

**RESEARCH ON SPOKEN LANGUAGE PROCESSING**  
Progress Report No. 28 (2007)  
*Indiana University*

**Audiovisual Perception of Spoken Words in Speech and Nonspeech Modes:  
Measures of Architecture and Capacity<sup>1</sup>**

**Nicholas A. Altieri and James T. Townsend<sup>2</sup>**

*Speech Research Laboratory  
Department of Psychological and Brain Sciences  
Indiana University  
Bloomington, Indiana 47405*

---

<sup>1</sup> This study was supported NIH Grants DC-00111 and DC-00012 to Indiana University. I would like to acknowledge Jeremy Loebach, Mario Fific, and David B. Pisoni for insightful comments.

<sup>2</sup> Indiana University, Bloomington. jtownsen@indiana.edu

## **Audiovisual Perception of Spoken Words in Speech and Nonspeech Modes: Measures of Architecture and Capacity**

**Abstract.** Contemporary models of audiovisual speech perception attempt to explain accuracy data based on curve fitting and optimization techniques (see Braida, 1991; Massaro, 1987). Research on audiovisual speech perception lacks a formal mathematical foundation because current models do not make predictions about reaction time or adequately describe how the audio and visual channels are processed in the black box. The double factorial paradigm (DFP) developed by Townsend and Nozawa (1995) uses systems factorial technology to provide a framework for investigating how different information channels are processed. In Experiment 1, participants were required to make one response if auditory information, visual information, or both forms of information were present and a negative response on target absent trials. Data from the audiovisual detection task with the word “base” as the stimulus showed that processing architecture was either coactive or parallel self-terminating. A second experiment again using the double factorial paradigm methodology (Experiment 2) required participants to distinguish between two spoken words: “base” and “face.” The data showed that processing was mostly coactive, but possibly parallel self-terminating in some cases. Processing capacity was limited in both experiments, indicating a lack of redundancy gain. Overall, these results suggest that the audio and visual channels are combined into a single processor, although inhibition or competition may exist between channels.

### **Introduction**

The cognitive or information processing approach to psychology seeks to understand in a mathematically rigorous fashion how information is processed in the “black box.” Given a certain number of distinct inputs to the system, the output or subject’s response is measured, but the psychologist would ultimately want to understand the cognitive mechanisms that produced the output. Speech perception for example, is a multimodal perceptual faculty that relies on auditory, visual, and even haptic information as inputs to the system—where word or segment recognition is the output (Fowler & Dekle, 1991; Sumbly & Pollack, 1954). Sumbly and Pollack demonstrated the importance of the contribution of visual information in speech perception by showing that the proportion of audiovisual gain remains identical across all signal to noise ratios. It is also well established that when listeners are presented with incongruent audiovisual stimuli, the resulting percept is different than either the audio or visual stimuli, as is the case in the McGurk effect. The auditory stimulus was the utterance /ba/, which was dubbed over a visually articulated /ga/, and in the majority of cases, subjects reported experiencing the “perceptual fusion” /da/ (see McGurk & McDonald, 1976).

Researchers have long investigated the output of the black box and established that fact speech perception is a multi-modal phenomenon. However, broad classes of models related to the way audio and visual stimuli are processed in black box such as serial, parallel, or coactive processing have not been falsified or investigated. The mechanisms that listeners use to extract and combine information from different modalities in real time are not understood. An investigation of the processing architecture (i.e., parallel or serial) in the “black box” would provide a fundamental foundation for scientific investigations of audiovisual perception.

Most research on audiovisual speech perception assumes that listeners somehow combine information from the individual modalities, without explaining how integration occurs in the “black box”

or a neurologically based model. The Fuzzy Logic Model of Perception FLMP is one class of models, which assumes *a priori* that audiovisual integration occurs in an optimal fashion, where the relationship between audio and visual information are multiplied and divided by the sum of the alternatives (Massaro, 2004). FLMP uses a formulation of Bayes' theorem to determine the probability that a certain syllable, word, or phoneme was processed given the available audio and visual parameters.<sup>3</sup> A second model referred to as the pre-labeling integration model (PRE) is founded upon multidimensional signal detection theory, and assumes that the unimodal information scores will be used optimally, and that the predicted AV scores should be greater than or equal to observed unimodal identification scores (Braida, 1991).

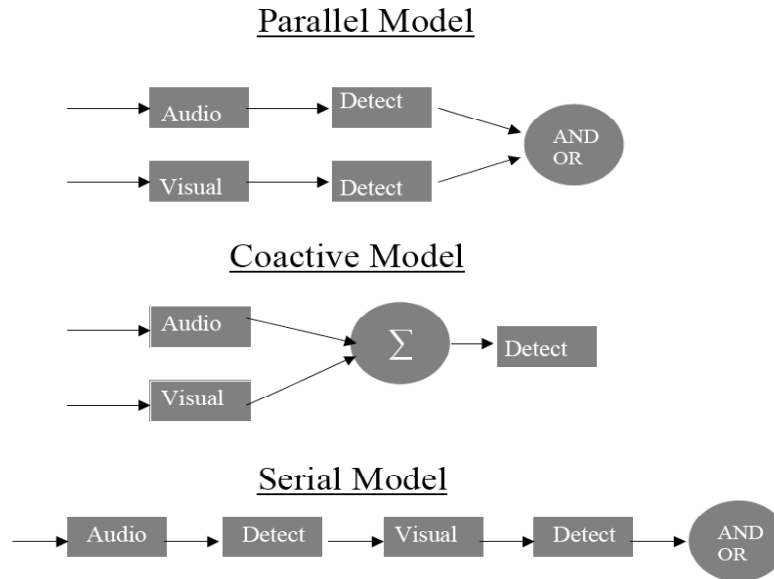
While FLMP and PRE account for confusion data when tested in audiovisual perception experiments (Grant, 2002; Grant, Tufts, & Greenberg, 2007; Massaro, 2004), they do not attempt to explain how cognitive systems process information from the audio and visual channels. The question is how are the audio and visual channels utilized and combined in real time to form a unified percept? A second and related point is that models of audiovisual perception do not make fine-grained predictions about reaction time data, which generally precludes mathematical modeling of dynamic processes.

Figure 1 shows two prominent conceptual accounts or neural representations of how integration might occur in an information processing system, along with a serial processing model where processing cannot begin on the second channel until it finishes on the first channel. The parallel model has independent channels where separate decisions are made on each channel. In this framework, the audio and visual speech streams are processed separately and simultaneously just prior to the decision stage. A separate decision is made on each channel or modality and a subsequent decision is made using an AND or an OR gate. Consider for example a case where a listener is given a task where they have to respond "yes" if presented with /ba/ in either the audio or visual modality. When /ba/ is presented, each channel accumulates information and if the auditory channel exceeds threshold, the listener responds "yes" regardless of whether the visual channel is finished accumulating information. In a coactive model, the information from each channel is combined into a common information processor that counts information from each source. Once the counter in this common processor exceeds threshold, a decision is made. Lastly, in the serial model, processing on the audio and visual components of /da/ or /da/ cannot occur simultaneously. If the auditory component is processed first, for example, then processing in the visual domain cannot begin until the audio channel is completely finished. If the system is self-terminating, then a decision can be made when the audio channel finishes, whereas if the stopping rule is exhaustive, both channels must finish.

FLMP and PRE do not make explicit predictions about serial, parallel, or coactive processing architecture. Massaro (2004) claims that the algorithms used in FLMP can implement either the parallel or coactive models depicted in Figure 1. A major undertaking in this project is to garner behavioral evidence to distinguish between the models depicted in Figure 1. Two general candidates for audiovisual speech recognition include coactive processing and parallel non-convergent processing, although serial processing will also be considered.

---

<sup>3</sup> In a two alternative forced choice task where the listener has to distinguish between /ba/ and /da/, the probability of a given value is a function of the audio and visual parameters:  $p(/da/ | A \& V) = aivj/[aivj + (1-ai)(1-vj)]$ .



**Figure 1.** Audiovisual Processing accounts. On top is a schematic representation of a parallel model with an OR as well as an AND gate. The coactive model below assumes that each channel is pooled into a common processor where evidence is accumulated prior to making a decision. The serial model at the bottom assumes that processing occurs one stage at a time. Processing cannot begin on the second channel on stage two unless processing on the first channel is completed.

