EPIC Models of Auditory and Visual Tasks

Final Report
Project N00014-16-1-2560

David E. Kieras & Gregory H. Wakefield
University of Michigan

Report No. FR-14/ONR-EPIC-20

Period Covered: 1 JUN 2016 – 31 DEC 2019

Requests for copies should be sent to: David E. Kieras, Electrical Engineering & Computer Science Department, University of Michigan, 3641 Beyster Building, 2260 Hayward Street, Ann Arbor, MI 48109-2121, kieras@umich.edu.

This page left blank

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
10/04/2020	Final Report	06/01/2016 - 12/31/2019

4. TITLE AND SUBTITLE
EPIC Models of Auditory and Visual Tasks

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-16-1-2560

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
David E. Kieras, Gregory H. Wakefield

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Division of Research Development and Administration
Ann Arbor, MI 48109

8. PERFORMING ORGANIZATION REPORT NUMBER
FR-14/ONR-EPIC-20

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 N. Randolph Street Suite 1425
Arlington VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This is the final report for a project on the further development and validation of the EPIC cognitive architecture for modeling human cognition and performance in the auditory and visual realm. The architecture was extended further to account for sound and speech phenomena, with emphasis on multichannel speech comprehension in a simple command-and-control task for which considerable empirical data is available. Militarily-relevant visual search tasks were modeled in considerable detail for both aggregate and individual performance. Because EPIC relies on simpler and more empirically-based visual mechanisms than current theories, the results are especially valuable for both science and application.

15. SUBJECT TERMS
Cognitive Architecture, Human Performance Modeling, Multichannel Speech, Visual Search

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UL	101	

This page left blank

EPIC Models of Auditory and Visual Tasks
Final Technical Report

ONR Grant N00014-16-1-2560

Period Covered: 1 JUN 2016 – 31 DEC 2019

David Kieras, Principal Investigator

Gregory H. Wakefield, Co-Principal Investigator

I. Project Data

A. David E. Kieras (PI), Gregory H. Wakefield (Co-PI)

B. University of Michigan

C. ONR Award No: N00014-16-1-2560

D. EPIC Models of Auditory and Visual Tasks

Reporting Period: 1 JUN 2016 - 31 DEC 2019

II. Major Goals and Approach

As the Navy takes advantage of new technologies and seeks to do more with fewer personnel, it is essential to design systems for human operators that are effective, reliable, and efficient. Computational cognitive architectures that synthesize psychological results and theory support practical predictions of human-system performance. This project extends the EPIC architecture in two areas: human audition, as applied to multi-speaker speech perception and spatialized audio, and human vision, as applied to visual search tasks typical of radar console operators.

Administrative Note: This project formally terminated on 30 DEC 2019, but work continued, with the plan being to prepare some technical reports and publications on the last set of results obtained before project termination, and some additional results obtained after the funding terminated. However, the disruption caused by the COVID-19 pandemic has forced a delay in completing these activities. In the body of this report, we note what work was delayed, and our plans for completing it in the near future.

Major Goals

Goal 1. Models of Multitalker Speech Tasks. The current model for two-talker speech tasks will be substantially refined to use fine-grain signal and timing information in the signals to explain effects in more detail, and to support generalization to multiple talkers and other classes

of competing signals. As part of this effort, phenomenological explanations for CRM results based on glimpsing with be operationalized within the more mathematically-rich domain of detection theory.

Goal 2. Modeling the Effects of Spatial Location. A preliminary model for how spatial location of sound sources is used in two-talker talks will be refined, generalized to three or more talkers, and the underlying model will be applied to the use of localized signals and cues for enhancing performance.

Goal 3. Extending the Visual Architecture. The current architectural components for vision and eye movements will be upgraded to increase the accuracy and generality of models of visual search.

Goal 4. A Preliminary Model of Auditory-Visual Integration. A basic model of auditory-visual integration will be proposed based on the results from modeling the use of localized sound in visual search.

Approach

The technical approach in the project is to develop the auditory and visual architecture by determining how to "black-box" low-level sensory and perceptual processing in a way that when combined with task strategy at the cognitive level, enables robust and useful models of how the visual and auditory information is used in a complex and practically significant task. Thus the modeling is a combination of symbolic processing at the cognitive level, and mathematical modeling at the sensory and perceptual level. This project seeks to "open up" the previous black boxes by representing lower-level sensory and perceptual processing and how they account for task performance, in both the auditory and visual domains.

