Atten Percept Psychophys (2016) 78:1712–1727
DOI 10.3758/s13414-016-1133-4

The McGurk effect: An investigation of attentional capacity employing response times

Nicholas Altieri[1] · Jennifer J. Lentz[2] · James T. Townsend[3] · Michael J. Wenger[4]

Published online: 17 May 2016

Abstract This paper proposes a novel approach to assess audiovisual integration for both congruent and incongruent speech stimuli using reaction times (RT). The experiments are based on the McGurk effect, in which a listener is presented with incongruent audiovisual speech signals. A typical example involves the auditory consonant/b/combined with a visually articulated/g/, often yielding a perception of/d/. We quantify the amount of integration relative to the predictions of a parallel independent model as a function of attention and congruency between auditory and visual signals. We assessed RT distributions for congruent and incongruent auditory and visual signals in a within-subjects signal detection paradigm under conditions of divided versus focused attention. Results showed that listeners often received only minimal benefit from congruent auditory visual stimuli, even when such information could have improved performance. Incongruent stimuli adversely affected performance in divided and focused attention conditions. Our findings support a parallel model of auditory-visual integration with interactions between auditory and visual channels.

✉ Nicholas Altieri
altinich@isu.edu

[1] Department of Communication Sciences and Disorders, Idaho State University, 921 S. 8th Ave., Stop 8116, Pocatello, ID 83209, USA

[2] Department of Speech and Hearing Sciences, Indiana University, Bloomington, IN, USA

[3] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA

[4] Department of Psychology, The University of Oklahoma, Norman, OK, USA

Keywords Speech perception · Attention: Selective · Reaction time methods

What cognitive mechanisms underlie speech recognition when audition is supplemented with visual information? The modern era of research into how auditory and visual speech cues interact began with Sumby and Pollack's (1954) seminal experimental work on audiovisual (AV) enhancement: They showed that visual cues provided by a talker's lip-movements facilitate auditory recognition across a range of signal-to-noise ratios. However, Massaro (1987a) proved that this outcome did not necessarily demonstrate integration since a single-channel model could theoretically predict the results.

In a critical study two decades later, McGurk and Macdonald (1976) reported a dramatic perceptual integration phenomenon that resulted from the presentation of incongruent auditory-visual speech signals. In what became known as the "McGurk effect," presentation of the auditory consonant/b/over a visually articulated/g/yielded a fused percept of/d/. Audiovisual fusions such as these occur when the perceptual system maps cues from conflicting signals onto a phonemic category distinct from either input signal.[1] Thus, the McGurk effect is a prime candidate with which to probe the mechanisms underlying integration.

Several studies of the McGurk effect have been carried out, with the majority using mean accuracy as the dependent variable. In these studies, performance in auditory and visual-only trials is compared to accuracy in audiovisual trials, usually via confusion matrices (e.g., Massaro, 1987a, 1998, 2004). These experimental designs and modeling efforts have shed considerable light on speech integration: Auditory and visual cues appear to interact in a multiplicative manner

[1] This cannot be demonstrated conclusively without also measuring responses to the single modality presentations (cf. Massaro 1987a, 2004).

according to Massaro's fuzzy logical model of perception (FLMP), even if the information is extracted independently. It is important to note that Massaro often analyzed data from individual listeners and sometimes supplemented accuracy with RTs by showed that processing times can be estimated via a measure of ambiguity in the FLMP framework (e.g., Massaro, 1998). Accuracy by itself, however, is limited in the set of questions it can address. In general, accuracy is more informative than RT in testing for various kinds of informational dependencies, while RT is stronger than accuracy for assessing characteristics of information processing architecture (see, e.g., Algom, Eidels, Hawkins, Jefferson, & Townsend, 2015; Massaro, 1987b; Townsend & Nozawa, 1995).[2] In the context of audiovisual integration, recent research using RT distributions has indicated that speech integration, using congruent signals, is perhaps best described by a parallel interactive process rather than coactivation or statistical facilitation (see Altieri & Townsend, 2011, for further discussion). The question remains: Do these findings change for McGurk stimuli, or is attention allocated differently across modalities?

Our study begins to fill the broad gap in knowledge regarding high-accuracy audiovisual speech performance. It uses RT distributions to investigate how integration efficiency, assessed using a measure of *capacity* (Altieri & Townsend, 2011; Altieri & Wenger, 2013; Townsend & Nozawa, 1995) varies as a function of audiovisual congruency demands on attention. We define *capacity* as the efficiency of which pieces of information are recognized, as a function of the number of pieces of information available to the observer (Townsend & Nozawa, 1995; Wenger & Gibson, 2004). Here, we manipulate "number of pieces of information" by varying number of modalities present (i.e., auditory visual vs. auditory only or visual only). The construct of capacity differs from comparisons using mean RTs. Specifically, capacity compares the RT distribution obtained from trials where both auditory and visual cues are presented, to the RT distribution specifying independent race model predictions. (The independent parallel race model predictions are derived from trials containing only auditory or visual (i.e., lip-reading) information). It is also important to point out that the capacity measure is rich in dynamic information since it depicts efficiency as a direct function of time. As defined by Townsend and Nozawa (1995), this type of parallel model possesses unlimited capacity, in that increasing the amount of perceptual work neither improves nor degrades performance; hence, they are called *unlimited capacity independent parallel* (UCIP) *models*. This model class is basically what investigators seem to mean by "parallel processing." Hence, another name for one is *standard parallel model*. Two critical advantages of our capacity methodology are (1) that it addresses performance at the level of the entire RT distribution, rather than just the mean, and (2), that it makes no parametric assumptions regarding the distribution of RTs (i.e., RTs need not follow a particular distribution, such as a standard normal or exponential).

Theoretical accounts of audiovisual speech integration

As previously discussed, quantitative models have been devised to predict accuracy (e.g., Braida, 1991; Massaro, 1987a) as well as RTs (Massaro, 1998) from speech recognition data; some even put forth evidence that integration occurs in parallel (e.g., Massaro, 1987b). Nonetheless, much of the discussion specifying the dynamic flow of information during audiovisual speech processing have been somewhat informal, and a typical approach has been to pose two extreme accounts against one another (see Bernstein et al., a, b, for discussion). For example, the *parallel* model posits that integration occurs late in processing, after consonant identification has taken place in the separate (parallel) visual and auditory channels. The opposing *common format* model assumes that auditory and visual phonemic information is integrated early in processing to form a single variable. Since integration occurs early according to this framework, data patterns should be similar regardless of whether attention is divided across modalities or focused on only one (e.g., the auditory).

Parallel channel models

The *parallel specific* model assumes separate, categorical decisions are made on the parallel auditory and visual channels prior to integration. In its purest form, this model assumes that the parallel channels do not interact at all, that is, they are neither mutually inhibitory nor facilitatory.

Another critical aspect of a model is when processing is assumed to cease. For instance, if the observer is searching for a target among distractors and one is present, it is most efficient for her or him to stop as soon as the target is located. This is called *self-termination*. An important special case of self-termination occurs when all items are targets; then, it is most efficient to stop as soon as the first item is completed. If the observer is told to respond to whichever target is found, we have the classic OR condition. The associated stopping rule is *first-terminating* or *minimum time* stopping. For instance, in our divided attention condition, the observer can stop as soon as either the auditory or the visual target is found. However, suppose processing can finish only when phonemic recognition occurs in both the auditory and visual modalities; in this case, observed RTs will be determined by the slower of the

[2] Note that it is possible to design experiments using accuracy that are useful in the study of processing architecture (e.g., Scharff, Palmer, & Moore, 2011; Shiffrin & Gardner, 1972; Townsend, 1981).

two channels. This is referred to as an *exhaustive* or *maximum-time* rule. An experimental condition where exhaustive processing is demanded is called an AND stopping rule condition. Evidence from audiovisual detection studies suggests that audiovisual speech perception generally utilizes an OR rule when conditions permit (e.g., Altieri & Townsend, 2011; Massaro, 1987b).

Braida's (1991) prelabeling integration model (PRE) is an example of a late integration, modality-specific model because it treats the two inputs as random observations on orthogonal dimensions. The model assumes that decisions are made about the value of each dimension separately. So, even without a temporal processing account, the model clearly interprets the dimensions as being perceived and decided on with no perceptual or other cognitive interactions. Another example of a parallel model where information is obtained in separate modalities is Massaro's FLMP. (The FLMP appears to assume that auditory and visual cues are extracted from the signal independently, and combined multiplicatively.)

To date, the evidence favors parallel interactive rather than independent parallel processing (e.g., Altieri & Townsend, 2011; Eidels et al., 2011; Massaro, 1987b; Townsend & Altieri, 2012; Townsend, Houpt, & Silbert, 2012). For example, Buchan and Munhall (2011) observed that selectively attending to the auditory modality significantly increased the proportion of auditory responses, but the influence of the visual modality on auditory processing could not be completely eliminated. Furthermore, engaging in dual tasks that tax cognitive resources or otherwise draw attention away from linguistic stimuli (Hessler, Jonkers, Stowe, & Bastiaanse, 2013; Tiippana, Andersen, & Sams, 2004) has been shown to attenuate the influence of the visual signal and bias responses to the auditory modality. In addition, evidence from our work suggests positive or negative dependencies between the two modalities depending on auditory signal clarity (Altieri & Townsend, 2011).

