Linguistic Mediation of Visual Search: Effects of Speech Timing and Display

Eric M. Chiu (echiu@ucmerced.edu)
Michael J. Spivey (spivey@ucmerced.edu)

Cognitive and Information Sciences
5200 North Lake Road, Merced, CA 95343 USA

Abstract

Recent studies have shown that instead of a dichotomy between parallel and serial search strategies, in many instances we see a combination of both search strategies utilized. Consequently, computational models and theoretical accounts of visual search processing have evolved from traditional parallel/serial descriptions to labels of “efficient” and “inefficient.” In the first experiment, we replicate previous findings with a between-subjects design, where incremental spoken language comprehension influences the efficiency of visual search processing. Next, we utilize a localist attractor network to simulate the results from the first experiment, and then employ the network to make quantitative predictions about the influence of subtle timing differences of real-time language processing on visual search. These predictions are then tested and confirmed in our second experiment. The results provide further evidence toward understanding linguistically mediated influences on real-time visual search processing and support an interactive processing account of visual search and language comprehension.

Keywords: visual search, linguistic mediation, speech, efficient, inefficient

Introduction

Many everyday objects and events generate multi-sensory inputs that appear concurrently or with some amount of overlap. Taking advantage of both the shared and unique information content in these signals, instead of processing them with isolated modules, can be advantageous for task performance and for learning (e.g., de Sa & Ballard, 1998). A similar situation arises in basic signal processing. For example, air traffic control is one of the many circumstances in which it is imperative that a sensor, such as radar, detects the presence of an external event. The current technology involved in this sort of event detection generally depends on a single sensor. This methodology can be very effective when the event to be detected produces a strong signal with a unique signature and where there are few other competing signals that may activate that sensor and confound operators. Nevertheless, this technology is very limited, because the threshold for activating the sensor must be very low if one is to be certain that an event is not missed due to a weak or ambiguous signal or a noisy environment. Unfortunately, this situation generates far too many signals, effectively making the information from such a sensor extremely difficult to manage. This predicament can be remedied by functionally coupling two or more sensors, each tuned to a different form of environmental energy (e.g. visible light and infrared light, radar and sonar, or light and sound). By specifying the criteria for activation and temporally synchronizing each sensor before activating their common central processor, the thresholds of these sensors can be set very low while still minimizing false-positives and accurately disambiguating events.

In human cognition, there are quite a few examples of perceptual interactions across vision, audition, touch, and language systems. The most famous illustration of dynamic and immediate integration of linguistic and visual processing is perhaps the McGurk effect. The McGurk effect is experienced when you see a televised face repeatedly saying “ga-ga,” but synchronized with the mouth movements the audio stream actually delivers “ba-ba,” which constructs a convincing percept of hearing “da-da” (McGurk & MacDonald, 1976). Historically the visual system was thought of as a functionally independent cognitive process (Fodor, 1983), but recent research demonstrates, instead, a dynamic and immediate integration of linguistic information with visual information. For example, ascribing a meaningful label to a novel visual stimulus, in a visual search task, significantly improves search efficiency and reaction times when compared to observers who do not have meaningful labels for the same stimuli (Lupyan & Spivey, 2008). These results demonstrate a top-down conceptual influence on visual recognition, implying that visual perception depends not only on what something looks like but also on what it means.

The present study extends finding like those above to explore the degree to which the incremental processing of spoken words in a full sentence can interact with concurrent visual search processes. Traditionally, two contrasting perspectives have driven the field of visual search on attention. The serial-processing perspective claims that observers allocate complete attentional resources discretely to individual objects, one at a time (Treisman & Gelade, 1980; Treisman, 1988). Conversely, neural mechanisms have been found to be mediated by biased competition in the extrastriate visual cortex, forming a compelling argument against the serial-processing perspective and for the parallel-processing perspective, which claims attention is better characterized as a function of partially active representations of objects simultaneously contending for probabilistic
mappings onto motor output (Desimone & Duncan, 1995; Desimone, 1998). It has been shown that single-feature visual search is relatively unaffected by the number of distractors, and often induces a perceptual “pop-out” effect. Conversely, studies of two-feature conjunction searches usually find a linear increase in reaction time as the number of distractors increase. However, as we will demonstrate these apparent dichotomous perspectives may not be from two contrasting fundamentals but rather a product of a single process. This single process is supported by observations of improvement in visual search tasks better described by a graded enhancement of feature salience (Olds, Cowan, & Jolicoeur, 2000).