The data to be modeled is both published data in the literature, and also raw data collected by our collaborators, often with our suggestions on manipulations, which is analyzed and then modeled in much more detail than is typical in the empirical literature.

A key feature of this approach is that there is no a-priori assumption of an attention-based limitation on perceptual processing; rather, performance limitations are attributed first to sensory-peripheral limitations, and second to task strategy. Central limitations on processing will be uncovered by this approach, rather than assumed in advance.

III. Concise Accomplishments

Goal 1. A neural-based point-process detector was integrated into a computational model of the peripheral auditory system. This detector was also modified to allow for automatic detection, as opposed to guided detection from the top down.

A non-speech paradigm for assessing human performance and modeling glimpsing has been developed using granular synthesis to create infrapitch signals and characterize their perceptual properties. The virtue of this new approach lies in the fact that the components that give rise to coherent auditory sources are directly observable from the algorithm used to create the sounds. This advance opens up a realm of possibilities for better understanding the triggers and acoustic features that induce the formation of an auditory source from an otherwise completely incoherent

signal. During the present funding period, the neural-based point-process detector was integrated into the analysis algorithms to represent infrapitch events.

Finally, two components of the auditory module were elaborated to establish a direct connection to the acoustics of speech signals and the auditory front-end. This resulted in a more flexible and generalizable definition of an auditory stream as well as a predictive model for the detection of linguistic content in two-talker speech environments.

Goal 2. As described in the 2017 Annual Report, the work on modeling a two-talker listening task was extended to handle conditions in which the talkers are spatially separated. A new state attribute, spatial position, was added to the stream tracker and the variability in observed position within each utterance was determined by a nonlinear sensory-theoretic mapping of minimum audible angle. Word detection functions were updated to depend on the larger of the left- and right-ear target-to-masker ratios, using head-related transfer functions weighted by the Speech Intelligibility Index. Although fits to the data from a spatial version of the two-talker CRM experiment are excellent, the data are sufficiently variable to preclude further refinement and justification of several modeling assumptions.

Goal 3. As described in the 2017 Annual Report, the main focus on complex visual search produced a model that accounted for many details of task performance, but the overall fit of the model to multiple aspects of the data was not satisfactory, which motivated a complete reanalysis of the data, but also clarified which low-level architecture modifications might be needed. Further modeling of the complex visual search task was deferred, while work was done on simple visual search tasks which provided a context in which low-level issues might be better explored. One such lower-level visual mechanism is *crowding,* which was added to the architecture. The resulting models accounted for both search time and errors in simple visual search tasks much more rigorously and parsimoniously than existing theories of visual search, and also showed how individual differences in search performance was due to a combination of individual visual ability and choice of task strategy. This line of work was very successful.

Goal 4. No substantive progress was made on this specific goal, because the expected adequate experimental data set was not developed.

IV. Expanded Accomplishments

Technical Approach

The EPIC Cognitive Architecture

This section presents a summary of the EPIC cognitive architecture. Extensive presentations of EPIC are available elsewhere (Meyer & Kieras, 1997a,b; Kieras & Meyer, 1997; Kieras, 2004; Kieras, 2016), so here only a brief sketch will be presented.

Figure 1 shows the overall structure of the EPIC architecture. In overview, EPIC provides a general framework for simulating a human interacting with an environment to accomplish a task. The EPIC architecture consists of software modules for the simulated task environment or device that interacts with a simulated human, which consists of perceptual and motor processor peripherals surrounding a cognitive processor. The device and all of the processors run in parallel with each other. To model human performance in a task, the cognitive processor is programmed