Common format coactive models

An alternative to the modality-specific parallel model is the *common format coactive* model. This model assumes that, early in the integration process, the auditory and visual inputs are transformed into a common format where they can be directly integrated. A simple representation of integration would be the summation of the two signals, and models of this type are known as *coactivation* models (Townsend & Nozawa, 1995). The summation results in a single evidence variable that can be compared to a single decisional criterion, thus obviating the need to consider OR or AND stopping rules.

Some studies do appear show coactivation in AV integration. For example, Green and Kuhl (1991) and Soto-Faraco et al. (2005) examined audiovisual integration using a Garner interference task. In both studies, results demonstrated that the visual signal influenced auditory processing by altering the response speed of the auditory percept. Such findings led the experimenters to argue that auditory and visual stimuli are integral or preattentive: that is, they are combined early in processing into a single common channel. Generally, observations that the McGurk effect occurs regardless of whether a listener is aware of the incongruence suggest that preattentive cognitive mechanisms mediate speech integration (e.g., Green & Kuhl, 1991; Soto-Faraco, Navarra, & Alsius, 2004; Sekiyama, Soshi, & Sakamoto, 2014). As another example, Massaro has provided evidence that the level of integration was constant across conditions of divided and focused attention. While such findings appear to add prima facie support for coactivation, the data are perhaps most consistent with parallel processing because separate auditory and visual information have been shown to be retained even after speech categorization (cf. Massaro, 1987b).

This issue of whether AV integration is parallel or coactive is further complicated by findings that the reliability of visual information affects the use of multimodal signals. Nahorna, Berthommier, and Schwartz (2012) showed that the strength of the McGurk effect could be reduced when the audiovisual test stimulus was preceded by a stream of syllables in which the visual speech cues conflicted phonetically or temporally with the auditory stimulus. Hence, the binding of auditory to visual speech information appeared to be affected by the prior predictability of the visual stimulus. In other words, when visual cues fail to provide relevant information, they appear to be weighted less during the early processing stages, perhaps even prior to the influence of top-down attention (e.g., Massaro, 1998; van Wassenhove, Grant, & Poeppel, 2005; van Wassenhove, 2013).

Formal coactive and parallel models of speech integration

The discussion so far suggests that the competing hypotheses of modality-specific and common format processing can be understood in terms of a contrast between parallel and coactive processing architectures. Specifically, we frame these contrasts using the formal analyses of these models provided by Townsend and colleagues (Townsend & Nozawa, 1995; Townsend & Wenger, 2004a, b). Figure 1 presents schematic representations of three models to be considered. Each model contains the same components of processing: (A)uditory and (V)isual feature analysis, a threshold on the accumulation of evidence, and finally a decision stage. The three alternatives are distinguished by the arrangement of those components. Critically, these three models can be empirically distinguished with respect to their predictions for measures of capacity (Townsend & Nozawa, 1995; Townsend & Wenger, 2004a, b) or integrated hazard ratios. The model in Fig. 1a represents the case of early integration; this is the common-format coactive. Figure 1b and c represent two versions of "late integration"; these are variants of the modality-specific parallel channels model. Figure 1a (coactive) illustrates that the auditory and visual processing are summed together prior to a decision. Figure 1b is a parallel system in which the detection of a target stimulus feature in either the auditory or visual channel can be made independent of the other channel. Figure 1c presents a parallel system in which attention can place greater weight on the auditory channel to effectively reduce the impact of the visual channel or vice versa. Complete attenuation of one channel (e.g., the visual) would lead to the prediction that auditory-visual performance would be identical to auditory-alone performance.

The models function as follows: In coactive models, sensory information is first processed in separate modalities. Next, the information is combined (i.e., summed together) in a common processor. This stage is crucial for coactive models because the categorization judgment is based upon summed or amalgamated sensory information. This summation of information generally elicits processing times that are substantially faster on audiovisual (or redundant target) trials compared to auditory or visual-only (single-target) trials (e.g., Townsend & Wenger, 2004b). In parallel "can't focus" models where independence assumptions are violated, auditory and visual cues are processed in separate modalities. Next, the information in each modality is compared to its threshold: An auditory or visual judgment is triggered, say, when enough information is accrued for the existence of the target phoneme. The final decision depends on the stopping rule used to combine the information from the separate channels and their associated decisional operators. In our case, we use an OR rule that stipulates that a final decision is made once threshold is reached in either the auditory or visual channel. Crucially, UCIP models predict slightly faster RTs when congruent audiovisual stimuli are presented relative to auditory or visual only (Miller, 1982). This statistical effect, known as the *redundancy gain*, occurs because targets are present in two as opposed to only one channel. Inhibition or facilitation between channels, when present, can have a modest influence on audiovisual RTs relative to auditory or visual-only RTs. Finally, the parallel "can focus" model (where the participant can focus on the auditory modality independent of the visual) contains the same basic set-up as the parallel cannot focus model. The difference is that processing can be ignored in, say, the visual modality. Therefore, this model predicts that audiovisual RTs should be identical to the unisensory auditory RTs.

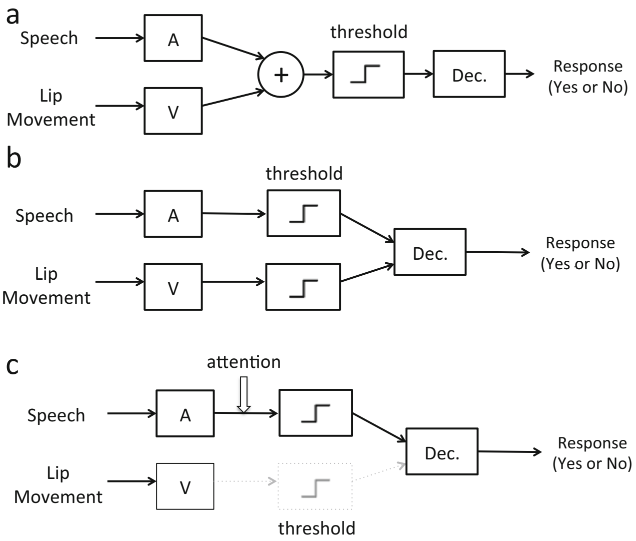

Fig. 1 Schematic representations of three alternative models: (a) a coactive model, (b) a parallel model in which the participant cannot focus attention, and (c) and a parallel model in which the participant can focus attention on the auditory modality. Note that both parallel models contain an OR decision rule that allows for detection as soon as enough auditory (or visual) information reaches detection threshold. The letters *A* and *V* represent auditory and visual inputs, respectively. The light and dotted lines for the V input (c) indicate that while a visual input is present, the listener can effectively ignore the content, thus making it irrelevant during the decision stage

Measures of capacity and relative efficiency

Our approach to characterizing *capacity* is based on the comprehensive characterization of this issue in the work of Townsend and colleagues (Townsend & Altieri, 2012; >Townsend & Ashby, 1978; Townsend & Nozawa, 1995; Townsend & Wenger, 2004a, b). Our capacity measures will be applied at the level of the individual participant to characterize the processing strategies used by a single observer. The principle motivation for this practice is that group average curves, such as integrated hazard functions, can obscure the functional form of the curve generated by an individual (Estes, 1956). Importantly, it is possible that different listeners will use differing strategies. Such variation may depend on hearing ability, cognitive status, and susceptibility to the McGurk effect; all of which are known to influence multisensory integration abilities (e.g., Altieri & Hudock, 2014; Erber, 2003; Sommers, Tye-Murray, & Spehar, 2005).

First, capacity is assessed by comparing the empirical RT distributions from trials in which both auditory and visual information are presented to the RT distributions obtained from trials in which auditory and visual-only information is present. This is done using integrated hazard functions, as follows. The hazard function for a response time distribution is a conditional probability function, defined formally as $h(t) = \frac{f(t)}{1-F(t)}$. It is interpreted as the probability that a response will occur in the next instant of time, given that it has not yet occurred. It reflects the intensity with which a process is proceeding, or the instantaneous amount of work being accomplished. As such, the hazard function is also referred to as the *intensity* function (e.g., Townsend & Ashby, 1978). The integral, the integrated hazard function can be interpreted as a cumulative measure of the amount of work accomplished. An important regularity regarding both the hazard function and its integral is that, if two processes are independent, then the sum of the hazard functions (and their integrals) is equal to the joint (combined) hazard function (and its integral) for the two processes (Luce, 1986; Townsend & Nozawa, 1995), when the two independent processes are joined by an OR decision rule. Townsend and Nozawa (1995) used this regularity to define the capacity coefficient for independent parallel systems using an OR decision rule:

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

The term $H_{AV}(t)$ in the numerator represents the integrated hazard function obtained from audiovisual trials, while the denominator is the sum of the integrated hazard functions from the auditory-only($H_A(t)$) and visual-only ($H_V(t)$) trials.

