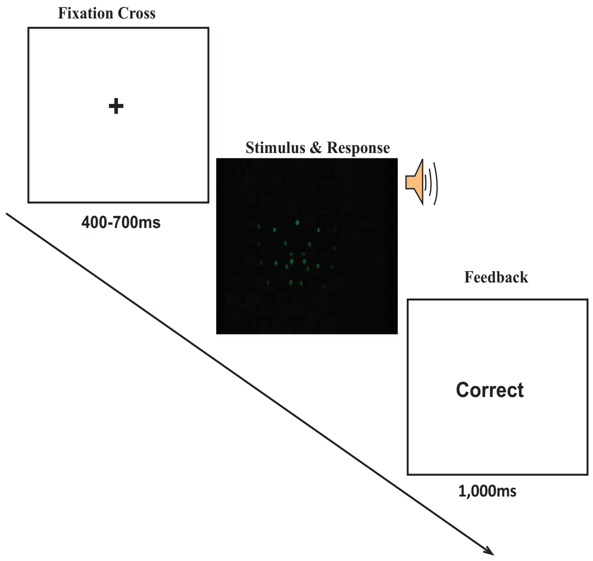

Fig. (1). This figure shows the trial structure, including sample stimuli.

2.4. Electrophysiological Recordings

EEG recordings were obtained from a high-density 128-channel electrode array placed on the participant’s scalp (NetStation; Eugene, OR). EEG data were collected on each of the four testing days. This was done to improve the same sample size in the data by increasing the number of trials, and thus, the signal-to-noise ratio in the EEG data. Recordings were re-referenced to the Cz electrode in the center of the scalp. One advantages for using Cz as a reference electrode is that it provides equal distances across left and right hemispheres. Eye blinks were monitored with two electrodes placed above and below the eye. Continuous data were recorded and sampled at a rate of 1 kHz. Two electrodes, one located under each eye monitored eye movements, and a set of electrodes placed near the jaw were used for off-line artifact rejection. Channel impedances were maintained at 50 K Ohms or less throughout each session.

After recording, the data were down sampled to 250 Hz. Noisy channels are typically identified by visual inspection and removed; however, none were removed in the course of this study. Trials with ocular artifacts such as eye blinks were removed automatically using a statistical thresholding function in EEGLab (http://sccn.ucsd.edu/eeglab/); over 90% of trials were retained in each condition. Baseline correction was carried out using an interval of 100 ms prior to the onset of the stimulus in each condition. Data were averaged into epochs across participants (100 ms prior to stimulus onset until 600 ms post stimulus) and organized by stimulus category: A-only (speech), V-only (speech), AV (speech), A-only (non-speech), V-only (non-speech), and AV (non-speech). In the A condition, the epochs were aligned with the onset of the auditory stimulus, in the V condition, with the onset of the visual stimulus, and in the AV, with the onset of the visual stimulus component. The waveforms were filtered from ½ to 30 Hz using an FIR filter in EEGLab. Data from each participant was used to generate the grand averaged waveforms across right and left parietal and temporal electrodes.

3. RESULTS

Mean accuracy scores were high overall, as predicted. The mean percent correct for the audiovisual speech condition was 95%, and the auditory-only mean accuracy was 97%. Visual-only accuracy was lower with an overall mean accuracy of 72%, reflecting the inherent difficulty associated with visual-only processing, especially under degraded conditions [25, 26]. The non-speech categorization conditions elicited similar accuracy scores for the audiovisual trials (95% correct), and auditory-only trials (96% correct). The mean accuracy for the visual-only trials was once again lower (61% correct), although accuracy was significantly above chance. A paired samples t-test was carried out to assess whether visual-only processing was significantly better in the speech versus the non-speech condition, and we observed a trend toward that effect (t(3) = 2.24, *p* = .115, SD_{Speech} = .16; $SD_{Nonspeech}$ = .15). While the results from the statistical test were non-significant, perhaps due to the small sample size, there is some evidence that linguistic information associated with the point light (i.e., when they are perceived as speech) displays may facilitate accuracy.[2] Next, the RT and capacity results shall be discussed before investigating the neural activation patterns in the speech and non-speech conditions.

3.1. Capacity

Fig. (**2**) shows the actual capacity values (y axis), calculated from the RT distribution using Equation 1, across the range of RTs (x-axis). Capacity is shown separately for each of the four participants, where each time bin was 5ms. Interestingly, *C(t)* generally indicated limited capacity for both speech and non-speech conditions, with values hovering around ½ or slightly above for a large range of RTs [7, 11, 27]. This result points to sluggish RTs in the audiovisual condition compared to parallel race model predictions, perhaps implicating limitations in neural resources or inhibitory cross-channel connections between auditory and visual circuits [7, 9, 11, 21, 28]. Other research on audiovisual recognition under high accuracy settings has shown that listeners typically fail to benefit from the visual signal in terms of RTs, and as a result, often exhibit limited capacity except when the auditory S/N ratio is low [7, 9].

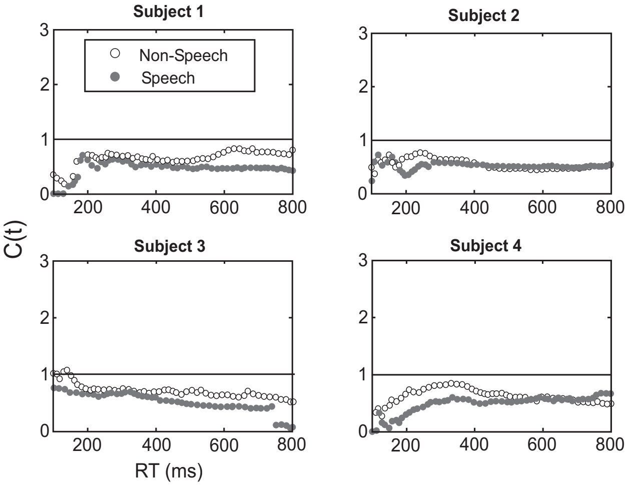

Fig. (2). Capacity values shown separately for each of the four participants. Capacity is displayed for both speech and non-speech trials in each panel. The closed circles represent the capacity values obtained from the RT distributions in "speech" trials, and the open circles represent the values obtained from the distributions from the 'non-speech' trials. Qualitatively, each participant showed a similar capacity profile, namely limited capacity for multiple time points (*C(t)* < 1 for both speech and non-speech conditions).