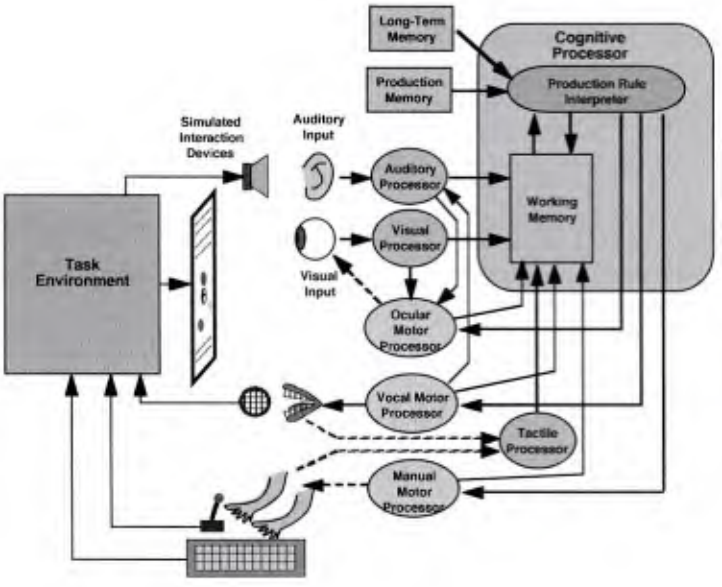

Figure 1. The EPIC architecture in simplified form. The simulated environment, or device, is on the left; the simulated human on the right.

with production rules that implement a strategy for performing the task. When the simulation is run, the architecture generates the specific sequence of perceptual, cognitive, and motor events required to perform the task, within the constraints determined by the architecture and the task environment. Monte-Carlo runs of the simulation produce predictions of human performance, both actual behavior sequences as well as statistical aggregates.

More specifically, the task environment (also called the simulated device, or simply the device) is a separate module that runs in parallel with the simulated human which is represented as a set of interconnected processors and simulated sensors and effectors. The cognitive processor consists of a production rule interpreter that uses the contents of production memory, long-term memory, and the current contents of a production-system working memory (PSWM) to choose production rules to fire. Production rules are simply if-then rules that represent the procedural knowledge of how to perform a task.

The cognitive processor runs on a 50 ms cycle. At the beginning of each cycle, the conditions of all of the rules are tested in parallel against the contents of PSWM, and those whose conditions match are fired and their actions executed. The actions can modify the contents of working memory, which may change which rules will match on the next cycle, or instruct motor processors to carry out movements. Auditory, visual, and tactile processors deposit information about the current perceptual situation into working memory; the motor processors also deposit information about their current states into working memory. The motor processors control the hands, speech mechanisms, and eye movements. All of the processors run in parallel with each other. The pervasive parallelism across perception, cognition, and action motivated the design of EPIC and is reflected in the acronym: Executive Processes Interact with and Control the rest of the system by monitoring their states and activity.

Models built in EPIC are "end to end" in the sense that they start with simulated sensory input and produce simulated physical movements; in practice, these are abstracted and simplified, but the constraint is that the model matches the basic requirements of the complete human task, rather than focus only on the internal purely cognitive processes, which has been the overwhelming focus of most cognitive architecture work.

The EPIC Philosophy. A unifying principle of EPIC modeling is a research strategy that leverages architectural commitments in EPIC to arrive at parsimonious, well-characterized, and accurate models of human performance. This research strategy can be stated as follows:

- As much as possible, characteristics of human performance will be attributed to perceptual and motor abilities and limitations rather than immediately postulating elaborate cognitive limitations or abilities such as limited central capacity or covert selective attention; this not only takes advantage of the large corpus of empirical results on sensation and perception and motor movements, all directly relevant to these perception- and motor-heavy tasks, but also helps ensure that the perceptual and motor mechanisms are close to the relevant data and well-characterized, rather than being obscured by unnecessary assumptions about cognitive processing.
- Because the architecture allows task strategy to be explicitly represented in a uniform way, it is possible to examine alternative cognitive strategies to distinguish underlying perceptual/motor constraints from task-specific strategy effects.
- Thus, to explain or predict a specific observed empirical effect, first priority is given to incorporating known properties of human perceptual and motor systems, and second priority is given to examining whether a different task strategy can account for the effect. If necessary, the adequacy or accuracy of the perceptual or motor mechanisms can be reassessed in the light of the empirical effects and whether strategy modifications can account for them, and justified changes made.
- Third, additional internal cognitive mechanisms would be entertained only to the extent that these first two classes of explanatory mechanisms cannot account for the effects.

EPIC's emphasis on perceptual-motor processes is a key feature compared to other cognitive architectures, and is what makes EPIC a good foundation for work in audition and vision. This emphasis emerged in the original ONR-supported project to develop EPIC beginning in 1992 (David Kieras and David Meyer, Co-PIs) with the goal of modeling multitask performance and mental workload. The perceptual-motor mechanisms and requirements emerged as fundamental constraints on task performance that must be included to produce comprehensive and accurate models of multitask performance.