Because $C(t)$ is a ratio, there are three possible values for the function: First, $C(t)$ can equal 1 for a range of RTs. This is equal to UCIP predictions and means that RT distributions obtained from audiovisual trials correspond exactly to the minimum time prediction derived from the RT distributions from the auditory and visual-only trials. Second, $C(t)$ may be less than 1 for that range of RTs, indicating that an increase in perceptual workload (audiovisual trials versus UCIP predictions derived from unisensory trials) degrades RT performance. This scenario describes inefficient audiovisual integration, suggesting that either there are inhibitory cross-channel mechanisms or limitations in the attentional resources and is referred to as limited-capacity processing (see Altieri & Wenger, 2013). Third, $C(t)$ may be greater than 1, meaning that performance is superior to that predicted by a standard parallel model. This finding describes highly efficient integration because RTs to AV stimuli are faster than would be predicted by statistical summation and UCIP predictions. Supercapacity is consistent with facilitatory cross-channel dependencies, and extreme supercapacity is consistent with coactivation (Eidels et al., 2011; Townsend & Wenger, 2004a; Wenger & Townsend, 2006).

We can also use the integrated hazard functions to assess the relative efficiency with which observers detect the presence of potential cross-channel interactions in conditions that require them to focus on only one modality. Comparing integrated hazard functions is useful for comparing processing rates across two types of experimental stimuli, specifically

$$Hazard\_Ratio = \frac{H_{S1}(t)}{H_{S2}(t)}. \tag{2}$$

Here *S*1 and *S*2 refer to two stimulus types. Using this metric, we can assess a number of questions, including whether congruent or incongruent information speeds up or slows down processing relative to a single modality (e.g., H_{AVC}/A_{Alone} or H_{AVI}/A_{Alone}, where AVC is AV congruent, and AVI is AV Incongruent).

Hypotheses for capacity and integrated hazard ratios

In the present study, we use capacity coefficients and integrated hazard ratios in a task requiring participants to detect the consonant/b/using auditory-only, visual-only, or audiovisual information. The audiovisual trials contained each factorial combination of the consonants/b/and/g/: $A_{/b/}V_{/b/}$, $A_{/b/}V_{/g/}$which normally yields a percept of auditory/d/, $A_{/g/}V_{/b/}$which yields a percept of the consonant cluster/b/+/g/, and $A_{/g/}V_{/g/}$. For the divided attention condition, observers were required to give a "yes" response if the target phoneme/b/is present in the auditory or visual modalities. In the focused attention condition, observers were required to only give a "yes" response if the target phoneme/b/occurs in the auditory modality.

Table 1 summarizes the predictions for the capacity coefficient and integrated hazard ratios for the coactive and standard parallel models in Fig. 1. Only the coactive (Fig. 1a) and the parallel model without attention (Fig. 1b) are considered here, as the divided attention experiment cannot elucidate the effects of attention on audiovisual perception. An early-integration coactive model predicts extreme supercapacity processing (i.e., $C(t) >> 1$), because if attention occurs after integration, processing will be impervious to attentional manipulations. A pure modality-specific model consistent with the UCIP model predicts that $C(t)$ =1. Thus, we consider predictions for these two models—coactive and UCIP—as they represent the extremes of the AV processing architecture.

With respect to the integrated hazard ratios, the coactive model predicts ratios magnitudes greater than 1, since for AV redundant stimuli evidence becomes fused, thereby contributing to very fast responses relative to unisensory stimuli. The parallel model predicts that the hazard ratios with a congruent AV numerator (e.g., A_BV_B/A_B) to be greater than 1 simply because there are more targets available for the congruent stimuli (i.e., if they "miss" the auditory signal, a visual cue is still available). If both targets yield similar reaction times, these ratios would have values around 2 for standard parallel models but much greater than 2 for coactive models. We can also assess model comparisons using integrated

Table 1 Capacity and integrated hazard ratio predictions for the divided and focused attention studies. The B and G represent integrated hazard functions computed for trials containing the auditory or visual phonemes/b/and/g/

Comparison	Coactive	Parallel (Can Focus)	Parallel (Cannot Focus)
Divided attention			
$A_BV_B/[A_BV_{B\,+}\,A_BV_B]$	$C(t) >> 1$	$C(t) \approx 1$	$C(t) \approx 1$
A_BV_B/A_B	*Ratio* >> 1	*Ratio* ≈ 1	*Ratio* > 1
A_BV_B/A_BV_G	*Ratio* >> 1	*Ratio* > 1	*Ratio* > 1
A_BV_G/A_GV_B	*Ratio* ≈ 1	*Ratio* ≈ 1	*Ratio* ≈ 1
Focused attention			
A_BV_B/A_B	*Ratio* >> 1	*Ratio* ≈ 1	*Ratio* > 1
A_BV_B/A_BV_G	*Ratio* >> 1	*Ratio* ≈ 1	*Ratio* > 1

hazard ratios through the incongruent pairings, such as the ratio A_BV_G/A_GV_B. If A_BV_G and A_GV_B yield fused percepts, this ratio should equal approximately 1 if both channels are weighted equally. On the other hand, differential weighting of A versus V would yield capacity ratios that differ from one.

Predictions for hazard ratios in the focused attention condition are also displayed in Table 1. Fewer comparisons are required for the focused attention experiment, as it (relative to the divided attention experiment) contains fewer stimulus configurations associated with a “yes” response. We consider two variants of the parallel models: one in which observers can attenuate the effects of the visual signal and one in which they cannot. The coactive model predicts that the integrated hazard ratios will be much greater than 1. For the parallel model, different predictions arise depending on whether the listener can focus on a single modality. If listeners are capable of selectively attending to auditory cues while ignoring visual information, the integrated hazard function ratios A_BV_B/A_B and A_BV_B/A_BV_G should equal 1 as the visual information should neither help nor hurt the perception. Conversely, if listeners are unable to ignore visual information, then the integrated hazard ratio without incongruent distractors (A_BV_B/A_B) should be greater than one. Here, the observer cannot ignore the helpful visual information. The ratio A_BV_B/A_BV_G would also be expected to be greater than 1. Because the visual/g/provides information conflicting with the auditory/b/, the listener may require more time to correctly identify the auditory component, thereby slowing the RTs in the condition in the denominator.

Method

Participants

Five right-handed college-age listeners recruited from the Indiana University (three females), Bloomington, community were recruited as participants. Participants were paid $8.00 per hour for participation. This study was approved by the Indiana University Institutional Review Board.

Stimuli

The stimuli were digitized movie clips of a female talker speaking the syllables/be/and/ge/and were obtained from the Hoosier Multi-Talker Database (Sherffert, Lachs, & Hernandez, 1997). The/be/movie clip consisted of fourteen 30-ms frames, and the/ge/movie files contained 16 frames. The stimuli were edited using Final Cut Pro HD and factorially combined on AV trials to create: $A_{/b/}V_{/b/}$, $A_{/b/}V_{/g/}$, $A_{/g/}V_{/b/}$, and $A_{/g/}V_{/g/}$trials. $A_{/b/}$, $V_{/b/}$, $A_{/g/}$, and $V_{/g/}$trials were also presented.

Procedure

In the divided attention condition, participants were instructed to give a “yes” response, by pressing the right mouse button as quickly and as accurately as possible if they perceived either an auditory/b/or visual/b/. On any other trial, they were instructed to give a “no” response by pressing the left mouse button. In the focused attention condition, participants were instructed to look at the computer monitor but attempt to ignore the visual signal. This task required them to give a “yes” response only if there was an auditory/b/and a “no” response otherwise. The response mappings are shown in Table 2. There were a total of 480 trials for each of the eight stimulus configurations.

Trials began with a fixation cross appearing in the center of the computer screen for 500 ms followed by the stimulus. RTs were measured from stimulus onset, and RTs greater than 3, 000 ms or less than 100 ms were removed from the data (<1 % of trials). The experiment took place over six 1-hour sessions on separate days. The first 3 days were either divided or focused attention blocks (counterbalanced across listeners). Participants were presented with 48 practice trials at the beginning of each day with feedback in order to facilitate the learning of correct response mappings. Feedback involved

Table 2 The "yes" and "no" response mappings for the different stimulus categories in the divided and focused attention experiments. The ∅ symbol indicates that nothing was presented in that modality

Auditory	Visual	Percept	Response divided	Response focused
/b/	/b/	/b/	Yes	Yes
/b/	/g/	/d/or/ th/	Yes	Yes
/g/	/b/	/b/and/ g/	Yes	No
/g/	/g/	/g/	No	No
/b/	∅	/b/	Yes	Yes
/g/	∅	/g/	No	No
∅	/b/	/b/	Yes	No
∅	/g/	/g/	No	No

either the word *correct* or *incorrect* being displayed on the computer monitor after each trial. Feedback was not provided on the experimental trials. Although participants acknowledged perceiving a/d/in $A_{/b/}V_{/g/}$trials, they were aware that $A_{/b/}$was present due to the task instructions and feedback.