A noteworthy result from the capacity analysis was the qualitative similarity between the speech and non-speech results. For both conditions, capacity was limited for all participants at most time points, indicating the lack of ability to take advantage of the visual signal in the time domain. This pattern of results was consistent with previous findings showing similar patterns of audiovisual gain (or lack thereof), between speech and non-speech signals [13]. A paired samples t-test comparing peak capacity values across participants did indicate overall higher capacity values for the non-speech relative to the speech condition (t(3) = 4.92, *p* = .016). This suggests a greater audiovisual disadvantage in the RT domain when the stimuli were interpreted as speech rather than simpler acoustical events. Interestingly, Stevenson and James [13] showed in an fMRI study that speech and non-speech objects elicit similar levels of inverse effectiveness in the BOLD signal (and accuracy), although on the other hand, portions of the pattern of neural activation were non-overlapping. This latter possibility will be further explored in the CDR results.

3.2. Mean Global Field Power

Mean global field power (mGFP) represents a measure of spatial standard deviation, where high global field power indicates similar fields or activation patterns [29]. The mGFP is displayed for the audiovisual, auditory, and visual-only conditions separately for each participant in (Fig. **3A**), while the composite results averaged across participants are shown in 3B. The results from the condition in which the participants were making speech judgments are shown in the left panels, while the right panels show results from the non-speech categorization judgments. Each condition produced peak activations, with the onsets for the slopes to the largest peaks generally beings observed immediately subsequent

[2] An alternative possibility is that observed differences might be associated with practice.

to100 ms post stimulus onset. This indicates that neural activation patterns responsible for producing mGFP peaks likely occurred in the 52-120 ms post stimulus interval. This interval is indicated by vertical lines in Figure 3B. Qualitatively, the results in (Fig. **3B**) indicate that the AV mGFP peak is suppressed relative to the A-only in the 52-120 ms time window.

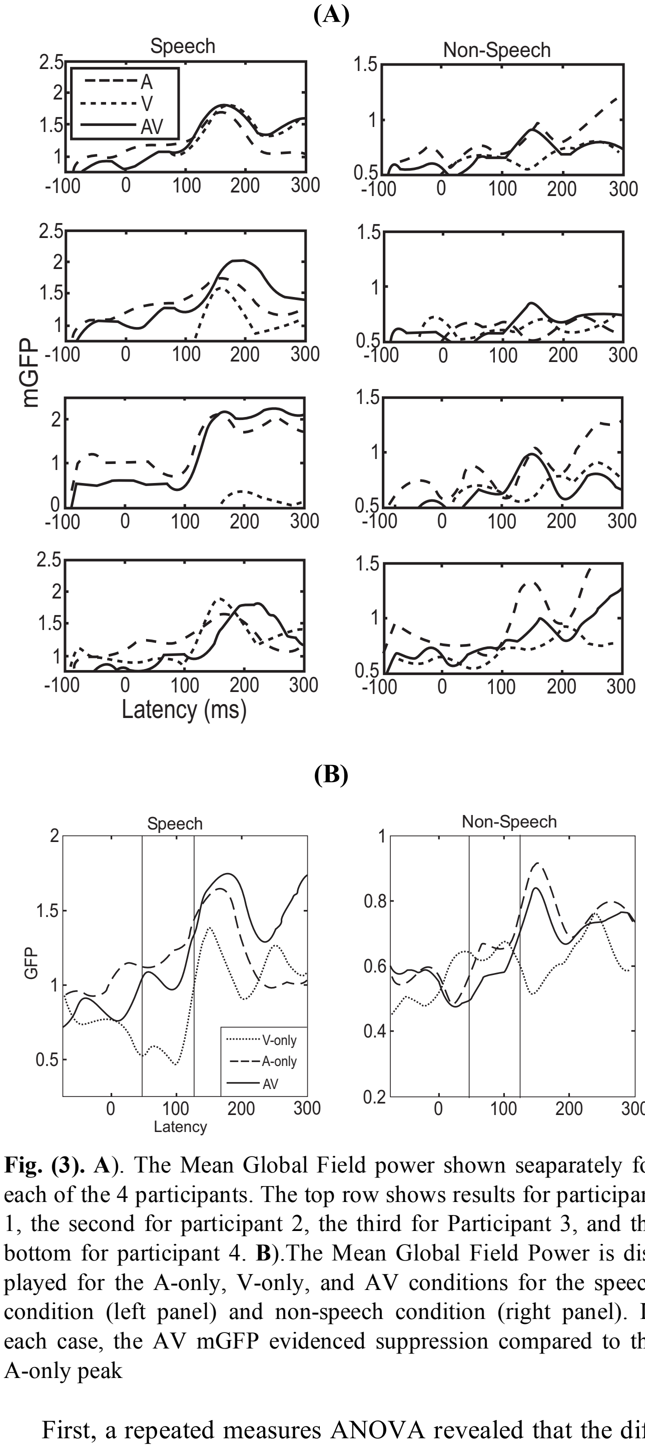

Fig. (3). A). The Mean Global Field power shown seaparately for each of the 4 participants. The top row shows results for participant 1, the second for participant 2, the third for Participant 3, and the bottom for participant 4. **B**).The Mean Global Field Power is displayed for the A-only, V-only, and AV conditions for the speech condition (left panel) and non-speech condition (right panel). In each case, the AV mGFP evidenced suppression compared to the A-only peak

First, a repeated measures ANOVA revealed that the difference in mGFP peaks between the AV and maximum of the unisensory mGFP peaks (AV vs. max{A, V}) marginally differed between the speech and non-speech conditions. Specifically, the mGFP in the speech condition showed evidence for greater multisensory suppression compared to the non-speech in that time window (F(1, 3) = 8.41, *p* = .06). Second, a Pearson correlation was carried out between multisensory gain, as measured by mGFP in the time window specified above, and maximum capacity values (an established behavioral index of integration [7]) across speech and non-speech trials. As predicted, the overall change in mGFP was significantly positively correlated with maximum capacity scores (r(7) = .84, *p* = .008). This result strongly suggests that capacity constitutes a useful behavioral index of integration efficiency in the neural domain. The upshot of these results is that they point to a systematic relationship between *C(t)* (efficiency in the behavior domain) and mGFP.