The basic approach is to represent fundamental sensory-perceptual and motor mechanisms with mathematical models in EPIC's perceptual and motor processors; the perceptual mechanisms deliver perceptual descriptions to the cognitive processor, which uses a production-rule representation to perform qualitative symbolic operations such as making inferences, selecting relevant information, and following task procedures. This hybrid combination of mathematical models for low-level phenomena and symbolic models for high-level cognition

works especially well in modeling how perceptual effects operate in the context of whole tasks. As such, EPIC has served as an efficient and elegant means to summarize a large body of experimental results in the perceptual, cognitive and motor domains as they relate to human performance in complex tasks. It serves as a testbed for optimizing human-computer interfaces, and, going forward, could provide an exploratory tool for integrating both biomimetic and novel sensory and motor systems with decision systems in autonomous systems such as robots.

The project goals cover both auditory and visual phenomena, and a preliminary look at their integration. A unifying principle of this work is that both areas of work will reflect two architectural commitments in EPIC that are also research strategies: First, as much as possible, characteristics of human performance will be attributed to perceptual limitations rather than cognitive mechanisms such as selective attention; this helps ensure that the perceptual mechanisms are well-characterized rather than being obscured by unnecessary assumptions about cognitive processing. Second, because the architecture allows task strategy to be explicitly represented in a simple way, it is is possible to work with alternative cognitive strategies, as well as different perceptual mechanisms, to help distinguish underlying perceptual abilities from task-specific strategy effects. In addition, because of EPIC's basic simplicity, the concepts in our models should be easily portable to other architectural approaches or related research communities. Newer architectures, such as ARCADIA, will likely benefit from our modeling concepts. In addition, as called out by Durlach (Durlach, Mason, Kidd, Arbogast, Colburn, & Shinn-Cunningham, 2003), a long-standing leader in psychoacoustics research, a theory-guided approach to the characterization and understanding of informational masking is likely to be the only way the field can make progress in the study of auditory perception in complex acoustic environments. We believe the structural separation within EPIC between cognitive strategies and sensory capacities presents the ideal framework within which to build such theories, as evidenced by our success in modeling speech perception in multi-talker environments.

The following sections are organized by project goal; the auditory sections are presented first, followed by the visual section. Each section will present additional information about the architecture mechanisms and their development in this project.

Goal 1: Modeling Multi-Channel Speech Processing

Summary of Previous Work

The Annual Reports for 2013-2018 summarize a set of key accomplishments which involved the use of some new empirical data collected by our AFRL collaborators. More details can be found in the previous Annual Reports and the Kieras & Wakefield (2014) Technical Report. The model elaborated in these reports draws on speech recognition data using the Coordinate Response Method (Bolia, Nelson, Ericson & Simpson, 2000), which can be considered to be a limiting case of the so-called cocktail party effect where the words spoken by two or more talkers align perfectly in time.

The standard EPIC model developed from this work consists of a black-boxed *perceptual* module and a cognitive-level listening strategy for correctly reporting the word contents of a *target* utterance in the presence of one or more *masker* utterances. In the perceptual module, an auditory stream is formed for each talker by tracking the pitch and intensity of each spoken word. Errors in the reporting of word content are due either to mis-tracked words (because a word is assigned to the wrong talker) or to missing word content (because a spoken word is masked by other talkers). Coupled with a sub-optimal task strategy, the model is able to account for 99% of the variance, although systematic errors remain.

The Annual Report from 2017 summarized the beginning of our efforts to open the perceptual black box by extending the tracking module to instantaneous pitch and intensity rather than average pitch and intensity at the time scale of a word. Recent innovations in statistical detection/classification were incorporated to yield a tracking system that utilizes Kullback-Leibler divergence (Kullback & Leibler, 1951) for collections of instantaneous pitches and intensities (sampled every 10 msec) to update track state. This extension eliminated the systematic errors of the standard model without degrading the fit to the data and the three free parameters associated with tracking.

As an initial exploration of the content-detection module, we considered the use of an auditory stream's instantaneous pitch to identify time-frequency regions where that stream dominates the masker's. By considering the relative strengths of projections onto the appropriate target or masker subspace, we were able to identify time-frequency "islands" for each talker which are akin to glimpses. We fully developed this approach in the current funding period.