Olds et al. (2000) presented single-feature visual search pop-out displays for very brief periods of time, less than 100ms in some conditions, before changing them to conjunction-search displays. Despite the fact that no observers reported experiencing the target “pop-out” for them, their response times nevertheless exhibited some facilitatory effects due to the very brief period of time during which the display had only single-feature distractors. Although their responses were not as fast as with pure pop-out displays, Olds and colleagues illustrated a graded effect of improved search efficiency and coined it “search assistance.” Findings like these, along with signal detection theory analyses of visual search data (Eckstein, 1998), and the absence of a bimodal distribution of search efficiencies (Wolfe, 1998), have led to the serial-parallel dichotomy in visual search being replaced by the notion of a continuum of search efficiency (e.g., Nakayama & Joseph, 1998).

A different kind of “search assistance” phenomenon comes from work by Spivey, Tyler, Eberhard, and Tanenhaus (2001). Observers in Audio/Visual Concurrent conditions, where the conjunction search display is presented concurrently with target identity via auditory linguistic queries (e.g. “Is there a red vertical?”), showed dramatically improved search efficiency compared to an Auditory-First Control condition, where the same spoken query of target identity was provided prior to visual display onset. The findings suggest that upon hearing the first-mentioned adjective in the spoken sentence, visual attention is able to begin the search with only that single feature, thus initiating the process with a highly efficient single-feature search. Upon hearing the second adjective, several hundred milliseconds later, the target can be quickly found among the attended subset of objects.

Subsequently, Gibson, Eberhard, and Bryant (2005) demonstrated that the improved visual search efficiency is affected by speech rate. With faster speech, the A/V Concurrent condition no longer provides an enhanced efficiency in conjunction search.

The purpose of the present study was to, first, replicate and extend previous findings and to, secondly, further examine the effects of precise timing of incremental delivery of natural linguistic queries during visual search processing.

Experiment #1

In this experiment, we replicated the design of Spivey et al.’s (2001) and Reali et al.’s Exp. 1, with the exception that we utilized a blocked between subjects design. One group was randomly assigned to the auditory-first control condition and one group of subjects was randomly assigned to the audiovisual-concurrent (A/V-concurrent) condition.

Methods

Participants One hundred and sixty-seven University of California, Merced undergraduates received course credit for participating in this experiment. Participants were randomly assigned to one of two slightly different conditions, 90 in the auditory-first control condition and 77 in the A/V concurrent condition. Fourteen participants in the auditory-first and 17 in the A/V-concurrent condition were unable to perform the task above an 80% accuracy and so their data sets were omitted from the analysis. Additionally, all incorrect responses and trials with reaction times 2.5 interquartile ranges from the median were also omitted from the analysis (IQR was used because of the superior resistance to outliers). The participants in this and all subsequent experiments were naive as to the purpose of the experiments, and all reported normal hearing and normal or corrected-to-normal vision.

Stimuli and Procedure The experiment was composed of two slightly different types of trials, auditory-first control trials and A/V-concurrent trials. Participants where randomly assigned to one of the conditions and participated in a 32 trial “practice block” that was not part of the final analysis before participating in a 96 trial “experiment block” that was used in the final analysis.

Participants in the auditory-first control condition were presented with the target identity via spoken query (e.g. “Is there a red vertical?”) prior to visual display onset. Participants in the A/V-concurrent condition were presented with the visual search display concurrently with target identity via the aforementioned auditory linguistic instruction. The same female speaker recorded all speech files, with identical 1-s preamble recording, “Is there a...” spliced onto the beginning of each of the four target queries. The two descriptive adjectives (color: “red” or “green” and orientation: “vertical” or “horizontal”) averaged 1.5s. Each stimulus bar subtended 2.8˚ X 0.4˚ of visual angle and neighboring bars were separated from one another by an average of 2˚ of visual angle.