3.3. Cortical Source Activity: Speech

The results for the CDR activation patterns (shown in Figures 4 and 5) are purely qualitative. Fig. (**4**) displays the source localization results in terms of CDRs. The left hemispheric results are shown in (Fig. **4A**), and the right in (Fig. **4B**) (results are collapsed across consonants "b" and "d" categories as in the other analyses). The results are plotted for the auditory-only, visual-only and audiovisual conditions and are plotted in 8 ms time steps, beginning at 52 ms through 100 ms, and in 20 ms time steps from 120-220 ms [15]. The data represent snapshots of single frames at the specified time points. Down sampling prevented the results from relying on high frequency potentials, and the CDR data were generally stable and showed continuity over time.

3.4. A-Only Activity

Temporal activation developed in the cortex between 52-60 ms, and included left hemispheric regions including primary auditory cortex, inferior prefrontal (e.g., Broca's area) and more posterior regions along the sylvian fissure (e.g., around Wernicke's area). Activity persisted in various temporal regions, including inferior prefrontal cortical areas, until approximately 100 ms (Fig. **4A**). Evidence for early activation was observable bilaterally in right temporal regions, but the activation failed to persist for more than 8-16ms (Fig. **3B**). Therefore, the CDR analysis indicates that the peaks in the mGFP (both prior and perhaps subsequent to 100 ms) of the auditory-only signal were mainly driven by left parietal/temporal cortical activity rather than right hemispheric activation. Overall, activation was more extensive and widespread in the left cortical regions compared to right. In summary, auditory activation appeared extensively, although briefly, across frontal and parietal areas. Early activation also spread to speech processing areas, including left frontal/temporal regions that may correspond to Broca's area. Major activation in both hemispheres appears to have been resolved at or around 100 ms (and the right hemisphere prior to 76 ms).

3.5. V-Only Activity

The visual-only CDRs suggest consistent activation of diverse processing regions ranging from inferior frontal to the anterior occipital lobe. Activation was observed in areas ranging from visual processing regions in the anterior occipital cortex to temporal association areas. Interestingly, the data show evidence for activation in both left parietal/temporal and inferior frontal activity, adding support to

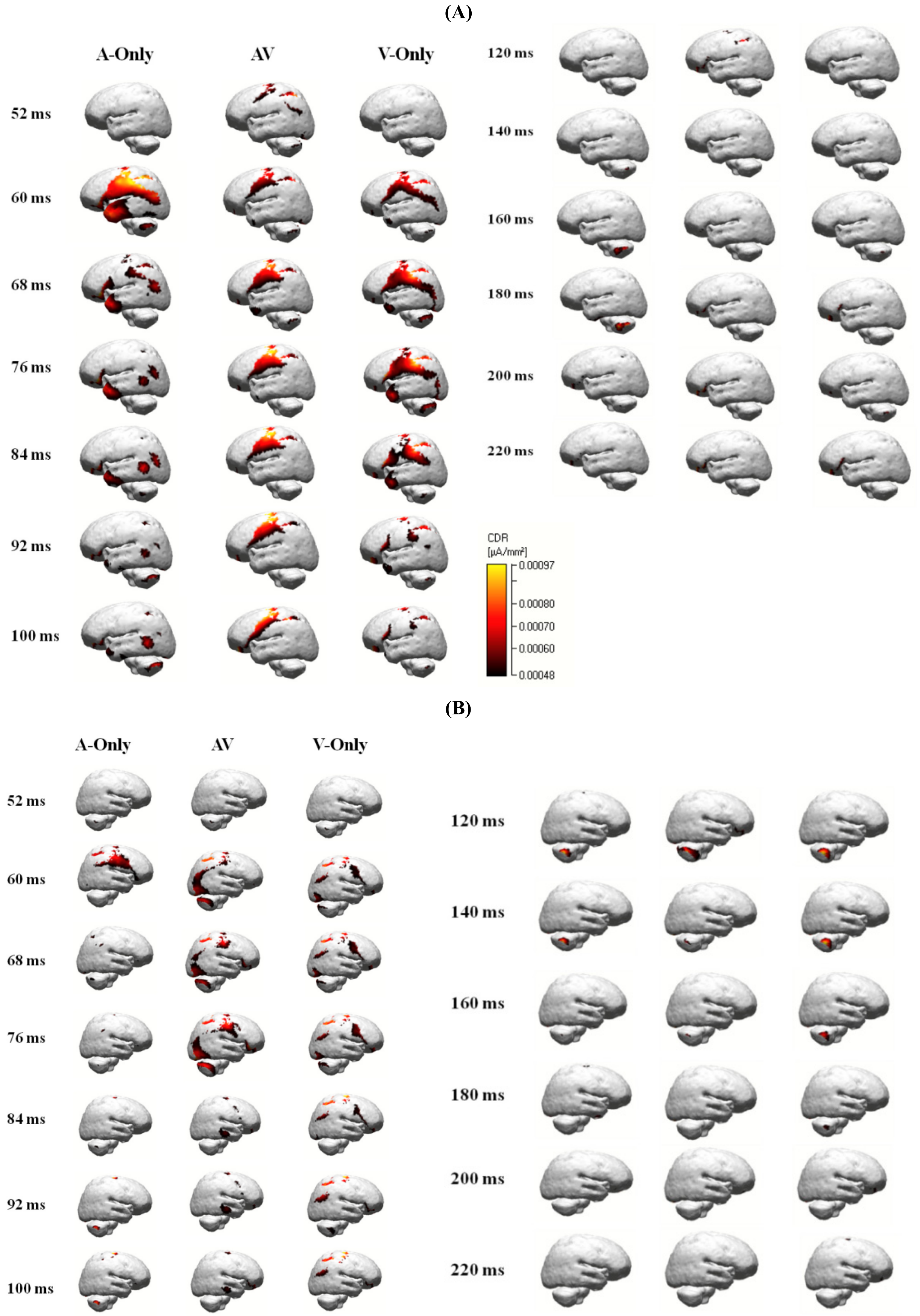

Fig. (4). Figure showing the dynamic CDR activation for the speech condition in 8ms time intervals from 52 to 100ms post stimulus and in 20ms intervals from 100 to 220ms post stimulus. Left hemispheric activation is shown in Fig. **3A** and right hemispheric in **3B**.

the hypothesis that visual speech may activate language pathways. These findings are therefore consistent with previous fMRI research showing that silent lip-reading can activate portions of the auditory cortex in language perception tasks [30] (although see [31]). The band of left cortical activation, ranging from the anterior occipital region to inferior frontal, remained consistently active until approximately 100 ms. This mirrors the time course of activation from the auditory-only condition, although the CDR showed more extensive activation in response to the visual-only condition in occipital cortical regions. The visual-only activation was sparser in the right anterior cortical region, but did show activation patterns across parietal (sensory-motor), temporal, posterior frontal, and occipital areas (Fig. **4B**). The posterior frontal activity persisted through 84 ms, while occipital and some temporal activity persisted through 100 ms. The overall pattern of activity was similar across the left and right cortical lobes, although more extensive areas were activated in the left hemisphere. Similar to auditory-only activity, activation in the visual-only condition appears to be resolved around or slightly post 100 ms.