### Audiovisual Speech Perception

While formal mathematical models have not been applied to distinguish coactive versus parallel processing, there has been discussion in the audiovisual speech perception literature pertaining to different processing architectures for the audio and visual channels. For instance, speech perception theorists from different schools of thought like motor theory (Liberman & Mattingly, 1985) direct realism (Fowler & Rosenblum, 1991), and other general processing theories (see Bernstein, 2005) differ in how they conceptualize audiovisual information processing. Motor theory and direct realism for instance, assume that the primitives of speech perception are articulatory gestures.<sup>4</sup> Rosenblum (2005) argues that the evidence of the importance of multimodal speech perception supports gesture based theories, and draws the conclusion that multimodal speech is the *primary* function of perception. He argues that information in the speech signal is present in every modality, and the perceptual processes involved in recognizing speech are “unconcerned” with regard to modality. Gesture based theories do not make explicit mathematical predictions with regard to the mappings between the auditory and visual channels. However, one way to illustrate this framework in the context of audiovisual perception is to conceptualize the information from the audio and visual modalities becoming “integrated” and combined into a single channel “early” in the decision process prior to word recognition (where the decision process considers only the sum of the information and not the information in the individual modalities), as depicted in the “coactive” model in Figure 1.

<sup>3</sup> In the case of motor theory, the motor gestures that produced the sounds are recovered by the listener using analysis by synthesis. For direct realism, information about gestures is carried by the speech signal and is perceived directly. For simplicity, motor theory and direct realism will be treated identically with regard to audiovisual perception in this paper.

Behavioral studies have provided some support for the view that speech perception is “unconcerned” with source modality, or that audiovisual integration occurs early, i.e., prior to word or segment recognition. Green and Miller (1985) demonstrated that visually perceived rate of articulation influences auditory segment perception. They used a McGurk paradigm to show that visual information about place of articulation can influence properties like voice onset time. Subjects were shown audiovisual clips of a talker saying a syllable that varied auditorially and visually on a continuum from /bi/ to /pi/. The corresponding visual information was played either fast or slow. They showed that slowly articulated syllables increased the percentage of time that subjects perceived /bi/ relative to /pi/. Because visual information influences the perception of features that are the components of word recognition, these findings indicate “early” integration of audiovisual channels, in which audio and visual information is combined into a single channel prior to word and segment recognition. They argued that the results were indicative of a decision process that has access to both auditory and visual information and combines the two sources of information prior to recognition.

Neuroimaging evidence from audiovisual speech perception tasks has suggested similar conclusions about the presence of coactive processing. Calvert and Campbell (2003) showed that silent lipreading tasks activate the primary auditory cortex. Subjects were presented with either sequences of still key frames or moving images of the same duration of a talker saying nonsense syllables. Subjects were instructed to look for a visible target syllable like “voo” in a sequence of other nonsense syllables. In contrast to resting conditions in which letters were superimposed on a resting face, sequences of still key frame images produced activation in the posterior cortical areas associated with the perception of biological motion. Activation was also observed in canonical speech processing areas including Broca’s area, the superior temporal sulcus (STS). However, moving images produced greater activation in these regions compared to still frames. They concluded that visual speech accesses areas traditionally believed to be auditory processing regions for language, which is possibly due to “dynamic audiovisual integration mechanisms” in the STS (Calvert & Campbell, 2003).

Super-additive activation in the STS has also been observed in congruent audiovisual speech perception tasks (Calvert et al., 1997), while incongruent audiovisual speech has yielded sub-additive activation in the STS (Calvert, Campbell, & Brammer, 2000). Super-additive activation occurs when the amount of activation recorded in a brain area in the bimodal condition is greater than the sum of the activation levels from each unimodal condition. The observation of super-additive levels of activation in the STS indicates the possibility that there are neurons and brain regions that only respond, or mostly respond to audiovisual input. The existence of neurons that respond selectively to audiovisual input provides at least some evidence that the brain might be implementing an information processing system analogous to the coactive model depicted in Figure 1 where the audio and visual components of the signal are combined into one channel prior to segment or word recognition.

Nonetheless, the conclusion that multi-sensory neurons are responsible for processing audiovisual speech is not uniformly accepted. The BOLD response is a measure of the blood oxygen level in a brain region and therefore represents an indirect measure of neural activity. fMRI designs also suffer from poor temporal resolution. Observations of super-additive levels of activation in the STS could be due to “commingled” unisensory neurons (Bernstein, Auer, & Moore, 2004; Meredith, 2002). That is, areas that are believed to respond only to audiovisual speech in reality contain large numbers of unisensory neurons. Furthermore, the STS responds not only to speech, but also to complex nonspeech gestures (Puce, Allison, Bentin, Gore, & McCarthy, 1998). When presented with pairs of moving eyes or moving mouths, bilateral activation was observed in the posterior STS, while the control stimuli consisting of moving checkered patterns did not activate the STS or surrounding areas. These data appear to indicate that the auditory and visual streams are not converging to a common processor, and therefore there is insufficient evidence for a coactive processing model.

Bernstein (2005) argued instead that while speech is part of a highly specialized cortical system, not all motor and perceptual areas of the cortex seem to be devoted to speech perception, as gestural theories would assume. According to Bernstein, auditory and visual speech stimuli might be processed separately and simultaneously and “converge” only after phonetic perception and word recognition. Bernstein reasons that multimodal perception of the speech signal involves separate and simultaneous analysis of the audio and visual inputs. According to this account, the information from the audiovisual speech streams is processed in parallel, where extensive unisensory processing occurs before the binding of auditory and visual speech representations. This view is analogous to the parallel model discussed in Figure 1, which differs architecturally from the coactive model where one common processor integrates audio and visual information prior to phonetic perception.

### **Double Factorial Paradigm: Assessing Architecture and Capacity**

Given the coactive and parallel models of integration in the context of Rosenblum (2005) and Bernstein’s (2005) respective analyses on audiovisual speech perception, it is pertinent to return to the purpose of this project by finding a way to distinguish between these two models. The double factorial paradigm (DFP) developed by Townsend and Nozawa (1995) is an experimental methodology that can be used to obtain behavioral evidence to distinguish parallel from coactive processing. The description of the coactive and parallel models in the speech perception literature, while of theoretical importance, requires a more specific mathematical formulation along with behavioral data if they are to be adequately distinguished due to conflicting and imprecise accounts discussed in previous paragraphs.

The methodology for assessing mental architecture involves a factorial methodology that captures potential interactions between factors. One statistic that has been used to analyze interactions is the mean interaction contrast, or  $MIC = RT_{ll} - RT_{lh} - (RT_{hl} - RT_{hh})$  (see Sternberg, 1969). In this formula, RT designates reaction time, and each subscript represents the level of one factor like presence or absence of a feature or brightness: h = high, which indicates fast reaction times and l = low, which indicates slower reaction times. The hh condition for example might represent audio and visual stimuli of a high level of clarity, which a listener would be able to identify more quickly than if the audio or visual portions (or both) were degraded or less salient. One shortcoming of the MIC is that it is a coarse measure representing only one point at each level (i.e., the mean or median of the distribution). Townsend and Nozawa (1995) developed a more sensitive measure that analyzes the curve of the entire distribution of reaction times referred to as the *survivor interaction contrast* (SIC). The SIC is defined as  $SIC(t) = S_{ll}(t) - S_{lh}(t) - (S_{hl}(t) - S_{hh}(t))$ . Notice that the SIC uses the same sequence of terms as the MIC, only this time survivor functions are used rather than mean reaction times. Let  $S(t) = 1 - F(t)$ , where  $F(t)$  is the cumulative distribution function of the density function  $f(t)$  of reaction times. The survivor function  $SIC(t)$ , is a distribution function indicating the probability that a process is still going on. If audiovisual stimuli is presented, then  $SIC(t)$  would indicate the probability that the word, phoneme, or stimulus has not been recognized and identified by the subject by time t.

The SIC function makes several predictions about processing architecture. For the type of parallel processing described by the non-convergent model which assumes that each channel has its own decision stage, the SIC function can be positive or negative depending on the stopping rule. A parallel model with separate decisions and an exhaustive stopping rule predicts a negative SIC curve. “Exhaustive processing” refers to a stopping rule in a parallel system where each channel must finish processing before a decision is made. The reason for underadditivity in parallel exhaustive models is because each element must be completed before the system terminates. In other words, the processing of the system is determined by the slowest element. On the lh or hl trials, the longest time tends to be closer to the longest time on the ll

trials. Thus, the difference between  $S_{ll}(t) - S_{lh}(t)$  is generally smaller than the difference between  $S_{hl}(t) - S_{hh}(t)$ .

The case is exactly the opposite for parallel minimum time self-terminating models (or horse race models), which terminate when the fastest element finishes. The SIC function for these models is positive since the difference between  $S_{ll}(t) - S_{lh}(t)$  is generally greater than the difference between  $S_{hl}(t) - S_{hh}(t)$ . The reason is because the  $lh$  trials have an element that takes less time to process.