In 2017-2018, we grounded the vague notion of *glimpsing* within the more mathematical framework of multiple looks as originally developed in signal detection theory. From the perspective of opening the black boxes of EPIC's auditory modules, we posed the concept of a *look* at the level of the VIII-th nerve with the hope of integrating the current body of research on auditory front-ends to provide a direct link between the acoustic signal and the construct of a *stream* as used in EPIC. The primary outcome was a mathematical form of a Poisson-driven multiple-look detector that emulated both the steeper slopes associated with the psychometric functions for *energetic* masking as well as the shallower slopes associated with those for *informational* masking. Unique to this detector was the mechanism of adaptively pooling firings from a family of neurons over the time-course of the signal of interest.

Within this framework, *glimpses* become concatenations of the more primitive construct of a *look*, which is thought to be related to the limits of temporal resolution, e.g., 3-5 msec. As such, the actual duration of a glimpse is likely to be defined operationally by the sensitivity and performance of the detector, its pooling strategy, and the signal conditions.

Beginning in 2016-2017, we have begun exploring *infrapitch* as a more fine-grained experimental tool for studying the processes involved in auditory stream formation and glimpsing. In its original formulation by Warren and Bashford (1981), *infrapitch* refers to the perceptual features of a periodically-repeated wideband noise. As the repetition rate decreases from 20 per second to one per second, perception changes from a single aggregate "motorboating" sound, to a "washing machine sound", and finally to temporally-isolated transients in a background of wideband noise. Our interest has been primarily focused on the regime of 1- to 1/3-repetitions per second based on the hypothesis that the perceptual formation of such temporally-isolated "figures" against a background of wideband noise reflects the build-up of *proto-streams* into full-fledged, coherent and identifiable *auditory streams*. In this sense, we propose that *infraptich* taps into *acoustic salience* whereby events, otherwise unexpected and unknown, are heard. Following some initial taxonomic characterization of the phenomenon in 2016-2017, we developed a family of signals in 2017-2018 based on granular synthesis, which allowed for more systematic control over the signal acoustics than is otherwise provided by wideband noise generators.

In addition to EPIC's modeling of pure-tone detection and infrapitch, our 2018-2019 research returned multi-talker speech communication and the black boxes employed in EPIC to model such. The discussion that follows considers each of these in turn.

Current work: Implementation of a pooling detector in an auditory model of the VIII-th nerve

Rationale

As noted in our 2017-2018 progress report, our efforts to open up the auditory module's black boxes have focused on a better characterization the statistical information present at the level of the VIII-th nerve. This approach was adopted for the following reasons:

- a variety of auditory front-ends are available that model the organization of sound at the auditory periphery,
- a detector posited at the VIII-th nerve "output" can leverage the codification these auditory front-ends provide and thereby reduce the potential for EPIC modeling artifacts/errors if a more ad hoc approach to coupling the acoustics of the signal to a time-frequency representation were to be used instead,
- such a detector can be designed and developed *in vitro* using reasonable inputs before it is integrated into the more complicated acoustics-transduction-neural-response system,
- parsimony argues that one should avoid the standard problems of Gaussian (or other) approximations by forcing the detector to deal with the strict dependence of the mean-to-sigma

ratio on number of "looks" for Poisson processes, given that such dependence is effectively baked into the auditory system beginning at the auditory periphery.

Simplifying assumptions (2017-2018)

Our work during the 2017-2018 period made a number of simplifying assumptions, including

- Poisson approximation.

  We ignored absolute and relative refractory periods in modeling the detector behavior. These clearly are important in neural processing. However, we noted that we are developing the detector based on Kullback-Leibler divergence, for which extension from Poisson to a dead-time Poisson process have already been worked out (Gruner & Johnson, 2001).

- High spontaneous response units.

  Neurons in the VIIIth nerve are differentiated into sub-populations by thresholds and spontaneous discharge rates. We considered a high spontaneous-rate unit with a discharge rate of 100 spikes/sec. This is a reasonable assumption as high spontaneous rate units are generally considered to be responsible for detecting sinusoids at the lowest amplitudes.

- Range of average discharge rate.

  As with spontaneous discharge rate, the driven range differs across the population of neurons. We considered a "middle of the pack" neuron that has a maximum average discharge rate of 300 spikes/sec.

- Ignore onset effects in the rate-intensity functions.

  At stimulus onset, the discharge rate often reaches as high as 700 spikes/sec, but rapidly adapts to the average discharge rate for that unit. We ignored such onset effects.