The CDR results are also supported by the mGFP analyses (left panel Fig. **3B**) showing a pattern of peak activation consistent with the auditory and audiovisual signals. Additionally, the accuracy levels in the visual-only condition indicate that participants were capable of processing the stimuli as language even though accuracy was lower compared to the auditory-only and audiovisual conditions.

3.6. AV Activity

Audiovisual activity in the left hemisphere region indicates reduction in activation compared to the auditory and visual-only conditions. Recall that consistent vidence for limited capacity ($C(t)$ < 1) emerged (Fig. **2**), pointing to the possibility that inhibition from visual brain regions constitutes this source of suppression. The greater extent of left hemispheric activation (CDR analysis) indicates that the mGFP was mostly driven by left hemispheric activity. One may observe additional evidence for suppression in the mGFP signals, where the peak from the audiovisual visual signal was suppressed compared to the auditory signal for early times (~60 ms) and the visual signal for later times. The audiovisual activity shows activation in areas consistent with the inferior parietal sulcus (IPS) and parietal cortex, including supramarginal gyrus (SMG), beginning approximately 52-60 ms post-stimulus. The activation in these regions was consistent and spread to areas encompassing the inferior frontal (Broca's area) and areas immediately superior to the sylvian fissure regions. Significant activation in these areas was generally consistent and persisted before dissipating around 100-120 ms. Activation was also observed superior to the sylvian fissure beginning approximately 68 ms in parietal brain regions.

The fact that significant activation was not observed in the STS itself could have resulted from multisensory suppression or inhibition. This hypothesis seems to be supported by the fact that auditory-only activation was observed in anterior regions of the STS from approximately 70, through 100 ms. Alternatively, it is possible that circuits in STS may have only been marginally involved in the processing of multisensory speech stimuli. Significantly, parietal activation for AV stimuli were more extensive in the temporal and IF and SMG regions, and lasted far longer especially compared to auditory-only processing. This implicates the involvement of these circuits in speech integration. One interpretation is that these circuits may have received facilitatory information from visual association areas while temporal sites may have been inhibited (later activity (> 120 ms) appeared minimal).

Right hemispheric results show evidence for a variable pattern of activation that is, not surprisingly, slightly different from the left hemisphere. Small regions of activation began to appear in parietal, temporal, and posterior frontal regions beginning around 60 ms post stimulus. Evidence for minimal activation persisted up until 92-100 ms. The main difference was that visual association areas in the posterior cortex evidenced the greatest and most persistent activation. The right hemisphere showed evidence for occipital activity and bilateral inferior frontal activation beginning around 60 ms and lasting through 76 ms. Overall, a broad band of activation from anterior to posterior regions was observed for auditory, visual, and audiovisual speech stimuli.

3.7. Cortical Source Activity: Non-Speech

Fig. (**5**) displays the source localization results in terms of current density reconstructions (CDRs). The left hemispheric results are shown in (Fig. **5A**), and right hemispheric results are shown in 5B. Once again, the results are collapsed against the /be/ "bay" and "day" /de/ categories.

3.8. A-Only Activity

Perceiving the auditory stimuli as a series of beeps rather than speech yielded an activation pattern that was qualitatively similar to the speech condition. In (Fig. **5A**), activation appeared across posterior temporal processing as well as parietal areas, beginning approximately 60 ms. The pattern of activation remained consistent, and various regions in the posterior frontal and temporal parietal cortices remained active until gradually dissipating around 100 ms (although significant temporal and inferior frontal activity appeared approximately 160-180 ms). Interestingly, non-speech stimuli evoked broader areas of posterior temporal activation for auditory non-speech stimuli compared to speech stimuli. This may implicate the involvement of more specialized circuitry (e.g., inferior frontal regions) for speech stimuli. Somewhat surprisingly, activation in the primary auditory cortex appeared to be virtually absent, and only appeared for early processing times in (Fig. **4A**). Presumably, the auditory cortex was a less crticial processing center compared to other areas, such as inferior frontal and posterior areas of the sylvian fissure. Primary auditory cortex (A1), for example, is involved in the processing of frequency information. Considering the task demands, and complexity of the stimuli then, it is not entirely unusual that A1 (and V1 for that matter) played a relatively minor role relative to certain association areas and more specialized language zones.

Bilateral right hemispheric activity was observed in the temporal cortex (including right STS) and surrounding areas including frontal regions, SMG, and extended into temporal association areas. Overall activation was sparser in the right hemisphere, although the extent of activation was more pervasive and longer lasting in the non-speech relative to the