Participants were instructed to respond to each display as quickly and accurately as possible by pressing the labeled “yes” button if the queried object was present in the display and the labeled “no” button if it was absent. Participants initiated each trial manually by pressing the space bar. A fixation-cross preceded the onset of the visual display in order to direct participants’ gaze to the central region of the display. The trials in each condition were broken up into two
blocks. Half of the trials were with target present and half with target absent; set sizes of 5, 10, 15, and 20 were used.

Results and Discussion

A hierarchal linear model (HLM), which accounts for the unbalanced repeated measures assumption, was used for this analysis and all following analyses. In order to fulfill the assumption of distribution normality, as reaction time response data is naturally positively skewed, the results reported here and for Exp. 2 will report descriptive statistics (slopes and intercepts) from an untransformed HLM while inferential statistics will be reported from a log transformed HLM.

The overall model for slopes of the reaction-time-by-set-size functions, including data from Exp. 2, was significant for target-present trials for the untransformed model, $\chi^2(4, N = 382) = 13.31, p = .01$, and for the log transformed model, $\chi^2(4, N = 382) = 26.19, p < .00$, meaning the slopes as a whole significantly differ from the auditory-first control condition for target-present trials. Similarly the overall model for slope, including data from Exp. 2, was highly significant for target-absent trials for the untransformed model, $\chi^2(4, N = 382) = 18.55, p = .001$, and for the log transformed model, $\chi^2(4, N = 382) = 104.96, p < .00$, meaning the slopes as a whole significantly differ from the auditory-first control condition for target-absent trials. The overall significance of both transformed and untransformed models for both target-present and target-absent trials supports our method of analysis and report.

In this experiment we reproduced previous findings demonstrated by Spivey et al. (2001) and Reali et al. (1999) with a between subjects design. In the auditory-first control condition, the reaction-time-by-set-size function was highly linear in both target-present, $r^2 = 0.971$, and target-absent trials, $r^2 = 0.996$, as typically observed in standard conjunctions search tasks. Similarly, the reaction-time-by-set-size function for the A/V concurrent condition was highly linear for target-present trials, $r^2 = 0.915$, and target-absent trials, $r^2 = 0.947$. Overall mean reaction time and subsequent y-intercepts were significantly slower in A/V concurrent conditions because complete notification of target identity was delayed by approximately 1.5s relative to the auditory-first control condition for both target-present, $t(59) = 3.28, p = .001$, and target-absent, $t(59) = 3.03, p = .003$, trials. Mean accuracy was 93.0% for the auditory-first control condition and 94.5% for the A/V concurrent condition, similar to previous observations. A HLM analysis revealed significantly shallower slopes for A/V concurrent condition than auditory-first condition in target-present trials, $t(75) = 5.5, p < .00$, and target-absent trials, $t(59) = 9.9, p < .00$, as observed by Spivey et al. (2001) and Reali et al. (2006). The HLM analysis of the untransformed data set also revealed significantly shallower slopes for A/V-concurrent condition than auditory-first condition in target-present trials, $t(75) = 2.02, p = .04$, and target-absent trials, $t(59) = 4.64, p < .00$.

Insofar, the results continue to indicate that by simply adjusting the timing of spoken query so that the two target-feature words were presented at the same time the visual display was visible allowed participants to find the target object in a way that was substantially less affected by the number of distractors, as congruent with Spivey et al. (2001) and Reali et al. (2006).

Model

To further investigate the influence of incremental information processing on visual search Spivey & Dale (2004) and later Reali et al. (2006) implemented a simple localist attractor network model that easily simulated a potential mechanism by which search process may be influenced by incremental linguistic input. Traditional implementations of Desimone and Duncan’s (1995) biased competition framework have focused on the level of individual firing rates of neurons (Reynolds & Desimone, 2001; Spratling & Johnson, 2004). By abstracting this framework to a level of unitized population codes that represent objects competing against one another Spivey and Dale (2004) evolved simulations of visual search reaction times. In the present implementation of this model, one feature vector of nodes represented the target property redness (positive activation) and nonredness (zero activation). Another feature vector represented the target property verticalness (positive activation) and nonverticalness (zero activation). Finally, an integration vector represented each objects’ likelihood of being the target. The lengths of these vectors vary depending on set size between 0 and 25 by intervals of 5; for a set size of 15 the length of both feature vectors and the integration vector would be 15.