Computational audition: University of Rochester EAR model

During the current funding period, each of these simplifying assumptions was eliminated by implementing the detector using the University of Rochester EAR model of the auditory VIII-th nerve (Carney, 2019). Over the years, Carney and her students have developed a flexible and well-documented framework within which to simulate human performance in a variety of auditory tasks. Their approach has considered primarily the average neural response at the levels of both the VIIIth nerve and inferior colliculus. We've taken advantage of their Matlab tools to extend and validate the pooling detector structure that was developed last year at a more molecular level that involves responses of individual nerves.

Our final implementation is based on UR EAR v2.1 and specifically utilizes Carney's most recent modules for the IHC (*model_IHC_BEZ2018.m*) and synapse (*model_Synapse_BEZ2018.m*). Accordingly, with respect to the simplifying assumptions from 2017-2018, the following have been addressed:

- Point-process model for neural discharge (replaces 2017-18 Poisson approximation)

  The EAR synapse module outputs a spike probability for each time sample, where the sampling rate is 100 kHz, to permit more fine-grained analysis of interspike intervals, for example. The module includes an absolute refractory period of 0.7 msec and a relative refractory period of 0.6 msec. Its outputs include the instantaneous spike probability and two analytic expressions for the instantaneous mean and the instantaneous variance of the firing rate. In our simulations, we used these mean and variance expressions.

- Arbitrary spontaneous discharge rate (replaces 2017-18 high spontaneous response units)

  The previous synapse model of the EAR package distinguished among low, medium, and high spontaneous rate units. The upgraded synapse model allows for arbitrary spontaneous discharge rates.

- Built-in IHC/BM nonlinearities (replaces 2017-18 fixed range of average discharge rate)

  The initial stages of the EAR emulate the compressive nonlinearities of BM/IHC sub-systems. We use these as written.

- Built-in transient response (replaces 2017-18 lack of onset effects in the rate-intensity functions)

  As with the modeling of I/O nonlinearities, UR EAR v2.1 emulates the onset transient behavior of VIIIth-nerve response.

Results

Figure 1.1 shows psychometric functions for a tone-in-noise detection task using the UR EAR model. The signal is a 100 msec 1-kHz sinewave and the noise is a wideband noise with a

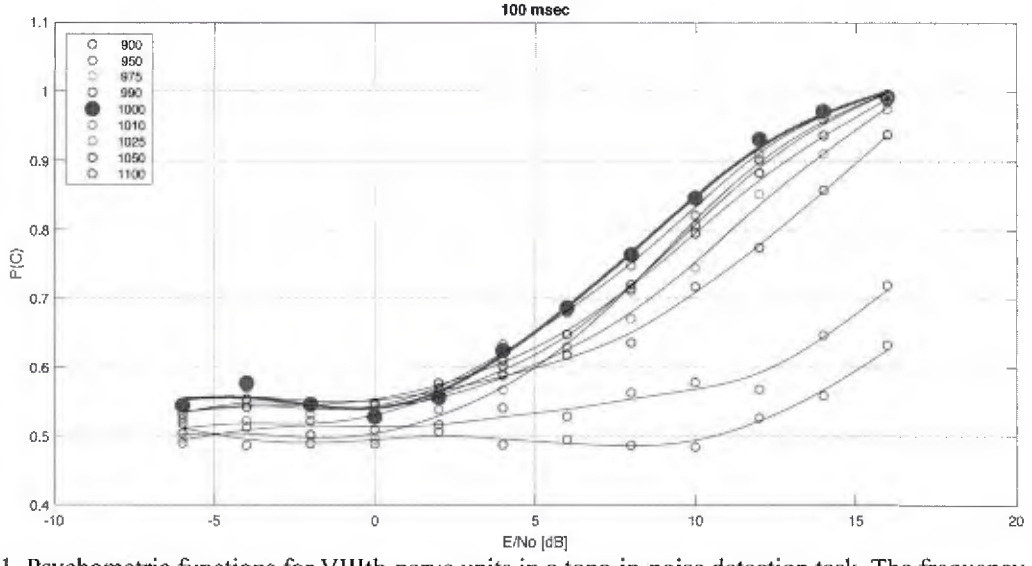

Figure 1.1. Psychometric functions for VIIIth-nerve units in a tone-in-noise detection task. The frequency of the sinewave is 1000 Hz and the duration is 100 msec. The parameter is the center frequency of each individual neuron.