Results

Summary statistics

Divided attention

For these analyses, accuracy data were analyzed using repeated measures ANOVAs. Mean audiovisual, auditory, and visual-only accuracy scores for the "yes" responses in the divided attention condition were generally high. The ANOVA results and mean accuracy scores are shown in Tables 3 and 4, respectively. Due to heterogeneity of the variances, accuracy scores were transformed using an arcsine transformation.

Overall, results revealed significant differences across stimuli, in which higher accuracy was observed in the congruent AV trials $A_{/b/}V_{/b/}$compared to the incongruent $A_{/b/}V_{/g/}$"McGurk" stimulus. More accurate responses were also observed for $A_{/b/}V_{/b/}$stimuli relative to $A_{/b/}$, and also relative to $V_{/b/}$. Next, mean RTs for each participant (for correct responses) were analyzed using repeated-measures ANOVAs. ANOVA results and mean RTs averaged across participants are shown in Tables 5 and 6, respectively.

Table 4 Mean accuracy in each condition

Experimental condition	Mean % correct
$A_{/b/}V_{/b/}$	99
$A_{/b/}$	91
$A_{/b/}V_{/g/}$	88
$A_{/g/}V_{/b/}$	92
$V_{/b/}$	92

First, we carried out a comparison of mean RTs in the congruent $A_{/b/}V_{/b/}$trials to the incongruent $A_{/b/}V_{/g/}$"McGurk" trials. Results averaged across participants indicated comparatively faster congruent AV RTs compared to incongruent "McGurk" trials. Next, we compared the congruent $A_{/b/}V_{/b/}$mean RTs to the single target trials. In contrast to the accuracy results, the results failed to show evidence for a facilitation in $A_{/b/}V_{/b/}$responses compared $A_{/b/}$or $V_{/b/}$.

Focused attention

Results for the repeated-measures ANOVAs and mean accuracy (% correct) are displayed in Tables 7 and 8, respectively. Accuracy in the focused attention trials for the "yes" responses was high. To investigate whether differences in accuracy were present across conditions, we carried out ANOVAs comparing mean accuracy for $A_{/b/}V_{/b/}$versus $A_{/b/}V_{/g/}$trials, and again for $A_{/b/}V_{/b/}$versus $A_{/b/}$trials. Results showed evidence for a significant difference between $A_{/b/}V_{/b/}$versus $A_{/b/}V_{/g/}$trials, although not for $A_{/b/}V_{/b/}$versus $A_{/b/}$trials.

Finally, we carried out an ANOVA comparing mean RTs across relevant conditions. These results are displayed in Table 9, and mean RTs averaged across participants are displayed in Table 10.

Results show a marginal though not significant trend toward faster congruent AV RTs compared to the incongruent trials. We also compared the congruent $A_{/b/}V_{/b/}$mean RTs to the A-only mean RTs. Results point to a modest slowdown in $A_{/b/}V_{/b/}$responses compared to the $A_{/b/}$trials. This indicates

Table 3 Repeated measures ANOVAs for mean accuracy

ANOVA test	*F* value	*df* between	*df* within	*p* value
$A_{/b/}V_{/b/}$vs. $A_{/b/}V_{/g/}$vs. $A_{/g/}V_{/b/}$vs. $A_{/b/}V_{/g/}$	3.22	4	16	.04*
$A_{/b/}V_{/b/}$vs. $A_{/b/}V_{/g/}$	70.53	1	4	<.001**
$A_{/b/}V_{/b/}$vs. $A_{/b/}$	35.42	1	4	<.001**
$A_{/b/}V_{/b/}$vs. $V_{/b/}$	14.04	1	4	.02*

* denotes significance at α = .05, ** at α = .01, and *** at α = .001.

Table 5 Repeated-measures ANOVAs for mean RT

ANOVA test	*F* value	*df* between	*df* within	*p* value
$A_{/b/}V_{/b/}$vs. $A_{/b/}V_{/g/}$	10.90	1	4	.03*
$A_{/b/}V_{/b/}$vs. $A_{/b/}$	<1.00	1	4	.33
$A_{/b/}V_{/b/}$vs. $V_{/b/}$	<1.00	1	4	.46

* indicates significant $p < .05$

that participants failed to benefit from congruent visual cues when they were not pertinent for the task. To test whether the McGurk effect was stronger in the divided as opposed to the focused attention condition, we carried out an ANOVA on the 2 × 2 interaction between congruency (congruent vs. incongruent) and attention (focused vs. divided). The interaction was significant, indicating a greater congruency facilitation in the divided (988 ms vs. 1,247 ms) compared to the focused (1,081 ms vs. 1,194 ms) attention condition.

Divided attention: Capacity

Unlike those applied to mean RT or mean accuracy, significance tests for the capacity coefficients are not computed using parametric measures. Traditionally, capacity has been assessed by comparing calculated capacity values to upper and lower bounds (Townsend & Eidels, 2011; Townsend & Nozawa, 1995; Townsend & Wenger, 2004a, b). These bounds have been translated into capacity space by Townsend and Eidels (2011), and appear in Fig. 2. These comparisons of $C(t)$ to theoretical bounds do not rely on parametric assumptions for RT distributions. In terms of statistical tests, Houpt and Townsend (2012) showed that semiparametric estimates of integrated hazard functions could be used to compute a Z statistic comparing data to race model predictions derived from unisensory trials. Figure 2 shows the values of the capacity coefficient (Eq. 1) for each participant. Recall that we only used the capacity coefficient for the congruent stimuli: $A_BV_B/(A_B + V_B)$. Note that in each of the following figures, capacity, and subsequently integrated hazard function ratio values are displayed for time points in which there is overlap between the audiovisual, auditory, and visual-only RT distributions. In Fig. 2, the solid line and dotted line represent the upper and lower bounds, respectively, for parallel independent model predictions.

Table 6 Mean RTs in each condition

Experimental condition	Mean RT (*SD*) ms
$A_{/b/}V_{/b/}$	988 (111)
$A_{/b/}$	1,071 (208)
$A_{/b/}V_{/g/}$	1,247 (192)
$V_{/b/}$	1,004 (110)

Table 7 Repeated-measures ANOVAs for mean accuracy

ANOVA test	*F* value	*df* between	*df* within	*p* value
$A_{/b/}V_{/b/}$vs. $A_{/b/}V_{/g/}$	8.79	1	4	.02*
$A_{/b/}V_{/b/}$vs. $A_{/b/}$	1.48	1	4	.18

* indicates significant $p < .05$

Table 8 Mean accuracy in each condition.

Experimental condition	Mean % correct
$A_{/b/}V_{/b/}$	97
$A_{/b/}$	94
$A_{/b/}V_{/g/}$	82

Results showed that capacity was limited (<1) for each of the participants. Notably, all capacity functions also increased as RTs increase, illustrating powerful dynamics in AV perception. The statistical tests from Houpt and Townsend (2012) indicated that capacity was significantly lower than UCIP predictions for all participants, $C(t) = 1$, $Z < -100$, $p < .0001$. None of the capacity functions approach the upper Miller bound, suggesting a strong rejection of a pure coactivation model for all individual listeners. The capacity functions for all participants, especially 3, 4, and 5 fell below the lower Grice bound for some time points, implying severely limited capacity. This was true especially for the fast RTs. Those data that fall below the Grice bound show than RTs are slower for $A_{/b/}V_{/b/}$than the faster of $A_{/b/}$or $V_{/b/.}$ Interestingly, RTs were actually harmed by the congruent audiovisual information in this high-accuracy setting. Thus, because capacity ranges from moderately limited to severely, we can conclude that there is no evidence for either a strongly coactive model or a parallel independent channels model with unlimited capacity.