At the beginning of the simulation, initial activation of each node in either feature vector is 1/N, where N is the number of nodes in the vector. Hearing “red” and “vertical” provides input to these feature vectors by multiplying each node by 1 if the object exhibits the appropriate property and by 0 if the object does not exhibit the appropriate property. Similar to a probability distribution, during the networks settling process each timestep begins with the normalizing of each feature vector to 1. The vectors are then noncumulatively averaged at the integration layer. The integration layer then sends point-wise multiplicative cumulative feedback to each of the feature vectors, applying a little cross talk from the other feature vector and biasing them. For each timestep (treated as 20ms, which represents the time difference between spoken adjectives), this cycle of normalization to integration to feedback repeats until a node in the integration layer exceeds some criterion activation, 0.95 in this case, at which point the target has been found and a settling time (i.e., RT) is recorded.

This normalized recurrence competition algorithm allows the integration layer to be updated and evaluated in parallel
at each timestep, rather than imposing a serial search of one object at a time, which reflects the human data and produces a strikingly linear increase in settling time as set size increases. It should be noted that this competition algorithm does not simulate target-absent trials, since termination of search is not likely the result of a representation winning a competition process (Chun & Wolfe, 1996). We have yet to come across an algorithm that accurately simulated target-absent search processing.

Simulation When simulating the auditory-first control condition the redness and verticalness vectors received input at the same time. The result was an RT X set size slope of 16 ms/item. In simulating the A/V-concurrent condition, the verticalness feature received its input 35 timesteps after the redness feature vector received its input allowing the network to pursue its settling first (the equivalent of 700ms). Under these circumstances, the RT X set size slope was reduced to 9.1ms/item. A constant of 800ms for auditory-first and A/V concurrent conditions is then added to the RT for perceptual registration and motor execution.

The activation pattern in the redness vector gradually increases probabilistic activation for the verticalness vector toward the red objects biasing activation away from the nonred objects. The model essentially simulates the graded subset selection of the red objects, which begins when the participant first hears “red” while the display is visible. So when the input to the verticalness vector is activated when participants hears “vertical,” it is multiplied by a number of nonred vertical nodes whose activations are so close to zero that the products are unable to effectively compete effectively creating a subset of salient red objects.

An investigation reveals that the model is highly correlated with the data collected from Exp. 1 as evident with a RMSE = 137.14 and a highly significant Pearson r-squared value, \( r^2 = 0.997, p < .001 \). Root mean square error (RMSE) can be interpreted as the standard deviation of the unexplained variance or error between the model and the human data, and thus has the valuable property of being in the same metric as the response variable (i.e., milliseconds). In this example a RMSE value of 137.14 ms is a relatively nominal difference and reflects a good model fit.

Predictions To test the model’s strength and to further investigate the mechanisms of linguistically mediated visual search we use the same localist attractor network model with a minor adjustment to make some predictions on a semi-concurrent condition, where the search display appears immediately after the first target-feature is mentioned but before the second target-feature is presented.

Similar to the A/V-concurrent simulations when simulating the semi-concurrent conditions, the redness feature vector received its input slightly before the verticalness feature received its input. This allows the network to begin settling the redness vector slightly before the verticalness vector is activated. Initial semi-concurrent simulations activated the verticalness vector 5 timesteps after the redness vector. Under these circumstances, the RT X set size slope did not differ from the auditory-first condition and remained at 16 ms/item. Additional investigation revealed that a semi-concurrent simulation where the verticalness vector received input 10 timesteps after the redness vector also had no affect on RT X set size slope but when 15 timesteps were allowed we saw that the slope decreased to 14.3 ms/item and furthermore when 20 timesteps were employed we observed a continual decrease to 13.1 ms/item.

Interestingly our localist attractor model, which easily simulated data from Exp. 1 predicted a surprising non-linear shallowing of RT X set size slopes as the timesteps increased from 5 to 20 between the activation of the first feature vector and the second. The RT X set size slope did not change until a 15-timestep duration was introduced but continues to shallow non-linearly when 20 timesteps were employed.