Coactivation might be considered a class of parallel models where the information from each channel is pooled into a single channel governed by a Poisson summation process. The survivor interaction function for Poisson summation models is negative at the beginning for low  $t$ , and becomes positive at later times  $t$ . The mean interaction contrast is positive. While the shape of the SIC function may not conform to intuition, it does make sense mathematically. The rate of coactive models is the sum of the rates of each channel—the sum of the audio and visual channels. For certain time  $t$ , the contrast will either be positive or negative. The SIC function is a function of the rate parameter and the curvature corresponds to the sign of the second derivative, which as stated above is negative for low  $t$ , and becomes positive as  $t$  increases (Townsend & Nozawa, 1995).

Finally, serial processing predicts an MIC of 0 regardless of whether the stopping rule is exhaustive or self-terminating. When processing in serial with a self-terminating stopping rule, the SIC( $t$ ) function is flat and equal to 0 at each point. Interestingly in the exhaustive case, the SIC( $t$ ) resembles an S-shaped curve with a negative region for early processing times and a positive region for later processing times (Townsend & Nozawa, 1995). The negative and positive regions of the curve are equal to each other in serial exhaustive model, and if we integrate over the curve, the total area is equal to zero.

### Capacity and Audio-Visual Gain in Speech Perception

A second feature of the DFP is its ability to assess the *capacity* of the system. Capacity is a measure that determines how the number of channels present affects the processing speed at a given time  $t$ . In other words, is there a cost, benefit, or no change in processing when both audio and visual channels are present (redundant target) relative to conditions when only the audio or visual channel is operating (single target)? If the processing rate is unaffected by increasing the number of channels, the system operates at unlimited capacity, if it slows down, then it operates at limited capacity, and if there is a benefit in processing rate, then it operates at super capacity.

Measuring processing capacity requires looking at the ratio of the integrated Hazard functions. The form of the hazard function is given below.

$$h(t) = f(t)/[S(t)] \quad (1)$$

Where  $f(t)$  is the probability density function, and  $S(t)$  is the survivor function which yields the probability that a process has not yet finished. The hazard function  $h(t)$  indicates the probability that a process will terminate at the next moment ( $t + 1$ ) in time given that it has not yet terminated at time  $t$ .

To calculate the capacity coefficient  $C(t)$  at each point in time, we calculate the integrated hazard function for the conditions where the subject is presented with the redundant target and divide it by the sum of the integrated hazard functions of the single target conditions (Townsend & Nozawa, 1995). The subscripts A and V indicate the audio and visual channels.

$$C(t) = H_{AV}(t)/[H_A(t) + H_V(t)] \quad (2)$$

The integrated hazard function  $H(t)$  is equivalent to  $\log[1 - F(t)]$  or  $\log[S(t)]$ , and in the field of physics it is used as a measure of the total energy consumed. The system operates at super capacity at a certain point in time  $t$  if  $C(t)$  is greater than 1 at that point, unlimited capacity if it equals 1, and limited capacity if it is less than 1 (Wenger & Townsend, 2000).

As previously stated, it is known that congruent audiovisual information about spoken words facilitates accuracy levels in perception (Sumbly & Pollack, 1954). However, the notion of processing capacity as defined above has generally been left unaddressed in the audiovisual speech perception literature, although research has been conducted investigating redundant target effects for nonspeech auditory and visual stimuli (see Berryhill, Kveraga, Webb, & Hughes, 2007; Miller, Kuhlwein, & Ulrich, 2004; Schroter, Ulrich, & Miller, 2007 for a discussion). Berryhill et al. (2007) presented subjects with congruent audiovisual stimuli (with a visual lead (SOA) of 0, 75ms, 150ms, and 225ms). The stimuli consisted of symbolic tokens of the numerals 1 and 2 presented in the visual modality, and auditory tokens of a talker saying 1 or 2, where the task of the participants was to determine whether '1' was presented or '2' was presented. Each trial was an audio only trial, visual only trial, or audiovisual trial (redundant target). They observed limited capacity, or lack of redundancy gain, when presentation of the audio and visual components was synchronized. When the lead (SOA) of the visual stimuli increased, capacity became less limited, and at SOAs of 150 and 225ms, a redundancy gain was observed.

In this study, the double factorial paradigm was applied in two separate experiments to test architecture and capacity in a control study where subjects were not required to attend to speech (i.e., nonspeech mode: Experiments 1A and 1B). A second Experiment (2) was conducted where subjects were required to distinguish between two spoken words. Both experiments used RT data to test audiovisual processing architecture and capacity using the formal framework of the double factorial paradigm. These experiments were designed to look inside the black box and begin to analyze whether processing of audiovisual components is parallel, coactive, or even serial in tasks where subjects were required to identify the presence of a talker or distinguish between spoken words of English. Experiment 1 was an audiovisual detection task using video clips of a single talker as stimuli. This experiment was a control study where subjects were exposed to a talker speaking a word of English. They were required to focus on the surface properties of the stimuli to judge whether a stimulus was present or absent, and were required to detect stimuli rather than engage in spoken word recognition. We compared the results (i.e., architecture and capacity) from Experiment 1 with the results from Experiment 2. The purpose was to assess whether the results from the speech perception experiment were particular to high-level cognition such as spoken word recognition, or whether they reflect general audiovisual processing mechanisms involved in simply identifying "complex" stimuli like the moving face of a talker.

## Experiment 1A

### Participants

Seven subjects (four females and three males) with normal or corrected vision were paid ten dollars per session for their participation. Data analysis was not conducted for one subject who only completed one session.

### Materials

The experiment was carried out in the Speech Research Laboratory (SRL) at Indiana University in Bloomington. The stimulus materials included an audiovisual movie clip of a female talker from the Hoosier Multi-talker Database saying the word "Base". A total of eight different stimuli were created

from this video clip: two audio files at two levels of saliency, two video files at two levels of saliency, and four audiovisual clips at each factorial combination of high-high, high-low, low-high, and low-low levels of saliency. The audio, visual, and audiovisual files were edited using Final Cut Pro HD version 4.5. The audio files were sampled at a rate of 48 kHz at a size of 16 bits. The high saliency audio files were presented at 57 dB and the low saliency audio files presented at a volume of 45 dB. The brightness level on the video files was manipulated to create two different levels of saliency. On the low saliency video files, the brightness was reduced 90 steps using the brightness video filter. This had the effect dimming and reducing the contrast of the video, making it more difficult to perceive the talker's articulators. Both audio and video files lasted for a total duration of approximately 1,600 milliseconds.

## Design and Procedure

Subjects were seated 14 to 16 inches in front of a Macintosh computer equipped with Beyer Dynamic-100 headphones. Each trial began with a fixation cross appearing in the center of the computer screen followed by either a target absent trial, or one of the eight possible stimuli: target present and target absent. One fourth of the trials were target absent trials in which no stimulus appeared after the plus sign on the center of the screen. The stimulus trials included either audio only, visual only, or audiovisual stimuli. Experiment 1 was an OR design where subjects were instructed to respond, as quickly and accurately as they possible by pressing the button labeled "Base" if they heard either the word "base" (audio only), saw the talker utter the word "base", or were exposed to a redundant target where both the audio and visual components of the word "base" were present. They were instructed to respond by pressing the button labeled "Nothing" if no stimulus appeared on the screen. There was a 750-millisecond delay between trials.

There were a total of 800 target-absent trials, 800 audio only trials, 800 audiovisual trials, and 800 visual only trials for a total of 3,200 trials per subject (1/4 Nothing, 1/4 A only, 1/4 V only, 1/4 AV). This included 200 trials in each redundant target condition (hh, hl, lh, ll). Participants were run for 40 blocks at 80 trials each with a break scheduled between each block. Participants also received sixteen practice trials at the onset of each experimental session that were not included in the subsequent data analysis. The experiment lasted approximately 45 minutes and was conducted over a course of 4 days.