Table 9 Repeated-measures ANOVAs for mean RT

ANOVA test	*F* value	*df* between	*df* within	*p* value
$A_{/b/}V_{/b/}$vs. $A_{/b/}V_{/g/}$	3.46	1	4	.14
$A_{/b/}V_{/b/}$vs. $A_{/b/}$	4.74	1	4	.01**
$A_{/b/}V_{/b/}$vs. $A_{/b/}V_{/g/}$ Focused vs. divided	5.10	1	4	< .01**

** denotes $p < .001$

*** denotes $p < .0001$

Table 10 Mean RT (*SD*) in each condition

Experimental condition	Mean RT (*SD*) ms
$A_{/b/}V_{/b/}$	1081 (108)
$A_{/b/}$	990 (72)
$A_{/b/}V_{/g/}$	1194 (151)

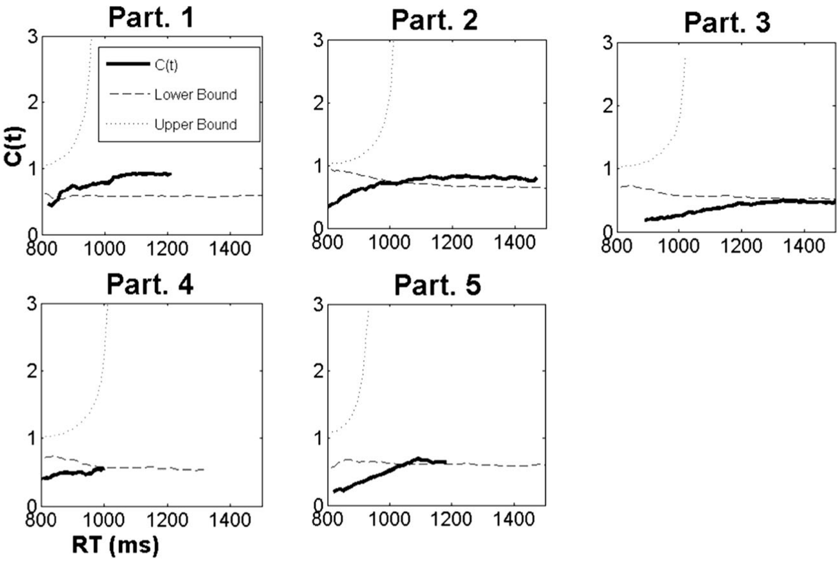

Fig. 2 Capacity, $C(t)$, shown for each of the five participants in the divided attention experiment. The solid and the dotted lines represent the upper bound on capacity and lower bound on unlimited capacity, respectively. Capacity values only exist for reaction times that are common to $A_{/b/}V_{/b/}$, $A_{/b/}$, and $V_{/b/}$

Although we can reject these particular models, our results could arise from a more sophisticated coactive or a parallel system. For example, systems that have inhibitory interactions between channels would be capable of yielding limited capacity (e.g., Eidels et al., 2011; Wenger & Townsend, 2006). However, a coactive system would have to include exceedingly strong inhibitory connections in order to produce the severely limited capacity exhibited by subjects 3, 4, and 5. Parsimony suggests, then, that these results are most consistent with a parallel system with limited capacity or a parallel system with inhibitory crosstalk.

Divided attention: Integrated hazard ratios

Next, we carried out comparisons involving the integrated hazard function ratios for each participant in order to supplement the capacity results and to deepen our inferences regarding which model may be more appropriate to describe auditory-visual integration. These analyses included the ratio of empirical integrated hazard functions for congruent A_BV_B (see Fig. 3) and incongruent A_BV_G (see Fig. 4) in the numerator. Using A_B and V_B is in the denominator allows us to assess the amount of audiovisual gain or interference provided over each modality.

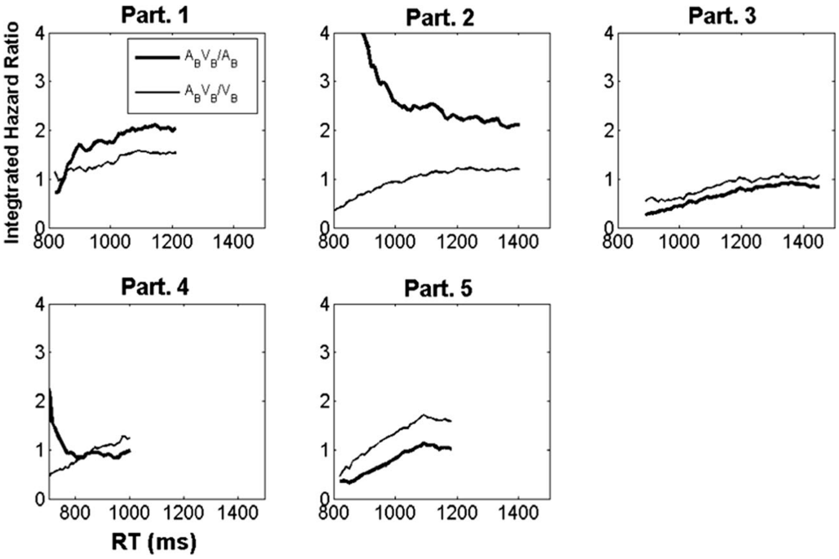

Fig. 3 Integrated hazard ratios comparing responses to congruent audiovisual signals, to auditory (filled circles) and visual-only (open circles) signals in the divided attention experiment

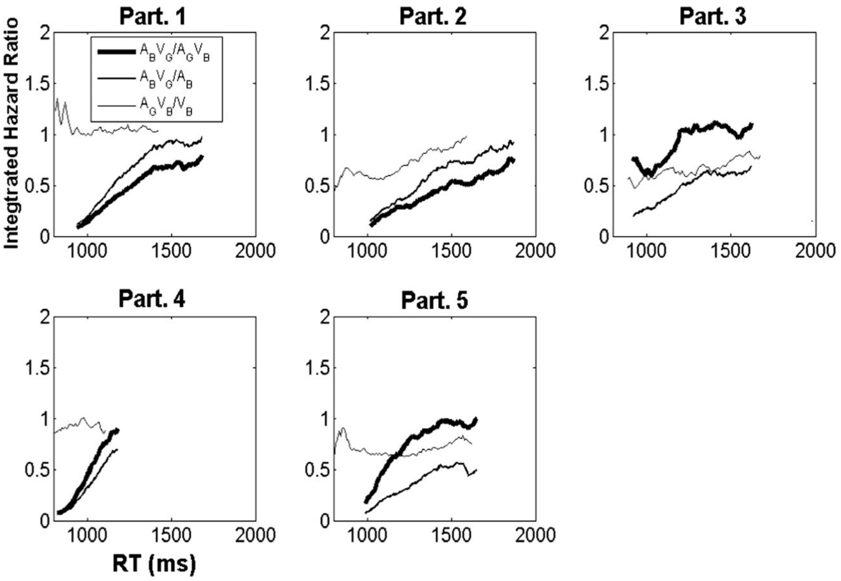

Fig. 4 Integrated hazard ratios comparing responses to incongruent audiovisual signals in the divided attention experiment: A_BV_G/A_GV_B (filled circles), A_BV_G/A_B (open circles), and A_GV_B/A_B (plus signs)

Cox proportional hazard regression statistical tests for the congruent ratios are shown in the first two columns of Table 11. Cox tests are semiparametric statistical tests that do not require the assumption of normality of RT distributions (see Altieri, Stevenson, Wallace, & Wenger, 2015; Wenger & Gibson, 2004; Wenger & Rhoten, in press). The Cox regression analysis was used to determine whether the underlying hazard functions from the two different trial types (e.g., A_BV_B vs. V_B) statistically differed from one other. We used Cox regression analysis, similar to Wenger and Gibson (2004) and Altieri et al. (2015) because the method has been established for testing differences in hazard function; furthermore, significant differences between two hazard functions implies that the integrated hazard functions will also differ.

Figure 3 and the first two columns of Table 11 illustrate the effects of the congruency benefit compared to each individual modality. Both demonstrate that for the A_BV_B/A_B comparison, Participants 1 and 2 showed ratios significantly greater than 1 indicating benefits provided by the congruent audiovisual stimulus ($A_{/b/}V_{/b/}$) over $A_{/b/}$alone. Notably, these two subjects also responded faster to $V_{/b/}$versus $A_{/b/}$stimuli, evidenced in the greater efficiency ratio for A_BV_B/A_B compared to A_BV_B/V_B. These two subjects also reveal a small redundancy gain over the visual modality, as both ratios appear to be greater than 1. These results are consistent with those for the capacity statistics. The other three participants evidenced similar RTs for $V_{/b/}$and $A_{/b/}$, although Participants 3 and 5 received greater benefit from $A_{/b/}V_{/b/}$over $V_{/b/}$.