Experiment #2

This experiment explores the predictions made by the localist attractor network model on what we call the semi-concurrent condition. We utilize four different stimulus onset asynchronies (SOA’s) to imitate the four different timestep durations used in our localist attractor predictions.

Methods

The methods followed that of Exp. 1 except that the timing of the visual display was slightly altered along with the auditory files, which were modified to include four SOA’s. Observers were randomly placed in one of four conditions, where they are presented with one adjective describing the target identity before onset of the search display and the other concurrently with onset of the search display.

The same audio files were utilized from Exp. 1 but with four SOA’s introduced between the end of the first descriptive adjectives, color, and the second adjectives, orientation (e.g. “Is there a red -SOA- vertical”). In these semi-concurrent conditions the search display was presented immediately preceding the first target descriptor mentioned after which subsequent SOA’s began before the second target descriptor was mentioned. We used SOA’s of 0ms, 200ms, 400ms, and 600ms, which were selected as estimates from the model with slight modifications to account for linguistic processing.

Participants A new sample of 275 University of California, Merced undergraduates participated in this experiment for course credit. Participants were randomly assigned to one of four SOA semi-concurrent conditions and only participated in that one condition. Forty-two participants were assigned to the 0ms SOA condition, 107 to the 200ms SOA, 66 to the 400ms SOA, and 60 to the 600ms SOA. Participants that did not meet a minimum 80% accuracy criterion were omitted from the analysis. Five participants in the 0ms SOA
condition did not meet this requirement, 11 in the 200ms SOA, 7 in the 400ms SOA, and 9 in the 600ms SOA.

**Results and Discussion**

The reaction-time-by-set-size function was highly linear in both target-present, $r^2 = 0.997$, and target-absent trials, $r^2 = 0.990$ for the 0ms SOA, also for both target-present trials, $r^2 = 0.966$, and target-absent trials, $r^2 = 0.983$ for the 200ms SOA, and again for both target-present trials, $r^2 = 0.969$, and target-absent trials, $r^2 = 0.999$ for the 400ms, and finally for both target-present trials, $r^2 = 0.971$, and target-absent trials, $r^2 = 0.972$, for the 600ms, as typically observed in standard conjunctions search tasks.

Overall mean reaction time and subsequent y-intercepts were significantly slower as SOA’s increased because complete notification of target identity was delayed by the approximate duration of the SOA for target-present trials, $\chi^2(4, N = 382) = 213.00$, $p < .00$, and for target-absent trials, $\chi^2(4, N = 382) = 339.00$, $p < .00$. Mean accuracy was 94.6% for the 0ms SOA condition, 93.5% for the 200ms SOA, 95.4% for the 400ms SOA, and 94.0% for the 600ms SOA, which are similar to previous observations of accuracy on this task (Spivey et al., 2001; Reali et al., 2006).

An analysis revealed significantly shallower slopes than compared to the auditory-first control condition for the 400ms SOA semi-concurrent condition for target-present trials, $t(58) = 4.48$, $p < .00$, and target-absent trials, $t(58) = 8.81$, $p < .00$, as well as for the 600ms SOA for target-present trials, $t(50) = 3.88$, $p < .00$, and target-absent trials, $t(50) = 8.32$, $p < .00$. Reaction time by set size slopes where not significantly shallower for the 0ms SOA for target-present trials, $t(36) = 1.78$, $p = 0.08$, and target-absent trials, $t(98) = 6.54$, $p < .00$. Interestingly reaction time by set size slopes for the 200ms SOA were not significant for target-present trials, $t(36) = 1.96$, $p = .05$, but were significant for target-absent trials, $t(98) = 3.87$, $p < .00$.

The findings are congruent with the localist attractor network predictions. As predicted the slopes of the reaction-time-by-set-size functions for target-present trials revealed that as SOA’s increase linearly by condition significantly shallower slopes appeared non-linearly beginning with the 400ms SOA and continuing with the 600ms SOA when compared with the auditory-first control conditions from Exp. 1 but not for the 0ms and 200ms SOA. Interestingly, the slopes of the reaction-time-by-set-size functions for target-absent trials revealed significantly shallower slopes for all but the 0ms SOA when compared with the auditory-first control condition.