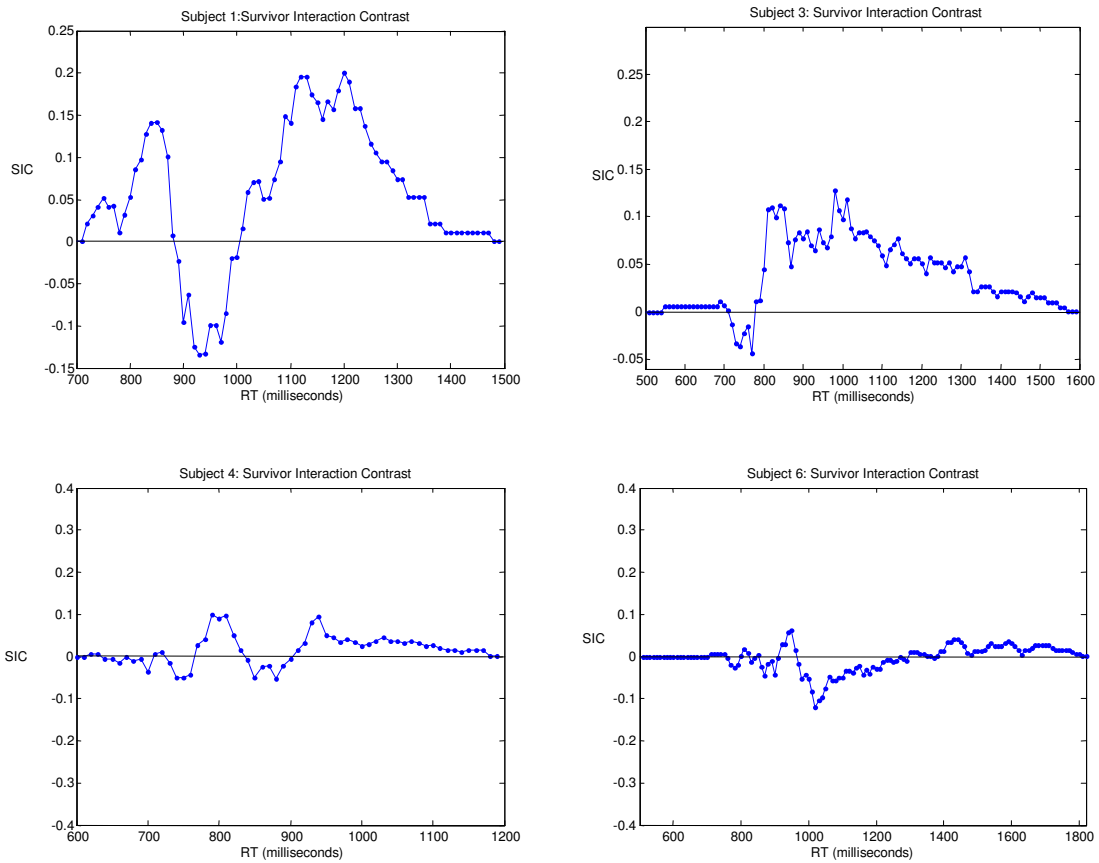
## Results and Discussion

Percentage of errors averaged across all participants was less than 2 %. Evidence of a speed-accuracy trade off was not observed. Therefore, only reaction time results will be presented.

The primary focus in Experiment 1A was on the set of SIC curves for each participant, which are distribution free (Townsend & Nozawa, 1995). Data from each participant was analyzed separately rather than averaged together since the results would have obscured individual differences and possibly led to different conclusions (see Townsend & Fific, 2004). ANOVAs and the mean interaction contrast (MIC) were analyzed in this experiment because they can help confirm or disconfirm interactions between the factorial conditions, which is an important tool for disconfirming serial processing. Serial processing would display a MIC of 0 (no interaction) and a flat SIC. The integral of the SIC curve is equal to the mean interaction contrast. Results of the SIC and mean interaction will be discussed together. Finally, the capacity coefficient,  $C(t)$ , which is a measure of the system's capacity at time  $t$ , was also of interest and will be addressed in subsequent analyses.

SIC curves for four participants who demonstrated selective influence appear in Figure 2. The MIC appears in Table 1 along with the ANOVA results for the four factorial conditions. A bin size of 10 milliseconds was used to calculate each survivor function in each experiment. Recall that each participant

completed 200 trials in each factorial condition, but errors and outliers (+ or - 3.0 SD from the mean) were eliminated from the analysis.



**Figure 2.** SIC curves for four subjects 1, 3, 4, and 6. These were the subjects who showed selective influence.

Subject 1’s ANOVA results shown in table 2 disconfirm serial processing. The SIC curve, while mostly positive, dips below zero yielding a small range of negativity around 900 ms. Since the SIC curve is not entirely positive, it fails to confirm parallel self-terminating processing behavior in this subject. One possible explanation, given the inconsistent curve and the positive interaction is that Subject 1 used dual processing strategies during the task, switching from parallel to serial.

Subject 3’s results reveal a positive SIC curve with negativity for early processing times and a positive MIC. This indicates coactive or possibly parallel self-terminating processing that finishes when either the audio or visual channel has reached a decision. The significant results provided by the ANOVA in Table 2 support this conclusion, along with the fact that the capacity coefficient discussed in the following section indicates severely limited capacity.

Subject 4’s results suggest either serial self-terminating or indeterminate behavior due to weak selective influence between the redundant target conditions. Subject 4’s SIC curve was neither positive nor negative and fit the line  $SIC(t) = 0$  with a root mean squared error of .019 and a sum of squared errors

of 1.57. Since the MIC was close to zero and the ANOVA did not even approach significance, we can tentatively accept the result that the behavior for this participant was serial self-terminating.

Subject 6's SIC curve was a flat line like subject 4's curve. Likewise, these results taken together with the MIC are indicative of serial self-terminating, or again indeterminate behavior due to weak selective influence. The fit to the flat line  $SIC(t) = 0$  had a root mean squared error of .017 and a sum of squared errors of 1.40. The corresponding ANOVA did not show a trend toward significance.

The SIC curves and ANOVA results from subject 3 suggested parallel or coactive processing. The SIC curves and ANOVA results from subjects 4 and 6 on the other hand, indicated serial processing. Subject 4's SIC curve was flat and the MIC was close to zero. It is possible in some instances that subjects process audiovisual material in a serial manner and self-terminate when a decision is made. Subject 6's SIC curve, similar to Subject 4's, was generally flat at  $SIC(t) = 0$ . The MIC was close to zero without a trend toward an interaction between channels. These ANOVA results added to the evidence that Subject's 4 and 6 processed the audiovisual stimuli in a parallel self-terminating manner.

Subject	df1	df2	<i>F</i>	<i>p</i>	Mic
1	1	181	11.019	.001	44.28
3	1	180	6.23	.013	41.90
4	1	171	.365	.546	7.28
6	1	181	.014	.905	3.45

**Table 1.** General Linear Model showing the level of interaction between the audio and visual channels. The Mean Interaction Contrast (MIC) is also displayed. This table shows the *F* value for the mean interaction, the *p* value (sig. = .05), and the mean interaction contrast.

The second part of this analysis involves examining the system's capacity. Specifically, we were investigating whether having both channels operating increases efficiency, decreases efficiency. Results of the measured capacity coefficient  $C(t)$  are compared with the bound for super capacity discussed previously in the introduction in addition to Grice's inequality (see Townsend & Nozawa, 1995).

The performance of each subject in the redundant target condition was compared with the predictions of an unlimited capacity parallel processing model (i.e.,  $C(t) = 1$ ). Figure 3 shows plots of the capacity coefficient for each of the six participants in Experiment 1A. The solid line at  $C(t) = 1$  is the bound for unlimited-super capacity. Data points above the line are indicative of super capacity, data points below the line are indicative of limited capacity, and data points hovering around the line indicate unlimited capacity. The boundary indicated by  $C(t) = 1/2$  represents the Grice bound for limited to extremely capacity. Grice's inequality is defined below:

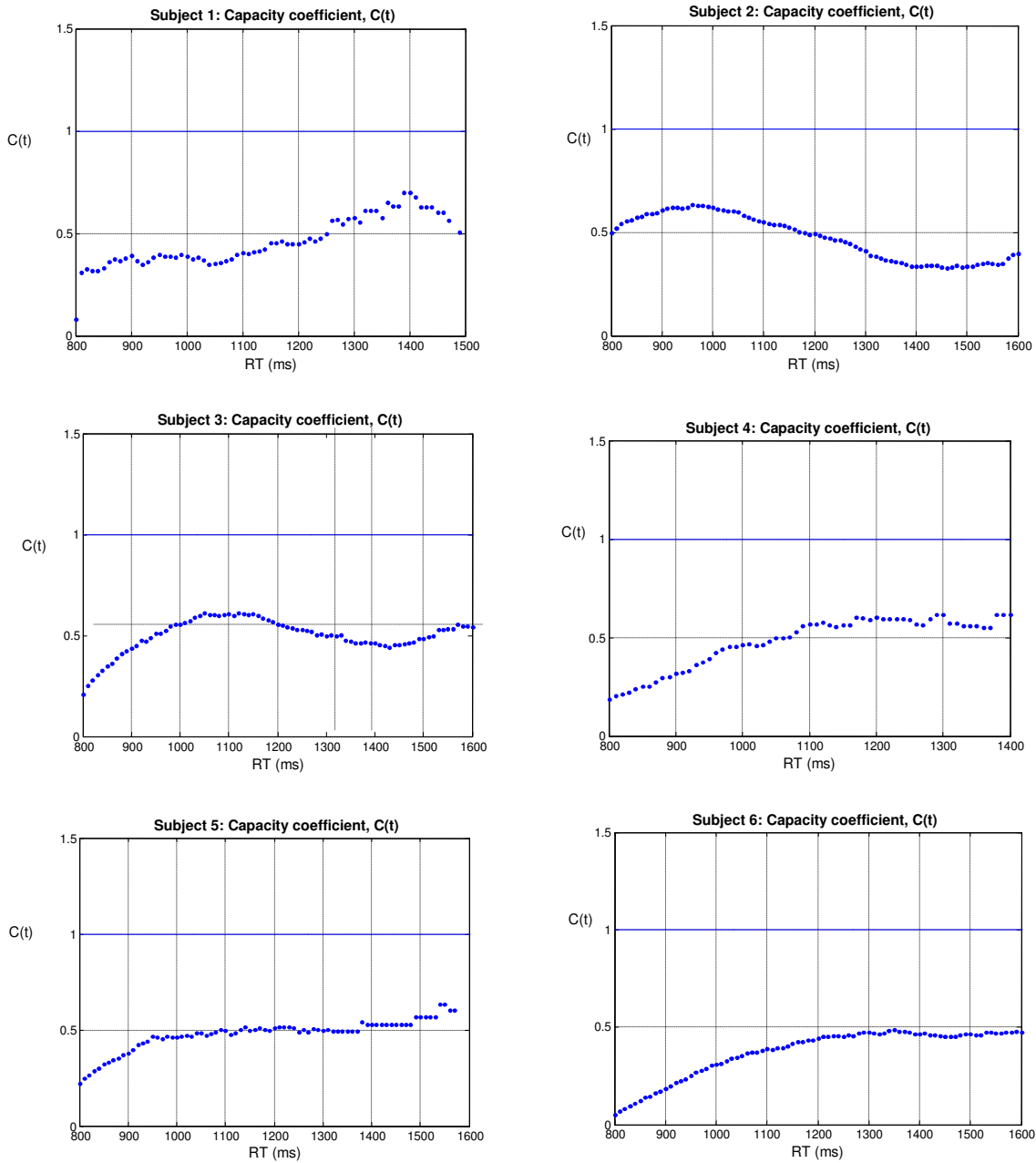
$$C(t) < \text{MAX}[HV(t), HA(t)] / [HV(t) + HA(t)] \quad (3)$$

The value in the numerator is the highest unimodal hazard function or the slower of the two processes. When the distributions of completion times for each channel are identical, Grice's inequality = 1/2.

The definition of "fixed capacity" is the average of the two single target integrated hazard functions (if we assume equal distribution parameters), which means that when two channels are operating, fixed capacity is  $C(t) = 1/2$ . Most of the data points fall below Grice's bound for extremely limited capacity and generally hover around  $C(t) = 1/2$ . Experiment 1 data support a limited to extremely

limited or fixed capacity model since  $C(t) < 1$  (where  $C(t) \sim 1/2$ ) for all six subjects across all time bins, even for small values of  $t$ .

The data from Experiment 1A indicated variable processing strategies for subjects. One possible reason for variability in processing strategies might have been the long exposure times of the stimuli combined with the simple experimental design. Therefore, the processing architecture data obtained in Experiment 1A is inconclusive. However, the capacity coefficient remained consistent across subjects, which supports the hypothesis that processing capacity is extremely limited in audiovisual detection tasks.



**Figure 3.** The Capacity coefficient for each of the six participants in Experiment 1A. Processing capacity was extremely limited for each subject.

## Experiment 1B

Experiment 1B was a modification of Experiment 1A. The audio and visual stimuli used in Experiment 1A (the female talker saying the word “Base”) were shortened where only the first five frames of the video and corresponding audio files were used. This manipulation had the effect of shortening the duration of the audiovisual stimuli from 1,624 ms to approximately 150–160 ms. The purpose of this manipulation was to improve selective influence by helping to reduce eye movements and variability in each of the redundant target reaction time distributions. The audiovisual files were cropped beginning at the onset of the word in “Base”. The SIC curves in Experiment 1A were highly variable. Of the four subjects showing selective influence of experimental manipulation, two yielded SIC functions that were basically flat, indicating parallel processing.

Since the stimuli lasted over 1,000 milliseconds in Experiment 1A, it was possible for subjects to move their eyes and therefore potentially shift processing strategies. The purpose of Experiment 1B was to eliminate variable processing strategies by manipulating the duration of the stimulus materials.

### Participants

Five participants (two males and three females) with normal or corrected vision were paid ten dollars per session for their participation.

### Materials

The materials were identical to those used in Experiment 1. The audiovisual files were shortened using Final Cut Pro HD version 4.5.

### Design and Procedure

The design and procedure was identical to task used in Experiment 1A.

### Results and Discussion

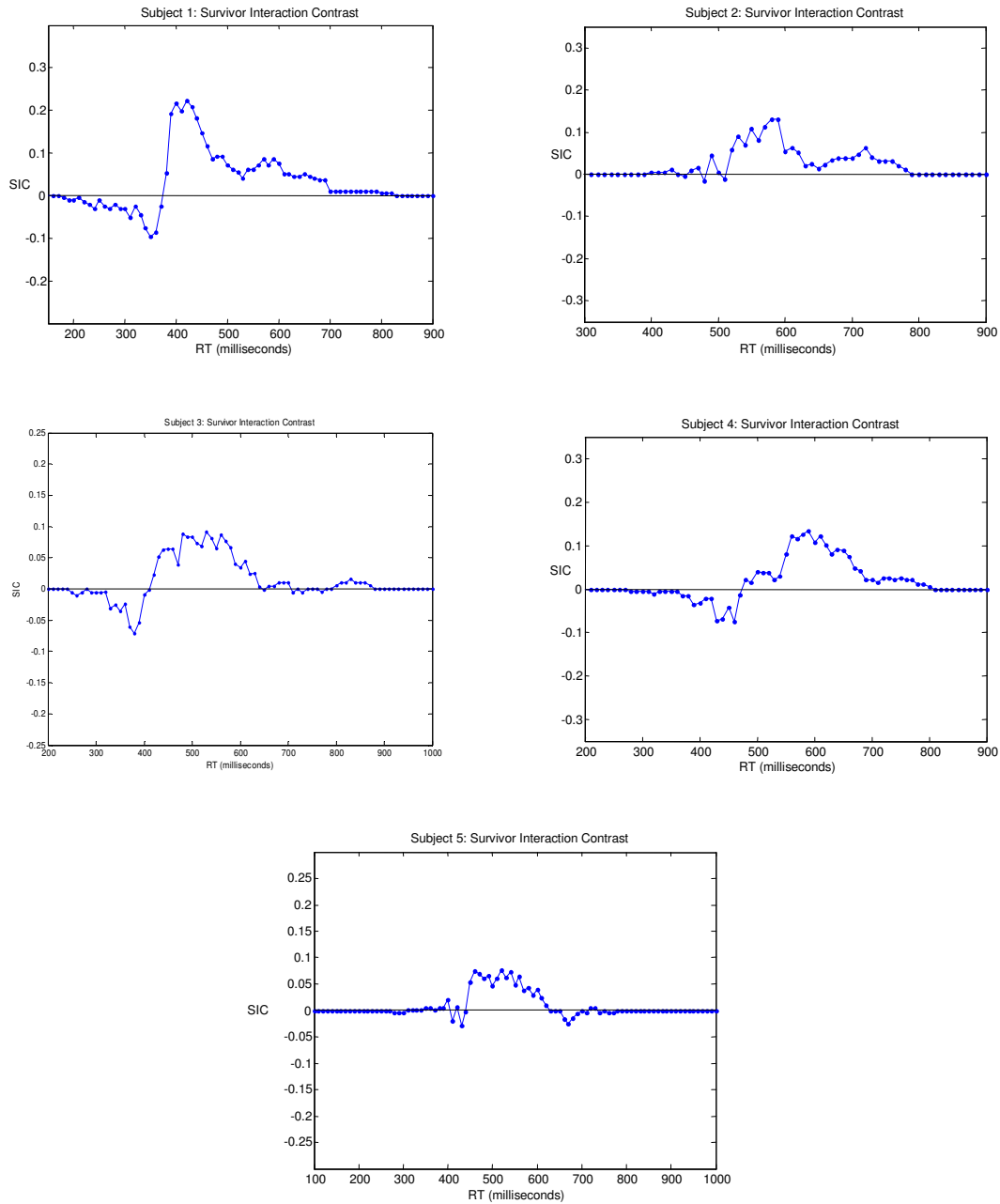
Percentage of errors averaged across all participants was less than 5%. As in the case of Experiment 1A, evidence of a speed-accuracy trade off was not observed. Therefore, only reaction time results are discussed.

Participants in Experiment 1B showed less between subject variability in the SIC curves. Participants in Experiment 1A on the other hand, either failed to show selective influence, or yielded SIC curves that were indicative of parallel self-terminating processing or coactive processing, or even serial-self terminating processing.