Recall that both coactive and UCIP models predict integrated hazard ratios greater than 1, with coactive predictions being much higher than 1. Data from Participants 3 and 4 do not support either of these models, as both congruent integrated hazard ratios tended to be less than 1. However, Participants 1, 2, and 5 showed evidence of AV advantages over at least one of the two modalities. For these participants, AV RTs appear to be driven by the faster of the two modalities, with only Participant 1 and Participant 2, to a lesser degree, showing any advantage over the fastest single modality (the smaller of the two efficiency

Table 11 Cox regression analysis (Allison, 1995), % Change (β; *p*), results for the divided attention study. Positive β values indicate that RTs for the stimulus specified in the numerator were faster than the stimulus specified in the denominator, and negative values indicate the reverse

Part.	A_BV_B/A_B	A_BV_B/V_B	A_BV_G/A_GV_B	A_BV_G/A_B	A_GV_B/V_B
1	46(.22; <.0001) ***	49(-.29 <.0001)***	88(-.87; <.0001) ***	52(-.35; <.0001) ***	39(.05; .48)
2	80(.78; <.0001)*	37(.01; .86)	71(-.66; <.0001) ***	49(-.29; <.0001) ***	52(-.34; <.0001) ***
3	50(-.31; <.0001) ***	42(-.13; .05)*	38(.02; .76)	59(-.47; <.0001) ***	55(-.41; <.0001) ***
4	39(-.06; .32)	38(.02; .72)	89(-.88; <.0001) ***	140(-1.34; <.0001) ***	14(-.14; .04)*
5	44(-.18; .006)**	52(-.35; <.0001) ***	51(-.33; <.0001) ***	139(-1.33; <.0001) ***	51(-.32; <.0001) ***

* indicates significant p < .05

** denotes p < .001

*** denotes p < .0001

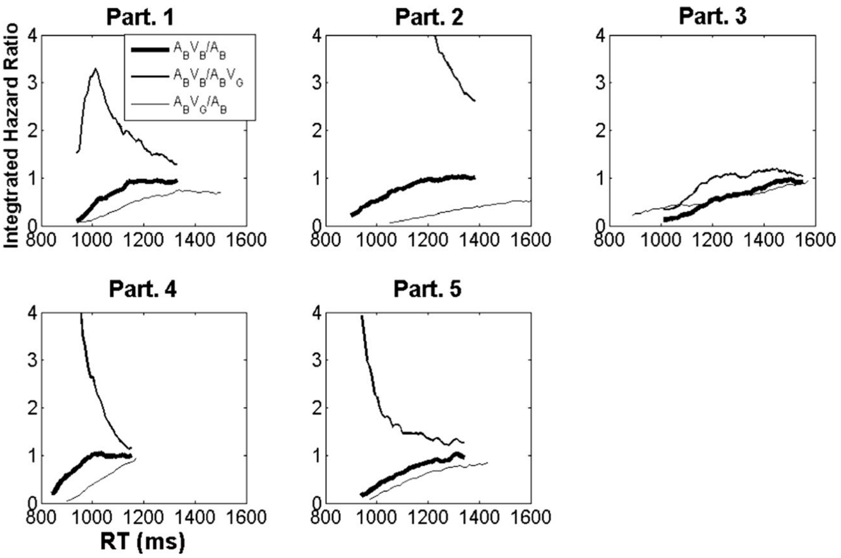

Fig. 5 Three integrated hazard ratios for the focused attention condition for each participant. The integrated hazard comparisons include: A_BV_B/A_B, A_BV_B/A_BV_G, and A_BV_G/A_B. The Cox regression analysis results are displayed in Tables 3 and 4

ratios is >1). Thus, there is further evidence to reject the coactive and UCIP models for four of the five participants, as 4 of 5 subjects received little to no benefit from the second modality.

Next, Fig. 4 shows three comparisons involving incongruent audiovisual speech information. These ratios allow assessment of how incongruent information in one modality interferes with the opposing modality. Cox regression tests are shown in the rightmost columns of Table 2, which support the results described below. First, the comparison involving A_BV_G/A_GV_B was carried out to determine whether incongruent visual information had a stronger effect on auditory processing, or alternatively, whether incongruent auditory information had a stronger effect on visual identification. A ratio of 1 would imply that there is symmetry in the influence of one modality on the other. The results show that in four out of five cases, visual distractors slowed auditory processing more than the other way around. That is, $A_BV_G/A_GV_B < 1$, and the McGurk perception of $A_{/b/}V_{/g/}$led to slower responses than the presumed clustered perception of $A_{/g/}V_{/b/}$. The exception was Participant 3, who showed no effect either way.

Second, we assessed whether an incongruent $V_{/g/}$slowed processing of $A_{/b/}$by evaluating the ratio A_BV_G/A_B. All participants showed evidence for slower processing for the $A_{/b/}V_{/g/}$trials when compared to $A_{/b/}$, consistent with the traditional accuracy-based result that $V_{/g/}$inhibits the perception of the $A_{/b/}$. Finally, we evaluated A_GV_B/V_B to determine whether and the extent to which the incongruent auditory $A_{/g/}$inhibited the detection of $V_{/b/}$. Testing this asymmetry constitutes an advantage of the detection approach since we can assess the relative influence of the visual modality on auditory perception, and the reverse. Strikingly, all participants except Participant 1 show evidence for slower processing to the $A_{/g/}V_{/b/}$stimulus when compared to $V_{/b/}$. Thus, there are effects in both directions: visual incongruence deleteriously affects auditory processing and auditory incongruence damages visual processing. However, the effect is greater for conflicting visual rather than conflicting auditory information.

One other notable finding in these data is that the incongruency effect is largest for short RTs. As subjects take longer to respond to the stimuli, the in-congruency effects tend to diminish, and efficiencies even approach 1 for the longest RTs. We also see similar effects in Fig. 2 where efficiencies for congruent stimuli increase with increasing RTs.

Table 12 Cox regression statistics [β(*p*)] from the focused attention condition

Part.	A_BV_B/A_B	A_BV_G/A_B	A_BV_B/A_BV_G
1	59(-.47; <.0001***)	84(-.83; <.0001***)	63(.53; <.0001***)
2	50(-.30; <.0001***)	128(-1.25; <.0001***)	117(1.16; <.0001***)
3	65(-.57; <.0001***)	61(-.51; <.0001***)	37(.01; .89)
4	52(-.35; <.0001***)	85(-.84; <.0001***)	76(.73; <.0001***)
5	61(-.50; <.0001***)	78(-.76; <.0001***)	52(.34; <.0001***)

** denotes p < .001

*** denotes p < .0001

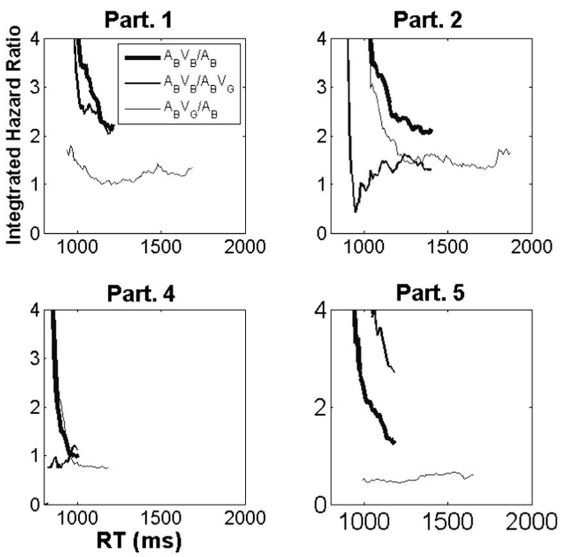

Fig. 6 Ratios obtained from the divided attention study, divided by the ratio obtained from the focused attention study: A_BV_B/A_B, A_BV_B/A_BV_G, and A_BV_G/A_B

Focused attention: Integrated hazard ratios

Figure 5 displays three integrated hazard ratios for the focused attention condition in the right panels separately for each participant. The integrated hazard comparisons include: A_BV_B/A_B, A_BV_B/A_BV_G, and $A_BV_G/A_{B.}$ The Cox regression analysis results are displayed in Table 12. Evidence for the ability of listeners to focus their attention on the auditory modality would be revealed in these hazard ratios all equaling 1, suggesting that the presence of the visual stimulus, whether it is $V_{/b/}$or $V_{/g/}$, would not influence the response. However, Fig. 5 demonstrates that for A_BV_B/A_B and A_BV_G/A, the efficiency ratios tend to be less than 1. That is, data from all observers revealed evidence for more efficient processing when only auditory information was present. It is particularly interesting that the irrelevant but congruent cue provided by $V_{/b/}$actually hurt performance. For A_BV_G/A_B, we see efficiencies much, much lower than 1 for all listeners, indicative of a strong inability to filter out the incongruent $V_{/g/}$stimulus.

The comparison between A_BV_B/A_B and A_BV_G/A_B allows a determination of the extent to which the conflicting but irrelevant visual information slowed processing relative to the congruent but irrelevant visual cues. All participants except Participant 3 evidenced significantly poorer efficiency when stimuli were incongruent than when they were congruent: The ratio A_BV_B/A_BV_G illustrates that subjects are faster for $A_{/b/}V_{/b/}$than for $A_{/b/}V_{/g/}$, indicative of faster RTs for congruent than for incongruent stimuli. Generally speaking, although the congruent information did not improve performance over that observed with a single modality, we still see a failure of attentional mechanisms to filter out the incongruent information. The incongruency hurt subjects much more than the congruent (but unhelpful) information.

Divided versus focused attention

To compare relative effects of attention, we compared integrated hazard ratios from the divided attention condition to the analogous ratio from the focused attention condition. The purpose of these comparisons was essentially to test an interaction across conditions in order to address the following question: To what extent was the difference between hazard ratios greater in the divided compared to the focused attention condition? Answering this question would allow us to determine whether the influence of the visual modality was greater when attention was divided. This test was carried out for the three integrated hazard ratios as shown in Fig. 6: DIV/FOC: A_BV_B/A_B, A_BV_G/A_B, A_BV_B/A_BV_G. To do this, we used a *Z* test based on Houpt and Townsend's (2012) capacity test statistics, modified to test for interactions: for example, the null hypothesis for the A_BV_B/A_B comparison was $[A_BV_B - A_B]_{Divided} - [A_BV_B - A_B]_{Focused} = 0$. Table 13 shows the results from the statistical tests.