An investigation reveals that the model is a good fit to the data collected from Exp. 2 with a RMSE = 157.02 and a highly significant Pearson r-squared value, $r^2 = 0.933$, $p < .00$. In this comparison a RMSE value of 157.02 ms is a relatively nominal difference and reflects a good model fit. When comparing the model’s fit with data from Exp. 2 with the model’s fit with data from Exp. 1 we see a slightly smaller RMSE for Exp. 1, 137.14, compared to Exp. 2, 157.02, indicating a slightly better fit of the model to Exp. 1. Conversely, a closer inspection reveals that the Pearson R-squared value is slightly more significant for Exp. 2, $p < 0.00$, than for Exp. 1, $p < 0.001$, imply the difference in RMSE is not substantial and the model is a well fit for both sets of data.

**General Discussion**

In an assortment of experimental conditions (Exp. 1-2), the results appear to indicate that because of the incremental nature of spoken language comprehension, observers in the A/V-concurrent condition from Exp. 1 and the 400ms and 600ms SOA semi-concurrent conditions from Exp. 2 could selectively attend to the subset of objects that exhibited the target feature that was mentioned first in the speech stream. A possible explanation could be that as the linguistic information is processed continuously with the visual display available search processing is able to enhance the salience of the group of items sharing the feature first mentioned and suppress the salience of the unmentioned objects. Thus, when the second feature is mentioned an easier search of a subset of salient objects can be made, as the unmentioned objects have been suppressed.

The results indicate that by simply adjusting the timing of spoken query so that the two target-feature words were presented at the same time the visual display was visible, A/V-concurrent condition, allowed participants to find the target object in a way that was considerably less affected by the number of distractors, as congruent with Spivey et al. (2001) and Reali et al. (2006). These results appear to support the notion that in the auditory-first condition, the search process may employ a conjunction prototype to find the target, thus forcing a serial-like process similar to an attentional spotlight sequentially comparing each object with the a conjunction template.

However, results indicate that in the A/V-concurrent condition, it appears that the incremental nature of auditory speech comprehension allows the search process to begin when only a single feature of the target identity has been heard. When the initial feature is recognized the search proceeds in a parallel-like process, biasing the mentioned objects by increasing their salience and decreasing those that were not mention. Now when the second feature is mention it can be used to find the target among a subset of objects with saliency already elevated, thus considerably improving the efficiency of search.

Computer simulations continue to give us insight into the underlining processes of human behavior. For example when the overall localist attractor network model is compared with data from both Exp. 1 and 2 we observe a good RMSE = 149.99, and a highly significant Pearson r-squared value, $r^2 = 0.904$, $p < 0.0001$. We can postulate that visual search processing when mediated by linguistic input is likely to utilize a mechanism similar to a normalized recurrence...
Conclusion

These findings demonstrate and support a dynamic and fluid interaction between visual perception/attention and real-time linguistic processing and comprehension. We demonstrated that when observers were placed in a semi-concurrent condition, where observers are presented with one adjective describing the target identity before onset of the visual search display and one concurrently with onset of the display, they exhibited a non-linear improvement in search efficiency, similar to an A/V-concurrent, when a 400ms SOA or more is provided. Furthermore, the good overall fit of our localist attractor network model to our post hoc simulation of Exp. 1 data and to our a priori predictions of Exp. 2 data support a search process analogous to a normalized recurrence competition algorithm.

These results parallel a rapidly growing body of work finding behavioral experimental evidence of functional and dynamic interaction between visual processing and linguistic processing (Anderson, Chiu, Huette, & Spivey, 2010). In previous studies the order of feature delivery was reversed to the orientation feature was identified first the effects found were marginal, which suggests an effect of linguistic fluency. It may be interesting to investigate the use of different words on this task. Furthermore, it is worth investigating whether the increase in efficiency observed with linguistic mediation is unique to incremental linguistic comprehension or simply the result of incremental processing of information?

References


Lupyan, G. & Spivey, M.J. (2010). Now you see it, now you don’t: Verbal but not visual cues facilitate visual object detection. PLoS ONE,.