spectrum level of 20 dB. Spontaneous rate for this particular case is 80 spikes/sec. Performance was simulated using a fixed-level two-interval psychophysical procedure in which the number of firings to the tone-in-noise was compared to that of a noise-alone trial. As discussed in greater detail in last year's progress report, this particular decision statistic is consistent with a matched filter for which we predict a slope of around 2.5%/dB. In the present case, the slope is a little higher (3.75%/dB), but still remains well below the much steeper psychometric functions observed in human signal-detection tasks (5%/dB). Another interesting result is that the slopes are shallower for the "off-frequency" case, where, for example, a neuron with a characteristic frequency of 950 Hz is both less sensitive for comparable E/No and incremental gains in performance with intensity are less than that for the on-frequency case. For comparison purposes, Figure 1.2 displays the psychometric functions for last year's point-process pooling detector. In all but the shortest signal durations, these functions achieve slopes of 5%/dB.

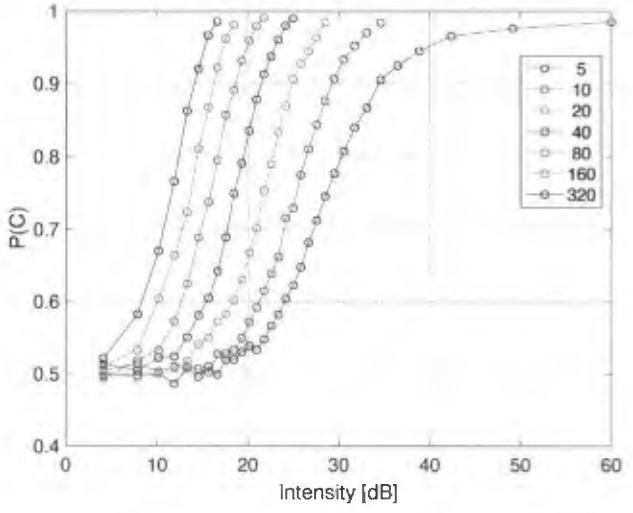

Figure 1.2. Psychometric functions for the point-process pooling detector developed during the 2017-18 period. For all but the shortest durations, these functions achieve the desired slopes of 5%/dB.

With a functioning auditory model in hand, we proceeded to develop a pooling scheme to replicate both the desired slopes of the psychometric functions as well as the time-intensity tradeoff shown in temporal integration. These yielded a form similar to that which was mathematically derived for last year's pooling detector. However, both of these pooling detectors rely on a external triggers to determine when a desired signal begins and ends.

Before proceeding with the main tasks of this year's research (opening the black boxes of EPIC's model for the CRM task), we addressed the issue of such triggers. From this effort, we developed a point-process detector that self-regulates by monitoring both a "wideband state" (by pooling neural discharges over brief "looks") and an "adaptive state" (by pooling over multiple looks based on iteratively smaller and smaller bands of neurons). The behavior of this system is inherently greedy: as long as the adaptive state suggests an improvement in the local SNR, the

adaptive state continues to evolve. However, once the incremental gain of the adaptive state is no better than that of the wideband state, the detector re-initializes.

Figure 1.3 provides an example of the behavior of this self-regulating detectors. When such detectors are initialized at frame 31, some of the 170 detectors continue to evolve by drawing upon increasingly narrower pools, while others wax and wane between the pooling and wideband states, and still others in the neighborhood of the sinewave monotonically continue to grow. In this context, a tone is detected once the local value of the accumulated multiple looks exceeds a threshold. The duration of such a tone can similarly be determined by computing the number of consecutive looks over which the detector remains in the adaptive state. What is attractive about this approach is that tones are detected "automatically", with the only top-down "cognitive" direction being whether the detected tone is in the neighborhood of what is expected. An intriguing feature of these temporally extended waxing/waning states is that it may explain why the detection of uncertain frequencies has such a small effect on detection thresholds (3 dB at best). Specifically, by widening the range of what constitutes an expected frequency, it is necessary to eliminate the potentially many false alarms that otherwise occur.