Table 13 Statistical tests using estimated integrated hazard functions comparing divided versus focused attention integration (*Z*(*p*))

Participant	A_BV_B/A_B	A_BV_G/A_B	A_BV_B/A_BV_G
1	39 (<.0001***)	.01 (*p* = .89)	39 (<.0001***)
2	106(<.0001***)	142 (<.0001***)	-191 (<.0001***)
3	305 (<.0001***)	-102 (<.0001***)	266 (<.0001***)
4	219(<.0001***)	275 (<.0001***)	-197 (<.0001***)
5	165 (<.0001***)	-89 (<.0001***)	218 (<.0001***)

** denotes p < .001

*** denotes p < .0001

First, we observe that DIV/FOC: A_BV_B/A_B across conditions were significantly greater than 1 for all participants. Thus, all participants benefited more from congruent visual cues when attention was divided compared to when they focused only on the auditory modality. Note, however, that $A_{/b/}V_{/b/}$provides the participant with two opportunities to say "yes" in the divided attention experiment but not the focused attention experiment. Hence, there may be statistical effects present that are typically associated with multiple targets (Miller, 1982), causing us to overestimate the benefit of dividing attention.

The most straightforward ratio to interpret between the divided and focused attention conditions is that of $A_BV_G/A_{B,}$ as for both experiments, only $A_{/b/}$is associated with a "yes" response. Therefore, when taking the divided hazard ratios over the focused ratios, values of 1 would indicate that attention has little effect on performance. If subjects were faster in the focused attention condition, we would expect that values would be less than 1. Values greater than 1 imply that dividing attention provides faster responses for the incongruent stimulus and that focusing impairs performance. There are clearly some large individual differences, but these ratios are near 1 for only one of the subjects, Participant 1 (see Tables 7 and 8), suggesting that only this participant was immune to the effects of attentional manipulation. Interestingly, Participants 2 and 4 demonstrated divided attention ratios *greater* than the focused attention ratios, suggesting that there was a cost to focusing their attention on the speed of AV processing. However, Participants 3 and 5 show the opposite effect, indicating that responses were faster to $A_{/b/}V_{/g/}$trials when compared to $A_{/b/}$in focused attention. The implication is that these subjects possessed a weak ability to focus their attention and were being less impaired by the incongruent information in the focused attention case.

Finally, we compared the A_BV_B/A_BV_G ratio across attention conditions (see Table 13). This comparison allowed us to examine whether the benefit of congruent audiovisual signals is greater when attention is divided rather than focused. The results were variable, but consistent with the results for the A_BV_G/A_B ratio: Participants 3 and 5 showed a greater effect in the divided attention condition, suggestive of some (albeit weak) ability to focus their attention and filter out the congruent visual stimulus. On the other hand, Participants 2 and 4 showed evidence of a weaker congruency effect in the divided attention condition.

While the influence of the visual signal varied across participants and conditions, the overall pattern of results shows that all of the participants were unsuccessful in completely focusing their attention on the auditory modality. There is some suggestion that there is a small effect of attention for some subjects, but on the whole, there is a global failure to inhibit both congruent and incongruent information. That people are impaired by congruent information also lends credence to a parallel interactive channel model: Congruent information would not be expected to have a negative effect in a coactive model.

General discussion

This study used an RT capacity statistic to quantify audiovisual integration and interference under generally high-accuracy conditions within a signal discrimination or detection paradigm. The general result was that capacity for congruent audiovisual cues was generally limited instead of extreme supercapacity, as predicted by coactive models. Furthermore, while some listeners showed evidence of being able to benefit from audiovisual cues relative to auditory cues alone, the gains were always substantially less than those predicted by parallel unlimited capacity models. Some participants were impaired by the presence of the second modality, even when its signal was congruent. Thus, we see a failure to benefit from the second modality.

Our approach to assessing integration differentiates it from accuracy-only approaches and most other attempts to investigate the effects of attention on audiovisual integration (Massaro, 2004; Bernstein et al., 2004a). One potential shortcoming of previous in measuring integration relates to the inherent difficulty of defining what is meant by "integration." For instance, integration has been operationally defined as susceptibility to the McGurk effect (Grant, 2002), audiovisual enhancement in accuracy scores (Bergeson & Pisoni, 2004; Sommers et al., 2005; Sumby & Pollack, 1954), superadditivity in the neural signal (Calvert & Campbell, 2003; Calvert, Campbell, & Brammer, 2000), the combining of the intermodality information (as in FLMP; Massaro, 2004), or the difference between obtained accuracy scores compared to model predictions derived from confusion matrices (Grant, 2002; Grant et al. 1998). A promising course of action has been to formalize these definitions and use them to test models against one another (e.g., Grant et al., 1998; Massaro, 1987a, 1998). Most relevant to future research is the development of models that can be tested against both accuracy and RT data.

The difference in performance between the present RT experiment, which supports limited capacity, and the nonspeech detection RT experiments, which is consistent with super-capacity predictions, constitutes an important finding. In other studies, we have observed a lack of super capacity in high-accuracy scenarios but super capacity in lower accuracy scenarios (e.g., Altieri, Pisoni, & Townsend, 2011; Altieri, Townsend, & Wenger, 2014). However, .even some of those studies have failed to provide evidence for the level of supercapacity in RT performance predicted by coactivation (Townsend & Wenger, 2004a; Altieri & Townsend, 2011; Altieri & Wenger, 2013). Other evidence suggesting that the relevant factor is not accuracy per se is support for coactivation in AV experiments using RT with nonspeech stimuli (Diederich & Colonius, 1991; Miller, 1982). Thus, it appears to not be high accuracy alone that yields this limited capacity result.

Could this result be due to the use of speech versus nonspeech stimuli? Such explanations call upon the complex nature of speech relative to the simple nature of the stimuli often used in the nonspeech tasks. Eramudugolla, Henderson, and Mattingley (2011) observed a congruency benefit that was similar in magnitude for both speech and nonspeech stimuli in an accuracy-based experiment. Ultimately, these questions require follow-up because our strong limited capacity result across subjects is at odds with much of the literature, including the FLMP and PRE models (although the predictions of those models are for accuracy).

In any case, the limited capacity result from the divided attention condition does allow us to reject a coactive model. The integrated hazard ratios from both divided and focused attention provide supporting evidence for this claim. Even though the hazard ratios for congruent cases are sometimes greater than 1 (which is predicted by the coactive model), they generally are only slightly larger than one. We would generally expect a coactive system to yield ratios >> 1 for A_BV_B/A_B and A_BV_B/A_BV_G. The asymmetry in the amount of influence the visual information has on auditory processing and the reverse (as also reported by Eramudugolla et al., 2011, for accuracy) would also imply that, should a coactive model hold, the individual channels would need to be weighted differentially or that there must be some form of asymmetric inhibitory channel interactions.

At this point, we cannot reject some forms of common format models. In fact, the attentional data that indicate a failure of attentional mechanisms to filter out congruent and incongruent visual stimuli provide support for a model in which attention cannot separate the individual channels from the combined percept. However, although one possible explanation for these data is an early coactive model with weighted channels and inhibitory connections between them, the inhibitory connections between channels would have to be exceptionally strong to yield severely limited capacity and efficiency ratios that are so small. Whether the common format model holds, we argue that models of audiovisual consonant identification, such as FLMP, PRE, or other related processing models, ought to incorporate these findings by weighting visual contributions to auditory recognition stronger than auditory contributions to visual recognition.

An alternative model is based on parallel processing, but is more complex than the standard parallel model. Recall that standard parallel processing assumes unlimited capacity and independent parallel channels. Because the standard parallel model, by definition, is unlimited capacity, the limited capacity results obtained for the congruent stimuli contradict this model. Based on the capacity results alone, however, we cannot distinguish between a model with parallel independent channels with limited capacity and a model with parallel channels with inhibitory cross-talk, as either of these factors could produce limited capacity.

The failure of attention to focus on the auditory channel in a parallel model could occur due to interactions between the channels. Furthermore, the dynamics revealed are consistent with time-varying interactions between channels: A limited attention resource would not yield such strong dynamics if the resource delivered capacity at a constant rate. Such an interpretation is consistent with the findings of Altieri (2014), who argued that a parallel interactive model can account for the McGurk effect and congruent audiovisual interactions. Additionally, the FLMP, which has historically provided the strongest account of the research on the McGurk effect, provides an account consistent with our findings supporting a parallel interactive model rather than coactivation in speech categorization experiments (see Massaro, 1987b).