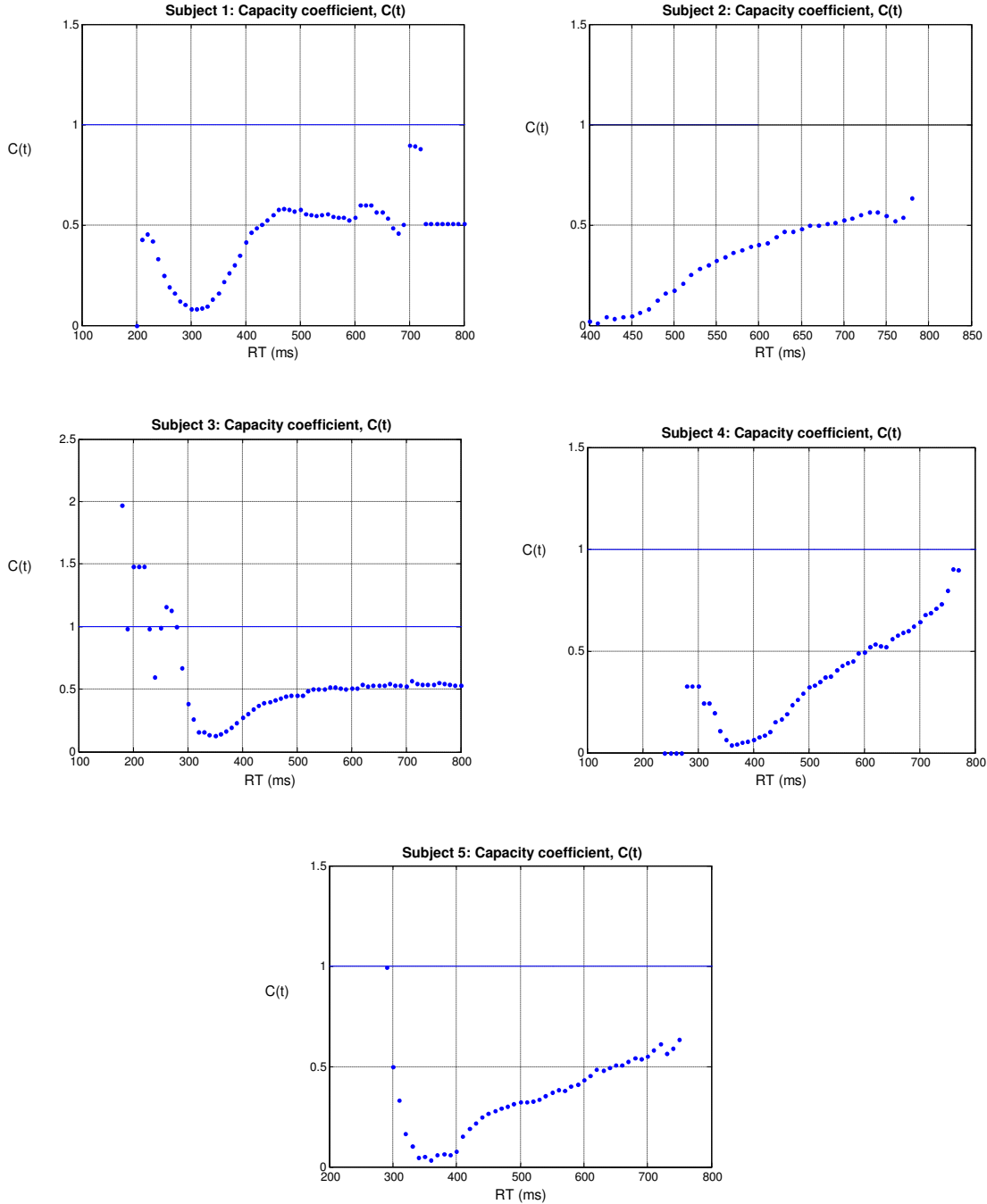
Subject	df1	df2	F	p	MIC
1	1	191	199.6	< .001	21.940
2	1	192	4.592	< .05	17.049
3	1	195	7.520	< .05	20.618
4	1	195	1.768	.185	14.162
5	1	195	7.670	< .01	10.00

**Table 2.** This table indicates the level of audio and visual channel interaction for each subject. The mean interaction contrast is also indicated.

Each of the five subjects demonstrated selective influence. SIC curves are shown for each of the five subjects in Figure 4 below. Each participant completed 200 trials in each factorial condition—the same amount of trials that were completed in Experiment 1A. Errors and outliers (+ or – 3.0 SD from the mean) were eliminated from the analysis. Table 2 displays the F values and MIC for each of the five subjects.



**Figure 4.** SIC curves for all five subjects in Experiment 1B. Each subject showed selective influence.



**Figure 5.** The Capacity coefficient for each of the five participants in Experiment 1B. Processing capacity was extremely limited for each subject.

The SIC for each subject was over-additive ( $> 0$ ), strongly suggesting parallel or coactive processing strategies. The SIC and MIC for subjects 2 and 5 was almost entirely over-additive, indicating parallel self-terminating processing. The SIC for subjects 1, 3, and 4 show negativity for early stages of processing. The MIC was positive for subjects 1 and 4, although the ANOVA on the interaction was statistically significant for subject 1 but not 4. The positive MIC supports the hypothesis that processing

was coactive for these subjects, but the case is weaker for subject 4 whose F value was not statistically significant. Negativity at early processing stages is indicative of coactive processing, while positive SIC functions as previously discussed indicate parallel self-terminating processing.

The capacity functions for each subject shown in Figure 5 differ slightly from those obtained in Experiment 1A. The capacity coefficient  $C(t)$  for each subject was below 1 indicating fixed or limited capacity. Data from each participant shows that the Grice inequality was violated across many points in time. The capacity data differ slightly from the data in Experiment 1A because Miller's inequality was violated in Subject 3's and Subject 5's data. In short, while largely consistent with the data obtained in Experiment 1A, capacity, at least for some subjects, was not as limited at early processing times.

## **Experiment 2**

Experiment 2 was designed to test architecture and capacity in a speech recognition task where participants have to distinguish between two words. Experiments 1A and 1B were control tasks where participants engaged in the detection, but not recognition, of audiovisual stimuli.

### **Participants**

Five female subjects with normal or corrected vision were paid \$10/session for their participation. Data from one subject was removed since that individual did not complete all experimental sessions.

### **Materials**

The stimulus materials included two audiovisual movie clips of a female talker from the Hoosier Multi-talker Database saying the words "Base" and "Face". The two words in this set are intended to be confusable, with only the onset phoneme (/b/ versus /f/) differing between them. A total of eight different stimuli were created from each video clip: two audio files at two levels of saliency, two video files at two levels of saliency, and four audiovisual clips at each factorial combination of high-high, high-low, low-high, and low-low levels of saliency. The audio, visual, and audiovisual files were created using Final Cut Pro HD version 4.5. The audio files were sampled at a rate of 48 kHz using 16 bit encoding. Pink noise was generated using Adobe Audition and mixed into each audio file to create two different signal-to-noise ratios, and hence two different levels of saliency. The two signal-to-noise ratios for both stimuli were 40 dB for the high condition and 0 dB for the low condition.

The brightness level on the video files was manipulated in the same way as in Experiment 1A and 1B. The audio and video files lasted for a total duration of 1,616 milliseconds for "Base" and 1,683 milliseconds for the word "Face". The beginning of each audio and video file was edited in Final Cut Pro in order to create identical onset times for the spoken stimuli.

### **Design and Procedure**

Subjects were seated in front of a Macintosh computer equipped with *Beyer Dynamic-100* headphones. Each trial began with a plus sign appearing in the center of the computer screen followed by the word "base" or "face." Trials are either audio alone, visual alone, or AV. Subjects were instructed to respond, as quickly and accurately as possible by pressing the button labeled "Base" if they either heard the word "base", saw a video of the talker saying "base", or both. Subjects were instructed to press the button labeled "Face" if they heard the word "face", saw a video of the talker saying "face", or both. There was a 1,000 millisecond delay between trials.

Each subject was presented with 3,360 total trials with 1,120 audio only trials (560 “base” + 560 “face”), 1,120 visual only trials, (560 “base” + 560 “face”), and 1,120 audiovisual trials (560 “base” + 560 “face”). Additionally, there were 280 trials in each redundant target condition (hh, hl, lh, ll). Table 3 below shows a diagram of the experimental design. Participants were run for 28 blocks at 120 trials each with a break scheduled between each block and each experimental session lasted approximately one hour. The experiment required four to five days to complete. Participants also received sixteen practice trials at the onset of each experimental session that were not included in the subsequent data analysis.