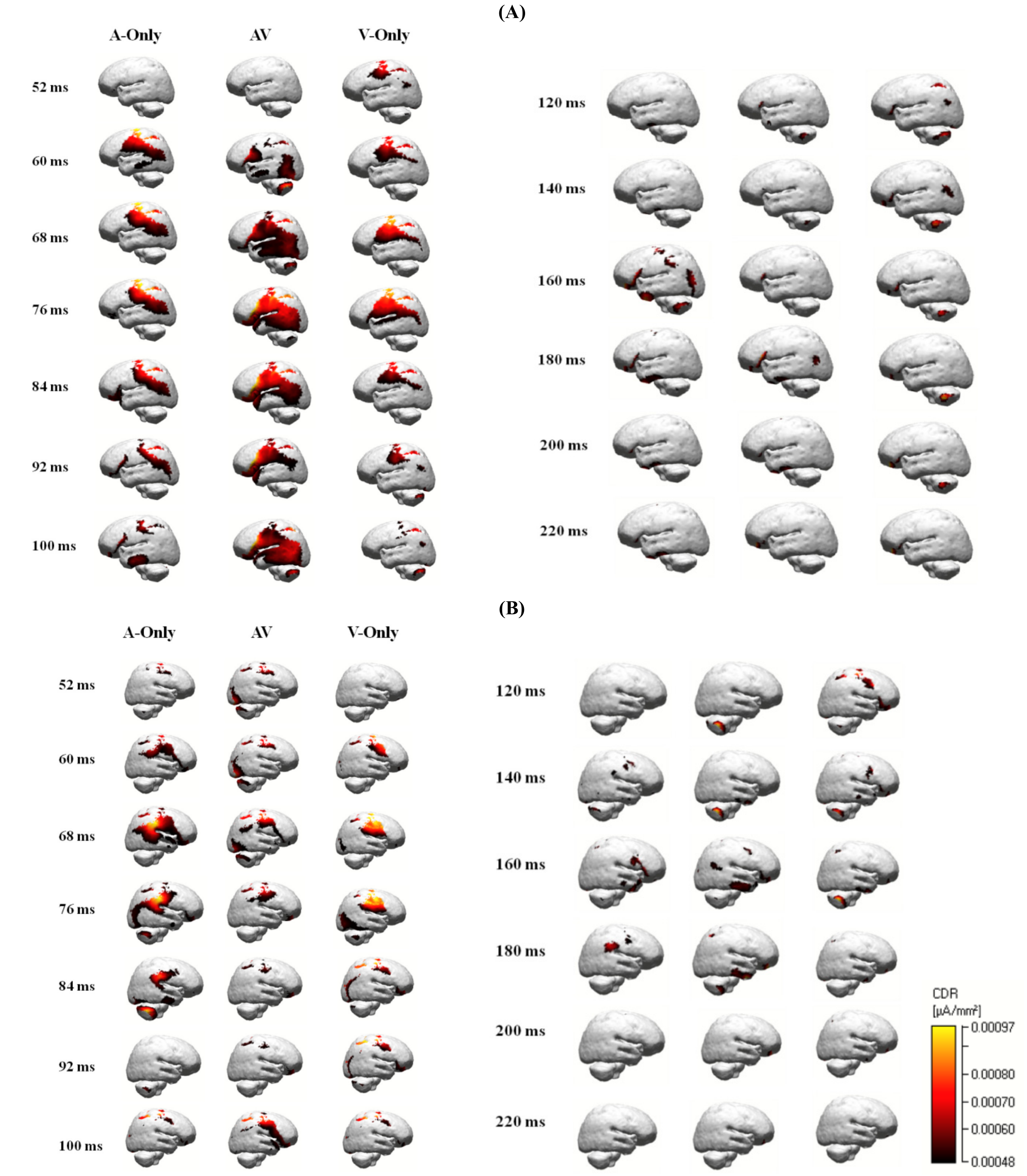

Fig. (5). Figure showing the dynamic CDR activation for the non-speech condition in 8ms time intervals from 52 to 100ms post stimulus and in 20ms intervals from 100 to 220ms post stimulus. Left hemispheric activation is shown in Fig. **4A** and right hemispheric in **4B**.

speech condition. This suggests that the perception of environmental sounds may recruit broader cortical circuitry. The extent of bilateral activation suggests that the peaks observed in the mGFP (right panel Fig. **2**) were driven by both right and left hemispheric activity. However, the mGFP was noisy for early latencies, perhaps indicating broad activation patterns.

3.9. V-Only Activity

Left hemispheric visual-only CDRs from the non-speech condition indicate an activation pattern similar to what was observed from the speech condition. Once again, activation above threshold was observed in visual cortical areas in portions of the occipital cortex beginning approximately 52-60 ms post stimulus. Activation was stable and consistent, with some small patterns of activity in parietal and occipital regions persisting through 140 ms. Activation was observed in the vicinity of areas consistent with auditory processing areas beginning approximately 76 ms, although activity around these particular areas was less pervasive compared to the speech condition and began dissipating at 92 ms. Nonetheless, it does indicate that the categorization of non-speech visual patterns can activate some auditory circuitry, especially when the visual dots have been associated with sound patterns after practice. Right hemispheric activation may be observed around right parietal, temporal association, and occipital areas. Activity in these regions was observable from approximately 60 ms (76 ms for occipital and temporal association areas), with minimal activity present through 140 ms. Activity in occipital cortices was present from 52-60 ms and began dissipating prior to 100 ms.

The CDR data pattern was consistent with the proposition that the peaks in the mGFP of the visual-only signal were driven by both left and right hemispheric activity. Finally, the mGFP for both auditory and visual-only non-speech conditions was also considerably lower than the speech conditions, which could reflect more extensive, less specialized, and non-overlapping cortical circuits involved in processing of non-speech stimuli.

3.10. AV Activity

The CDR analysis of the audiovisual non-speech trials revealed a broad pattern of left hemispheric activation. A broad band of activation emerged in temporal, inferior frontal regions, and parietal/occipital regions began at 60 ms. Beginning at 68 ms, more intense activation began to develop along inferior frontal areas, parietal areas, temporal cortices along the sylvian fissure including primary auditory cortex, the SMG, and classical multisensory areas such as the STS and STG. The activation was stable before beginning to dissipate around 100 ms post stimulus. Interestingly, the CDR activation pattern exhibited a general sequence of activation similar to the audiovisual speech condition, although the extent of the activation is considerably broader. Once again, this implicates broader and non-overlapping cortical structures involved in multisensory integration of non-speech stimuli, perhaps due to the engagement of less specialized circuits. Lastly, audiovisual activation, though broader, was more suppressed compared to the unisensory activation for early processing times. This might be reflective of the fact that audiovisual integration requires the recruitment of more cortical resources, which requires more time [32]. Right hemispheric activation appeared in occipital and right parietal/temporal location beginning from 52-60 ms. The activity was stable in most areas past 100 ms, although considerably less extensive than the left hemispheric region. From 68-76ms (and again at 100 ms), activation expanded to include regions in the parietal and inferior frontal cortices. Overall, right hemispheric activation was less extensive compared to the left further implicating left cortical structures in the integration of multisensory stimuli, particularly in right handed individuals.

In summary, activation appeared more spatially extensive in both right and left hemispheres in the processing of multisensory non-speech stimuli compared to speech. As predicted, our results show a similar pattern of multisensory enhancement in terms of behavioral RT measures of efficiency (i.e., *C(t)*), but with the recruitment of both overlapping and non-overlapping cortical structures. This implicates the recruitment of broader neural resources that are less specialized in multisensory non-speech recognition, as predicted. According to this hypothesis, qualitatively similar levels of integration efficiency might be achieved in speech versus non-speech stimuli via a tradeoff that involves greater recruitment of language processing areas for linguistic stimuli, and more diffuse cortical regions for non-speech integration. This hypothesis was further supported by mGFP analyses showing higher A and AV peaks in the speech compared to the non-speech condition, and hence, a stronger underlying component devoted to linguistic processing.