Overall, we find convincing evidence for early interactions between auditory and visual channels. We suggest that a parallel interactive "preattentive" model would be the most parsimonious account of the results from our listeners, because interchannel inhibitory connections would have to be exceptionally strong to yield the observed severely limited capacity in coactive systems (Eidels et al., 2011). While the preattentive inhibitory model appears appropriate for most typical-hearing adults, other models may better describe other listeners. For example, individuals with autism appear to be less susceptible to the influences of visual speech and hence the McGurk effect (e.g., Stevenson et al., 2014); hypothetically, they may better conform to the predictions of the parallel postattentive model. On the other hand, those who are hard of hearing may rely more heavily on the visual signal compared to participants in this study and thus show stronger evidence of inhibition from the visual channel. Taken together, these findings are consistent with the notion that information presented to listeners must be differentiated from information processed by listeners (e.g., Massaro, 1987a, 1998).

Further work will be needed to determine the influence of interaction between channels within the context of the different potential architectures. One promising way to address this question would be to use the techniques described in Townsend and Nozawa (1995) that were specifically designed to assess the architecture underlying integration of two channels. These tools may allow us to take steps in establishing the nature of cross-channel interactions (e.g., Altieri & Wenger, 2013).

In summary, the present use of RT has allowed us to not only define integration but also measure it using UCIP models as a benchmark. This approach also expanded upon previous literature (e.g., Altieri & Townsend, 2011) by quantifying integration as a function of attention and congruency.

Acknowledgments This work was supported by the National Institute of Health (Grant No. DC-00111), the National Institute of Health Speech Training Grant (No. DC-00012), and also by NIMH 057717-07 and

AFOSR FA9550-07-1-0078 grants to James T. Townsend. We wish to acknowledge David B. Pisoni, Jerome Busemeyer, and John Kruschke for past discussion and comments.

References

Algom, D., Eidels, A., Hawkins, R. X. D., Jefferson, B., & Townsend, J. T. (2015). Features of response times: Identification of cognitive mechanisms through mathematical modeling. In J. Busemeyer, J. Wang, J. T. Townsend, & A. Eidels (Eds.), *Handbook of computational and mathematical psychology*. Oxford University Press; Oxford, UK.

Allison, P. (1995). *Survival analysis using the SAS system: A practical guide*. Cary, NC: SAS Institute.

Altieri, N. (2014). *A dynamic theory of audiovisual speech integration* (2nd ed.). Scholars Press. OmniScriptum GmbH & Co. Germany.

Altieri, N., & Hudock, D. (2014). Variability in audiovisual speech integration skills assessed by combined capacity and accuracy measures. *International Journal of Audiology, 53*, 710–718.

Altieri, N., Pisoni, D. B., & Townsend, J. T. (2011). Behavioral, clinical, and neurobiological constraints on theories of audiovisual speech integration: A review and suggestions for new directions. *Seeing and Perceiving, 24,* 513–539.

Altieri, N., Stevenson, R. A., Wallace, M. T., & Wenger, M. J. (2015). Learning to associate auditory and visual stimuli: Capacity and neural measures of efficiency. *Brain Topography, 28,* 479–493.

Altieri, N., & Townsend, J. T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Frontiers in Psychology, 2*(238), 1–15.

Altieri, N., Townsend, J.T., & Wenger, M.J. (2014). A measure for assessing the effects of audiovisual speech integration. *Behavior Research Methods, 46*, 406–415. doi:10.3758/s13428-013-0372-8.

Altieri, N., & Wenger, M. (2013). Neural dynamics of audiovisual integration efficiency under variable listening conditions: An individual participant analysis. *Frontiers in Psychology, 4*(615), 1–15.

Bergeson, T. R., & Pisoni, D. B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 153–176). Cambridge, MA: MIT Press.

Bernstein, L. E., Auer, E. T., Jr., & Moore, J. K. (2004a). Audiovisual speech binding: Convergence or association? In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processing* (pp. 203–223). Cambridge: MIT Press.

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004b). Auditory speech detection in noise enhanced by lip-reading. *Speech Communication, 44,* 5–18.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology, 43A*(3), 647–677.

Buchan, J. N., & Munhall, K. G. (2011). The influence of selective attention to auditory and visual speech on the integration of audiovisual speech information. *Perception, 40*(10), 1164–1182.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15*(1), 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.

Diederich, A., & Colonius, H. (1991). A further test of the superposition model for redundant signals effect in bimodal detection. *Perception & Psychophysics, 50,* 83–86.

Eidels, A., Houpt, J. W., Altieri, N., Pei, L., & Townsend, J. T. (2011). Nice guys finish fast and bad guys finish last: Facilitatory vs. inhibitory interaction in parallel systems. *Journal of Mathematical Psychology, 55*(2), 176–190.

Eramudugolla, R., Henderson, R., & Mattingley, J. B. (2011). Effects of audio-visual integration on the detection of masked speech and non-speech sounds. *Brain and Cognition, 75,* 60–66.

Erber, N. P. (2003). Use of hearing aids by older people: Influence of non-auditory factors (vision, manual dexterity). *International Journal of Audiology, 42,* 2S21–2S26.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2), 134–140.

Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective. *Journal of the Acoustical Society of America, 112*(1), 30–33.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*(5), 2677–2690.

Green, K. P., & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 278–288.

Hessler, D., Jonkers, R., Stowe, L., & Bastiaanse, R. (2013). The whole is more than the sum of its parts—Audiovisual processing of phonemes investigated with ERPs. *Brain and Language, 124,* 213–224.

Houpt, J.W., & Townsend, J.T. (2012). Statistical Measures for Workload Capacity Analysis. *Journal of Mathematical Psychology, 56*, 341–355.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Massaro, D. W. (1987a). 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: Erlbaum.

Massaro, D. W. (1987b). Integrating multiple sources of information in listening and reading. In D. A. Allport, D. G. MacKay, W. Prinz, & E. Scheerer (Eds.), *Language perception and production: Shared mechanisms in listening, speaking, reading and writing* (pp. 111–129). New York, NY: Academic Press.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

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: MIT Press.

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

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14*(2), 247–279.

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of American,* 1061–1077.

Scharff, A., Palmer, J., & Moore, C. M. (2011). Evidence of fixed capacity in visual object categorization. *Psychological Bulletin and Review, 18,* 713–721.

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: A heightened McGurk effect in older adults. *Frontiers in Psychology, 5,* 323.

Sherffert, S., Lachs, L., & Hernandez, L. R. (1997). The Hoosier audiovisual multi-talker database. *Research on Spoken Language Processing Progress* (Report No. 21). Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Shiffrin, R., & Gardner, G. T. (1972). Visual processing capacity and attentional control. *Journal of Experimental Psychology, 93,* 72–78.

Sommers, M., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26,* 263–275.

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: Evidence from the speeded classification task. *Cognition, 92,* B13–B23.

Soto-Faraco ,S., Navarra, J., & Alsius, A. (2005). Assessing automaticity in audiovisual speech integration: Evidence from the speeded classification task. *Cognition, 92*, B13–23. PMID 15019556 doi:10.1016/j.cognition.2003.10.005

Stevenson, R. A., Siemann, J. K., Woynaroski, T. G., Schneider, B. C., Camarata, S. M., & Wallace, M. T. (2014). Arrested development of audiovisual speech perception in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 4*(6), 1470–1477.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 12–15.

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*(3), 457–472.

Townsend, J. T. (1981). Some characteristics of visual whole report behavior. *Acta Psychologica, 47,* 149–173.

Townsend, J. T., & Altieri, N. (2012). An accuracy-response time capacity assessment function that measures performance against standard parallel predictions. *Psychological Review, 199*(3), 500–516.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan and F. Restle (Eds.), Cognitive Theory Vol. III (pp. 200–239). Hillsdale, NJ: Erlbaum Associates.

Townsend, J. T., & Eidels, A. (2011). Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin & Review, 18,* 659–681.

Townsend, J. T., Houpt, J. W., & Silbert, N. H. (2012). General recognition theory extended to include response times: Predictions for a class of parallel systems. *Journal of Mathematical Psychology, 56,* 476–494.

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*(4), 321–359.

Townsend, J. T., & Wenger, M. J. (2004a). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111*(4), 1003–1035.

Townsend, J. T., & Wenger, M. J. (2004b). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review, 11*, 391–418.

van Wassenhove, V. (2013). Speech through ears and eyes: Interfacing the senses with the supramodal brain. *Frontiers in Psychology, 4*, 388. doi:10.3389/fpsyg.2013.00388

van Wassenhove, V., Grant, K., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 1181–1186.

Wenger, M. J., & Gibson, B. S. (2004). Using hazard functions to assess changes in processing capacity in an attentional cuing paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 708–719.

Wenger, M. J., & Rhoten, S. E. (in press). Characterizing and quantifying human bandwidth: On the utility and criticality of the construct of capacity. In J. W. Houpt & L. M. Blaha (Eds.), *Mathematical models of perception and cognition* (Vols. 1–2).

Wenger, M. J., & Townsend, J. T. (2006). On the costs and benefits of faces and words: Process characteristics of feature search in highly meaningful stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 45*, 283–323.