Audio	Visual	Correct Response
A <sub>Base</sub>	V <sub>Base</sub>	Base
A <sub>Base</sub>	∅	Base
∅	V <sub>Base</sub>	Base
A <sub>Face</sub>	V <sub>Face</sub>	Face
A <sub>Face</sub>	∅	Face
∅	V <sub>Face</sub>	Face

**Table 3.** This table shows each stimulus-response category (Base and Face) along side each factorial condition.

## Results and Discussion

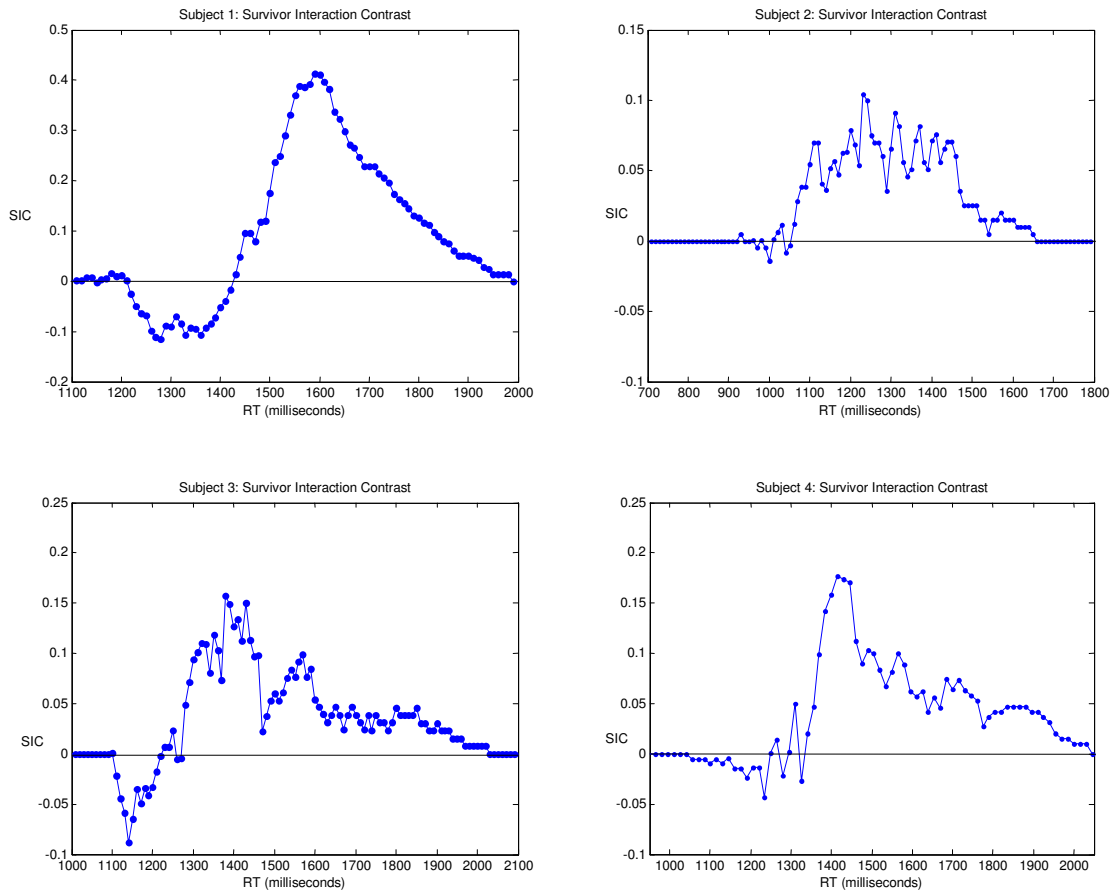
Percentage of errors averaged across all participants was less than 10 %. The error rate was likely higher in Experiment 2 due to the increased complexity of the task requiring subjects to distinguish between two similar spoken words of English. Each subject was close to or above 90 % accuracy across conditions. Evidence of a speed-accuracy trade off was not observed in the redundant target condition.

As in Experiments 1A and 1B, the initial analysis consisted of an investigation of the SIC curves and corresponding ANOVAs. Each subject showed selective influence. SIC curves for 4 subjects in Experiment 2 appear in Figure 6. ANOVA results and the MIC are shown in Table 4. The different nature of the tasks in Experiments 1 and 2 was the reason that subjects failed to show selective influence (or showed weaker selective influence) in the former experiment but not the latter. Although the duration of the stimuli remained the same between Experiments 1A and 2, participants were required to simply detect the presence of a moving image or sound in the first experiment, whereas in Experiment 2, the task was more likely to be cortically driven requiring them to distinguish between two words. More evidence was able to accumulate in each channel in Experiment 1 because the stimuli remained on for a longer time (compared with shorter durations in Experiment 1B), resulting in a smaller difference in completion times between the high and low conditions.

Recall that each participant in Experiment 2 completed 28 blocks consisting of 120 trials. Overall, the data demonstrate consistent processing between subjects. Each subject’s SIC curve is over additive (greater than 0) at most points. Furthermore, each subject’s MIC is positive and the corresponding one-way ANOVA indicates either a strong trend, or a significant positive interaction between the audio and visual channels. The positive SIC curve with the MIC and ANOVA results indicate parallel processing while observing a minimum time or self-terminating stopping rule.

Subject	df1	df2	F	p	Mic
1	1	263	3.34	.12	38.0
2	1	261	2.99	~ .05	21.4
3	1	261	3.87	< .05	36.7
4	1	201	.71	< .50	22.1

**Table 4.** This table shows the F value for the mean interaction, the p value (sig. = .05), and the mean interaction contrast for Experiment 2.



**Figure 6.** SIC curves for each the four subjects in Experiment 2. Each subject showed selective influence.

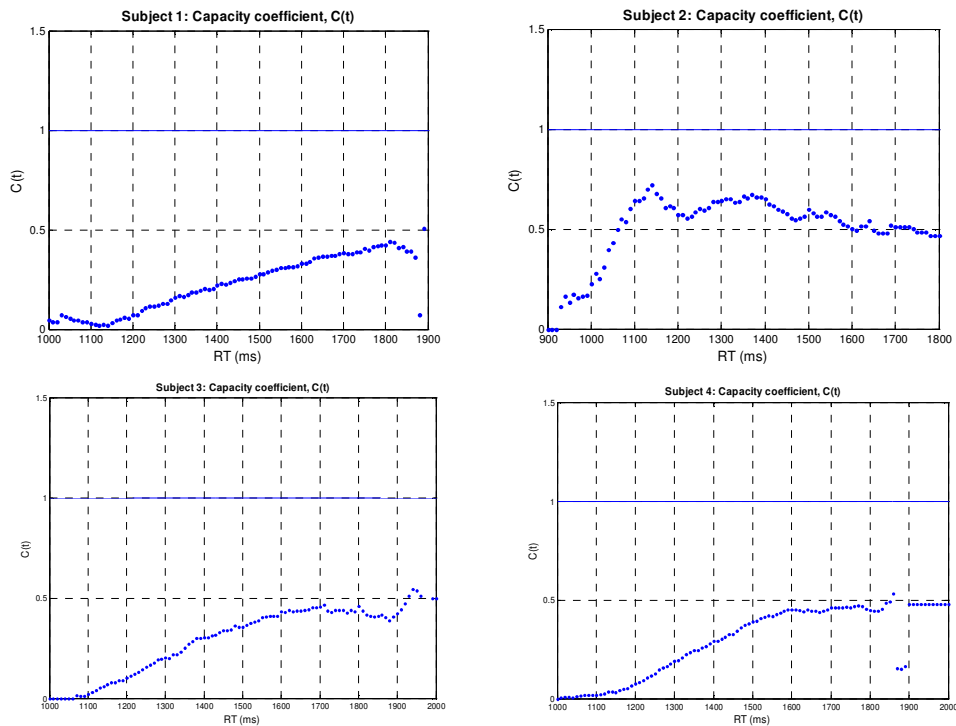
Subject 1’s SIC curve and ANOVA suggests coactive processing in both conditions. The curve is mostly over additive with a degree of negativity for reaction times around 1300 milliseconds. Secondly, the mean interaction is positive (MIC = 38), although the p value from the ANOVA indicates only a trend toward a significant positive interaction ( $p \approx .10$ ).

Subject 2’s SIC curve was entirely over additive, strongly suggesting parallel self-terminating processing. Secondly, this participant’s MIC was positive and the ANOVA indicates a marginally significant interaction between the audio and visual channels.

Subject 3's results indicated coactive processing, both in the over additive SIC curve showing negativity for early processing times, and the positive MIC. Results from the ANOVA also indicated a significant interaction with  $p < .05$ .