We now discuss our results within the context of a general framework for proposed circuitry in multisensory integration. Adjustments to the framework may be required to account for differences between speech, non-speech, and other types of stimuli.

4. DISCUSSION AND CONCLUSION

These data replicated several important findings implicating the involvement of multiple brain regions, in addition to the left SMG and STG, in audiovisual integration [15, 33].[3] Similar to Bernstein *et al.*'s observations [15], support for left frontal/parietal activation was obtained for early latencies in response to both speech and also non-speech stimuli. STS activation was not the earliest and did not occur with the greatest intensity in response to either type of AV stimuli. Brief left anterior STS activation was observed for auditory stimuli, although both anterior and posterior STS activation was observed for audiovisual (non-speech stimuli), and for a longer duration.

To summarize, differences in activation patterns between the auditory and visual-only conditions indicate that multiple brain circuits underlie phonetic processing. Critically, these processes engage circuits in diverse regions, ranging from the pre-frontal cortex to visual processing areas in the occipital lobe (see [15, 41]). In particular, greater and more extensive posterior and anterior occipital activation was observed in the visual compared to the auditory-only conditions, which was generally inferior frontal and temporal/parietal.

[3] Peak mGFP activation (~150 ms) unexpectedly occurred subsequent to the maximum activation in the CDR model (~ 60-100 ms). A possible explanation is that the CDR dispersion was affected by noise. Even though our data sets included several hundred trials, the signal averaging may have induced noise resulting in a lower S/N ratio than observed by Bernstein *et al.* [15]. This leads to two viable suggestions for future studies. First, include a larger number of subjects with fewer trials (and average the data). Alternatively, use a small sample size as was done in this study, but include several thousand trials and focus on reporting data for each individual participant.

Consistent with the results of other tasks requiring active perception and responses, activation was observed in temporal regions. One caveat relates to the observation that the activated circuits for visual speech and non-speech processing showed a high degree of similarity. This observation was consistent with the hypothesis that similar brain regions may become activated in response to visual language and complex non-speech stimuli [21]. As an illustration, the STS itself responds to a variety of multisensory stimuli such as complex non-speech gestures, as well as patterns of moving dots (Figs. **3** and **4**) [34]. Puce and colleagues [34], for example, observed that bilateral activation in posterior brain regions, including the posterior STS, in response to moving eyes and mouths (non-speech gestures). Conversely, stimuli with vastly different spatio-temporal dynamics, such as moving checker patterns, failed to activate the STS or surrounding areas.

In light of these findings and previously reported data (e.g., [15]), a generalized framework of audiovisual integration for complex stimuli shall be discussed in what follows.

4.1. Proposed Circuitry of Audiovisual Integration

This study builds upon previous findings by now comparing CDR activity in response to stimuli that participants interpreted as speech, and control stimuli employing identical dynamics that listeners did not interpret as speech. Interestingly, the results reported in this study provided exploratory evidence for the involvement of non-overlapping brain regions in the integration of speech and non-speech stimuli. This result was expected in light of findings from previous neuroimaging studies comparing activation patterns for speech and non-speech stimuli [13]. Crucially, while speech and non-speech AV stimuli elicited early frontal activity as well as activity around the primary auditory cortex, the AV non-speech stimuli yielded a greater range of peak activation in the proximity of the superior temporal sulcus and the lower/posterior portions of the temporal lobe. Interestingly, while non-overlapping circuits in key association regions were implicated in the integration of speech and non-speech stimuli as predicted (see [13]), the capacity analysis was consistent with previous accuracy data showing qualitatively similar multisensory benefit for speech stimuli and non-speech stimuli such as However, the capacity data were sensitive enough to demonstrate quantitative differences between speech and non-speech stimuli.

ERP studies comparing behavioral responses (RTs) with scalp map activations have provided evidence for early audiovisual neural interactions in the posterior cortex in conjunction with violations of the race model inequality, most likely resulting from facilitatory interactions [35]. Nonetheless, comparing speech data with studies that employed poorly controlled non-speech stimuli can be highly problematic simply because the surface level visual and acoustical properties of non-speech stimuli naturally differ from the auditory and visual components associated with language (see [13, 19, 20, 35]).

The observed spatio-temporal activation patterns for audiovisual speech and non-speech stimuli involved multiple cortical association areas, including some that displayed simultaneous activity. One of the broader implications of our findings, in terms of the hypothesized circuitry of integration, is that the STS contains only one portion of the association areas of the brain circuits critical for the dynamics of multisensory integration [15]. Nonetheless, previous research has implicated the STS the main brain area underlying audiovisual convergence [13, 33, 36] (see also [37]). Significantly, neuro-imaging studies of speech perception and the McGurk effect have failed to implicate the STS as a major center for multisensory processing [38]; the CDR results in this study, while exploratory, may yet indicate that the STS plays a less major role in such perceptual processes.

Fig. (**6**) provides a proposed diagram of the neural circuitry hypothesized to underlie the dynamics of synchronous

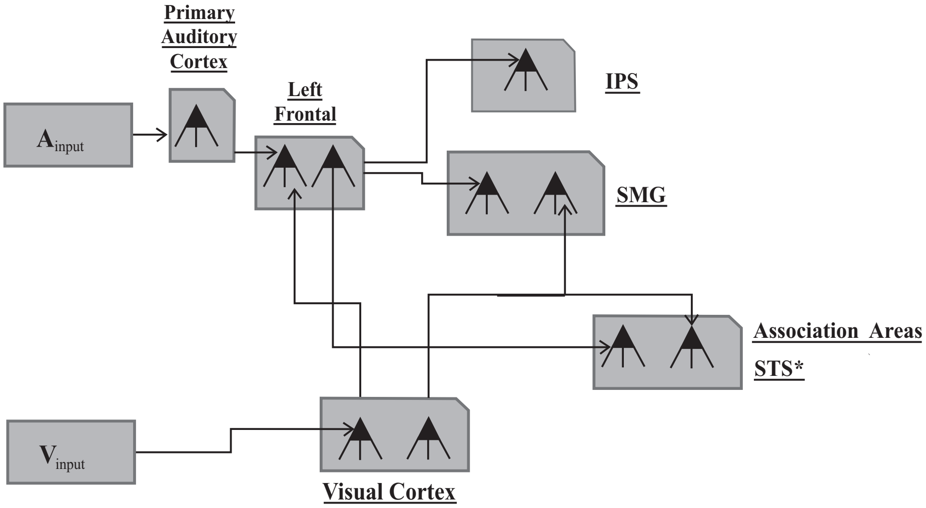

Fig. (6). A diagram of a general framework of integration. Spatiotemporal differences in integration may occur depending on stimuli. Auditory and visual inputs enter segregated pathways and are processed in parallel. Visual speech interacts with auditory information in early processing stages in left temporal and posterior frontal regions (and possibly circuits associated with the STS). Arrows indicate proposed uni and bi-directional connections.

audiovisual integration.[4] Rather than being initiated in the STS, early stages of audiovisual convergence (around 60-80 ms) appear to be resolved first in (inferior) frontal language zones, and subsequently, in temporal and parietal association areas such as SMG and perhaps STS. Studies using intracranial recordings provide evidence for similar early activation in the primary auditory cortex, although conscious stimulus discrimination probably does not begin to unfold until after 100 ms post stimulus [39]. Posterior left frontal AV activity persists longer than the auditory-only condition, implicating the involvement of these regions in integration. Other key areas including the interparietal sulcus (IPS) and supramarginal gyrus (SMG)/angular gyrus (AG) were activated concurrently to or immediately following activation in frontal and temporal regions likely via feed-forward connections in response to both speech and non-speech stimuli. As the results suggest, the foci of these activations may be affected simply by whether or not the inputs are cognitively interpreted as speech. More extensive activation in posterior temporal and parietal regions (e.g., including the STG/AG and SMG) appeared to occur for non-speech inputs. However, while the extent of the CDR activation tended to be more focused and less dispersed in response to speech stimuli, the degree of activation was generally greater. Finally, the left STS (superior to the STS in the case of speech stimuli) began to show some evidence for activation immediately subsequent to the aforementioned regions starting around 68 ms and persisting for tens of milliseconds.

Significantly, emerging evidence implicates earlier and more temporally persistent integration sites than STS, including SMG, and greater temporal/parietal circuitry [15]. During sensory interactions, visual activation appeared to spread from visual areas in the anterior occipital region to parietal regions superior to the sylvian fissure. Broader temporal regions were involved in non-speech integration. In short, the pattern of activation was consistent with the hypothesis that segregated auditory and visual inputs undergo processing in their respective cortices before visual information interacts with auditory processing through cross-modal connections in both earlier and later stages of processing [7, 32, 40-43].

4.2. Measuring Integration

Besides pinpointing the locus of audiovisual integration, one of the major issues addressed concerns how to quantify “integration”. Some have argued that a key signature of multisensory integration relates to the *inverse effectiveness* of neural responses [14]. This principle stimulates that the neural response to multisensory stimuli must be greater than both unisensory stimuli, and furthermore, that multisensory enhancement increases as the intensity of the unisensory stimuli decreases. There are several difficulties associated with assessing multisensory convergence using fMRI and electrophysiological methods. One such difficulty relates to the fact that the BOLD signal from the fMRI, as well as surface electrodes, record from a large number of neurons in which unisensory auditory and visual, and multisensory neurons are intermixed [40-44]. Effects of audiovisual suppression/enhancement in the EEG (e.g., [32-41]) or BOLD signal may result from purely statistical reasons.

One way around these difficulties involves obtaining statistically motivated measures of integration, such as *C(t)*, and examining the extent to which they co-vary with neural signals [7]. In our case, we compared *C(t)* to mGFP and also the activation patterns obtained from CDRs. Capacity provides a measure of efficiency relative to the baseline predictions of a parallel independent model without interactions. Deviations from the value of *C(t)* = 1 at any point in time indicates the presence of multisensory interactions, whether they be inhibitory or excitatory.

While RTs occur much later than neural interactions responsible for language perception (e.g., [45]), there are several advantages in associating EEG measures with behavioral measures of efficiency. First, the EEG signal provides fine-grained temporal resolution unlike BOLD signals obtained from fMRI (8 ms in CDRs). Second, the EEG signal itself may be construed as a neural measure of energy expenditure [32] The mGFP (mean spatial standard deviation) also provides supplementary insight to *C(t)*—although direct comparisons of capacity with mGFP and EEG may be difficult due to differences in time scale. In the mGFP analysis in (Fig. **3B**) for instance, one may observe that the AV peak was suppressed compared to the A peak at latencies less than 100 ms and again at approximately 150-200 ms for the non-speech stimuli. Suppression could be associated with visual inhibition resulting from early from visual processing areas [7, 40]. Nonetheless, this conclusion will require more evidence from future studies systematically varying the strength of the auditory and visual signals.

In conclusion, integration in the neural domain was assessed by focusing on CDR and mGFP analyses comparing auditory and visual-only, and audiovisual activation. A qualitative analysis of the data indicated a broad range of activity associated with integration. Audiovisual integration patterns included areas such as the inferior posterior frontal cortex and SMG. Evidence also implicated canonical integration sites in temporal areas such as the STS and surrounding areas.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENT

This research was supported by funds provided to the Visual Neuroscience Laboratory at The University of Oklahoma, by the Dept. of Psychology, bridge funds through Michael J. Wenger, and URC grant (Idaho State University) awarded to Nicholas Altieri.

REFERENCES

[1] McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 1976; 264: 746-8.

[4] The spatio-temporal circuitry shown in (Fig. **5**) is broad enough to encompass integration processes associated with speech and non-speech inputs. While there may often be considerable overlap in brain regions associated with integration, we did observe the involvement of more extensive brain areas in non-speech integration, including the STG, and broader regions of the left primary auditory and temporal cortex.

[2] Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. J Acoustic Soc Am 1954; 26: 12-5.

[3] Erber NP. Interaction of audition and vision in the recognition of oral speech stimuli. J Speech Hearing Res 1969; 12: 423-5.

[4] MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. Br J Audiol 1987; 21: 131-41.

[5] Ross LA, Saint-Amour D, Leavitt V, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex 2006; 17: 1147-53.

[6] Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. J Acoustic Soc Am 1998; 103: 2677-90.

[7] Altieri N, Townsend JT. An assessment of behavioral dynamic information processing measures in audiovisual speech perception. Front Psychol 2011; 2: 1-15.

[8] Summerfield AQ. Preliminaries to a comprehensive account of audio–visual speech perception. Dodd B, Campbell R, Eds. Hearing by eye: the psychology of lip reading. Hillsdale, NJ: Erlbaum 1987; pp. 3-52.