The negativity for early processing times in Subject 4's data also suggests coactive processing. The MIC was positive for this subject adding further evidence for coactive architecture rather than serial exhaustive. We did notice that the  $F$  value was low and the  $p$  value did not approach significance suggesting that the power was too low for this particular subject to draw strong conclusions. Nonetheless, these data are consistent and strikingly suggest coactive or parallel processing since the SIC curves and interactions were overwhelmingly positive, with a range of negativity for early processing times for three out of four subjects.

Figure 7 displays the capacity  $C(t)$  plots for all three subjects. The solid flat line at  $C(t) = 1$  represents the bound for super capacity. The plots were consistent in showing that capacity at all points in time was extremely limited. The Grice Lower bound was violated for each subject at nearly every point in time, while the bound for super capacity was not surpassed at any point in time. These data, as in Experiments 1A and 1B, are indicative of an extremely limited capacity or fixed capacity coactive model. In order for a coactive model to predict extremely limited capacity, strong inhibition between auditory and visual channels is necessary. Independent coactive models always produce violations of the unlimited—supercapacity bound and do not violate Grice's inequality for extremely limited capacity (see Townsend & Nozawa, 1995).



**Figure 7.** The Capacity coefficient for each of the four participants in Experiment 2. Processing capacity was extremely limited for each subject, as was the case in Experiments 1A and 1B.

## General Discussion and Conclusion

Experiments (1A & 1B) and 2 were designed to test different models regarding how listeners process audiovisual stimuli in real time. Both versions of Experiment 1 were designed to test how participants processed audiovisual speech stimuli in a detection task, while Experiment 2 examined how participants process audiovisual speech stimuli in a word discrimination task where they were required to distinguish between real spoken words. Recall that previous work on audiovisual speech perception was generally unconcerned with dynamic models of audiovisual perception, and primarily sought to account for accuracy data (Braidá, 1991; Massaro, 2004). In particular, previous research did not attempt to account for how information from the auditory and visual channels was utilized by the “black box” prior to or during the decision process.

Models of dynamic audiovisual speech perception are relevant to current work in the field. Theorists of direct perception and motor theory (Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985), and contrasting theories (Bernstein, 2005) make different claims about how audiovisual information is used during perception and word recognition. Mathematical tools founded upon factorial methodology which make specific claims about reaction time distributions, are an appropriate tools to begin investigating these claims. Factorial methodology was employed to assess the processing architecture and capacity in a detection task and word discrimination task. Our primary focus was analyzing the SIC and determining what form of processing emerged. In addition to investigating the shape of the SIC curves, we looked at the capacity coefficient to determine whether processing time increased, decreased or remained the same when two channels were present relative to the cases when only one channel was present. Experiments 1 and 2 began to reveal how the audio and visual channels are integrated. Data show that the main candidates for processing architecture are parallel with a self-terminating decision rule, or possibly coactive with extreme capacity limitations.

Data from the detection task in Experiment 1A revealed inconsistent results. SIC curves were either inconclusive as to the nature of processing taking place due to the fact that selective influence between conditions was either weak or not present. Processing appeared to be parallel self-terminating, while one subject showed coactivation and the rest demonstrated either serial or indeterminate processing. Experiment 1B, a modified version of Experiment 1A with shorter stimulus durations produced clearer results. Processing for each subject was most likely parallel self-terminating for 2 subjects, while 3 participants showed architecture consistent with coactive processing. Capacity between these two experiments was consistent, where the capacity coefficient  $C(t)$  was overwhelmingly negative for each of the subjects.

Data from the word discrimination task in Experiment 2 showed that processing was either coactive or parallel self-terminating. Capacity coefficients obtained in Experiment 2 revealed extremely limited capacity, which was consistent with the capacity measured in Experiment 1A and 1B. Extremely limited capacity is observed in serial models, and parallel models with negative inhibition, but is not typical of coactive models (Townsend & Nozawa, 1995; Townsend & Wenger, 2004). Hence it is important to begin understanding why coactive architecture indicated by the negativity in the SIC(t) function was observed in conjunction with extremely limited capacity in Experiments 1B and 2. The fact that capacity was extremely limited might indicate strong inhibition or competition between the audio and visual channels. Inhibitory links between channels might begin to explain why extremely limited capacity was observed in conjunction with coactive processing. However, simulations have demonstrated that coactive processing models are usually super capacity even with negative inhibition between channels (Townsend & Nozawa, 1995; Townsend & Wenger, 2004).

It is worth mentioning that extremely limited capacity was observed even though previous studies have consistently observed “audiovisual enhancement” in accuracy scores when audiovisual conditions were compared to audio only conditions (Sumbly & Pollack, 1954). Audio and visual processing channels might simultaneously engage in inhibition (slowing the system down) while enhancing the quality of the information at the decision stage.

It is also important to continue investigating the nature of the limitations in processing capacity obtained in these experiments. If limitations in audiovisual processing capacity result from between channel inhibition, it would be worthwhile to understand how this inhibition might be manipulated or offset. Recent research involving discrimination of the numerals “1” and “2” in the visual modality with congruent speech stimuli indicates that manipulating the SOA (the lead of the visual stimuli in milliseconds) might decrease capacity limitations. At SOAs of 150 milliseconds or more, “redundant target effects” (i.e., supercapacity) were observed, which might indicate coactive processing (Berryhill et al., 2007).

Another worthwhile future direction will be to explore capacity and processing architecture using incongruent audiovisual stimuli as in the McGurk effect. The use of incongruent audiovisual stimuli will allow investigators to explore how audiovisual inhibition as indicated by the capacity coefficient  $C(t)$  might be enhanced or otherwise altered, and explore whether processing architecture remains consistent with cases where the audio visual are congruent in both AND as well as OR experimental designs.

## References

- Bernstein, L.E., (2005). Phonetic perception by the speech perceiving brain. In D.B. Pisoni & R.E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 79-98). Malden, MA: Blackwell Publishing.
- Bernstein, L.E., Auer, E.T., & Moore, J.K. (2004). Audiovisual speech binding: Convergence or association? In G.A. Calvert, C. Spence & B.E. Stein (Eds.), *Handbook of Multisensory Processing* (pp. 203-223). Cambridge, MA: MIT Press.
- Berryhill, M., Kveraga, K., Webb, L., & Hughes, H. C. (2007). Multimodal access to verbal name codes. *Perception & Psychophysics*, *69*, 628-640.
- Braida, L.D. (1991). Crossmodal Integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology*, *43A*, 647-677.
- Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593-596.
- Calvert, G.A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*, 57-70.
- Calvert, G.A., Campbell, R., & Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649-657.
- Fowler, C.A., & Dekle, D.J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 816-828.
- Fowler, C.A., & Rosenblum, L.D. (1991). Perception of the phonetic gesture. In I.G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (pp. 33-59). Hillsdale, NJ: Lawrence Erlbaum.
- Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding. *Journal of the Acoustical Society of America*, *112*, 30-33.
- Grant, K.W., Tufts, J.B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing impaired individuals. *Journal of the Acoustical Society of America*, *121*, 1164-1176.

- Green, K.P., & Miller, J.L. (1985). On the role of visual rate information in phonetic perception. *Perception and Psychophysics*, *38*, 269-276.
- Lieberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception. *Cognition*, *21*, 1-36.
- Massaro, D.W. (1987). Speech perception by ear and eye. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53-83). Hillsdale, NJ: Lawrence Erlbaum.
- Massaro, D.W. (2004). From multisensory integration to talking heads and language learning. In G.A. Calvert, C. Spence & B.E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 153-176). Cambridge, MA: The MIT Press.
- McGurk, H., & McDonald, J.W. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- Meredith, M.A. (2002). On the neuronal basis for multisensory convergence: A brief overview. *Cognitive Brain Research*, *14*, 31-40.
- Miller, J., Kuhlwein, E., & Ulrich, R. (2004). Effects of redundant visual stimuli on temporal order judgments. *Perception & Psychophysics*, *66*, 563-573.
- Puce, A., Allison, T., Bentin, S., Gore, J.C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience*, *18*, 2188-2199.
- Rosenblum, L.D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51-78), Malden, MA: Blackwell Publishing.
- Schroter, H., Ulrich, R., & Miller, J. (2007). Effects of redundant auditory stimuli on reaction time. *Psychonomic Bulletin & Review*, *14*, 39-44.
- Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. *Acta Psychologica*, *30*, 276-315.
- Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212-215.
- Townsend, J.T., & Fific, M. (2004). Parallel versus serial processing and individual differences in high-speed search in human memory. *Perception and Psychophysics*, *66*, 953-962.
- Townsend, J.T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*, 321-359.
- Townsend, J.T., & Wenger, M.J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review*, *111*, 1003-1035.
- Wenger, M.J., & Townsend, J.T. (2000). Basic response time tools for studying general processing capacity in attention, perception, and cognition. *The Journal of General Psychology*, *127*, 67-99.