[9] Altieri N, Townsend JT, Wenger MJ. A parallel interactive processing explanation of audiovisual integration and the McGurk effect. Under revision.

[10] Tiippana K, Puharinen H, Möttönen R, Sams M. Sound location can influence audiovisual speech perception when spatial attention is manipulated. See Perceiv 2011; 24: 67-90.

[11] Townsend JT, Nozawa G. Spatio-temporal properties of elementary perception: an investigation of parallel, serial and coactive theories. J Math Psych 1995; 39: 321-60.

[12] Wenger MJ, Gibson BS. Using hazard functions to assess changes in processing capacity in an attentional cuing paradigm. Human Perception and Performance, J Exp Psychol 2004; 30: 708-19.

[13] Stevenson RA, James TW. Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. Neuroimage 2009; 44: 1210-23.

[14] Stein BE, Meredith MA. The Merging of the Senses. Cambridge, MA: MIT Press 1993.

[15] Bernstein LE, Auer ET, Wagner M, Ponton CW. Spatiotemporal dynamics of audiovisual speech processing. Neuroimage 2008; 39: 423-35.

[16] Liberman AM, Mattingly IG. The motor theory of speech perception revised. Cognition 1985; 2: 1-36.

[17] Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science 1981; 212: 947-50.

[18] Tuomainen J, Andersen TS, Tiippana K, Sams M. Audio-visual speech perception is special. Cognition 2005; 96: B13-22.

[19] Miller JO. Divided attention: evidence for coactivation with redundant signals. Cog Psych 1982; 14: 247-79.

[20] Miller JO. Timecourse of coactivation in bimodal divided attention. Percept Psychophys 1986; 40: 331-43.

[21] Eidels A, Houpt J, Altieri N, Pei L, Townsend JT. Nice guys finish fast and bad guys finish last: a theory of interactive parallel processing. J Math Psych 2011; 55: 176-90.

[22] Rosenblum LD, Johnson JA, Saldaña HM. Visual kinematic information for embellishing speech in noise. J Speech Hearing Res 1996; 39: 1159-70.

[23] Sherffert S, Lachs L, Hernandez LR. The Hoosier audiovisual multi-talker database. In: Research on Spoken Language Processing Progress Report No.21, Bloomington, Speech Research Laboratory, Psychology (Department), Indiana University 1997.

[24] Ding L. Reconstructing cortical current density by exploring sparseness in the transform domain. Phys Med Biol 2009; 54: 2683-97.

[25] Altieri N, Pisoni DB, Townsend JT. Some normative data on lip-reading skills. J Acoustic Soc Am 2011; 130: 1-4.

[26] Bernstein L, Demorest ME, Tucker PE. What makes a good speech reader? first you have to find one. Campbell, Ruth; Dodd, Barbara; Burnham Denis, Eds. Hearing by eye II: Advances in the psychology of speech reading and auditory-visual speech (Hove, England: Psychology Press/Erlbaum: UK, Taylor and Francis 1998; pp. 211-27.

[27] Grice GR, Canham L, Gwynne JW. Absence of a redundant-signals effect in a reaction time task with divided attention. Percept Psychophys 1984; 36: 565-70.

[28] Townsend JT, Wenger MJ. A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. Psychol Rev 2004; 111: 1003-35.

[29] Skrandies W. Global field power and topographic similarity. Brain Topograph 1990; 3(1): 137-41.

[30] Calvert G, Bullmore ET, Brammer MJ, *et al.* Activation of auditory cortex during silent lipreading. Science 1997; 25: 593-6.

[31] Bernstein LE, Auer ET, Jr., Moore JK, Ponton CW, Don M, Singh M. Visual speech perception without primary auditory cortex activation. Neuroreport 2002; 13: 311-15.

[32] Altieri N, Wenger M. Neural dynamics of audiovisual integration efficiency under variable listening conditions. Under revision.

[33] Miller LM, D'Esposito M. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. J Neurosci 2005; 25: 5884-93.

[34] Puce A, Allison T, Bentin S, Gore JC, McCarthy G. Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci 1998; 18: 2188-99.

[35] 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. Cog Brain Res 2002; 14: 115-28.

[36] Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 2000; 10: 649-57.

[37] Calvert GA, Campbell R. Reading speech from still and moving faces: the neural substrates of visible speech. J Cog Neurosci 2003; 15: 57-70.

[38] Jones J, Callan D. Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. Neuro Rep 2003; 14: 1129-33.

[39] Yvert B, Fischer C, Bertrand O, Pernier J. Localization of human supratemporal auditory areas from intracerebral auditory evoked potentials using distributed source models. NeuroImage 2005; 28: 140-53.

[40] Altieri N, Pisoni DB, Townsend JT. Behavioral, clinical, and neurobiological constraints on theories of audiovisual speech integration: a review and suggestions for new directions. See Perceiv 2011; 24: 513-39.

[41] van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci USA 2005; 102: 1181-6.

[42] Besle J, Fort A, Delpuech C, Giard M-H. Bimodal speech: Early suppressive visual effects in human auditory cortex. Eur J Neurosci 2004; 20: 2225-34.

[43] Besle J, Bertrand O, Giard MH. Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. Hearing Res 2009; (1-2): 143-51.

[44] Laurienti PJ, Perrault TJ, Stanford TR, Wallace MT, Stein BE. On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. Exp Brain Res 2005; 166: 289-97.

[45] Näätänen R. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). Psychophysiology 2001; 38: 1-21.

[46] Wenger MJ, Negash S, Petersen RC, Petersen L. Modeling and estimating recall processing capacity: Sensitivity and diagnostic utility in application to mild cognitive impairment. J Math Psych 2010; 54: 73-89.

[47] Murray MM, Brunet D, Michel CM. Topographic ERP analyses: a step-by-step tutorial review. Brain Top 2008; 20(4): 249-64.

[48] Fuchs M, Wagner M, Kohler T, Wischmann H-A. Linear and non-linear current density reconstruction. J Clin Neurophysiol 1999; 16: 267-95.

[49] Wagner M, Fuchs M, Wischmann H-A, Ottenberg K, Dössel O. Cortex segmentation from 3D MR images for MEG reconstructions. In: Baumgartner, Deecke C, Stroink L, Williamson G, EDs SJ. Eds. Biomagnetism: fundamental research and clinical applications. Elsevier Science IOS Press; The Netherlands: Amsterdam 1995; pp. 433-8.

Received: July 05, 2013 Revised: September 20, 2013 Accepted: September 26, 2013