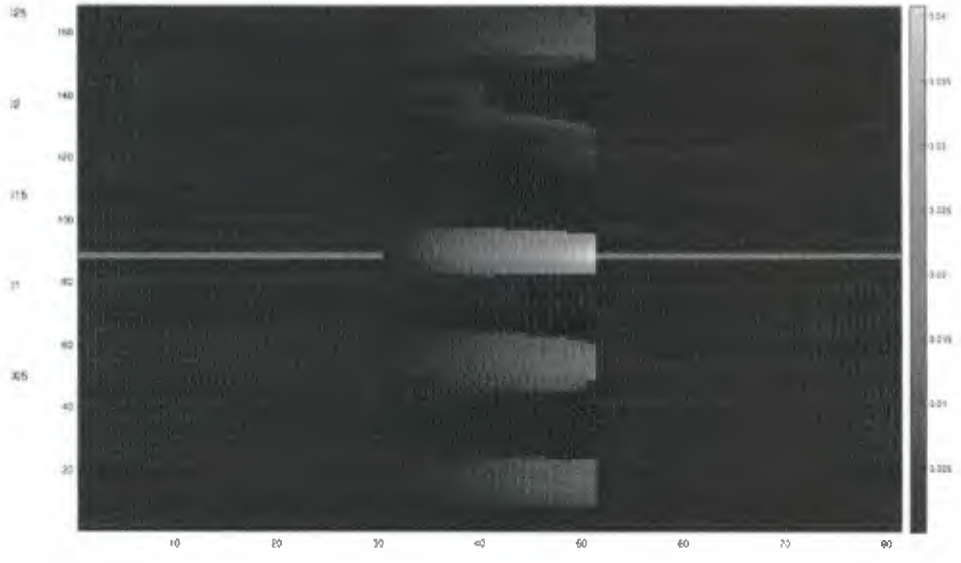

Figure 1.3. Example of the behavior of a self-regulating version of the pooling point-process detector. For demonstration purposes, the display shows the time-frequency surface of a tone-in-noise before the self-regulating detector is initialized at frame 31 and after it is (arbitrarily) terminated after frame 52. The tonal signal is displayed as the horizontal line around channel 85.

Conclusions and Project Relevance

The primary goal of this year's work was to demonstrate that the simplifying assumptions used to derive the form of the point-processor pool detector (2017-2018 outcomes) are not overly restrictive when based on more accurate models of the VIII-th nerve. When integrated into the UR EAR model, one of several state-of-the-art models of cochlear transduction, we have shown that the point-process pooling detector can emulate auditory detection of pure tones in noise. In

addition, the detector was modified to include a mechanism for self-regulation, thereby avoiding the need for external start/stop triggers.

How to reconcile the long time constants observed in temporal-integration experiments with the short time constants observed in temporal-resolution experiments has been a long-standing issue in psychoacoustics. An important contribution to this discussion is our results summarized in this section: a parsimonious accounting models both as an adaptive pooling of VIII-th nerve activity across multiple fibers. In our development, we also observed how the slopes of the psychometric functions can vary under different detection strategies based on these primitive adaptive-pooling units. We believe this result is key to expressing the heuristic concepts of informational and energetic masking in more formal mathematical language. While this is beyond the scope of the current project, it is highly relevant to modeling the transition from the two-talker to four-talker CRM tasks, where the underlying psychometric functions continue to steepen as the number of talkers increases.

COVID-19 delay. A draft of a brief JASA paper outlining the basic mathematical structure of the adaptive-pooling detector had been tabled while work proceeded on modeling the CRM task at a finer temporal and acoustical scale. With the outbreak of COVID-19 in early March, both adding teaching and administrative responsibilities at the University of Michigan shelved this and the completion of a technical report on EPIC modeling of the CRM task. These two tasks are expected to resume once Winter semester is over (May 1) and the final work will be submitted to ONR as soon as possible. In addition, there are some lessons to be learned in what front-end models of the auditory nerve can and cannot do well with regard to modeling the types of psychophysical experiments of interest in the current project. It is hoped that a shorter, technical document will be completed that documents our experiences and observations.

Current work: a potential experimental paradigm for characterizing auditory source formation

Background

At the end of last year's report, we noted that the most important outcome of our work on this topic was a much more reliable body of methods for generating infrapitch sounds and relating their character to segments (temporal and spectral) of the granular sound. Because these sounds have been constructed using known sound "atoms", we don't require statistical signal processing to characterize the short-time spectral properties of the granular noises. Instead, we can "read off" these properties directly since we know exactly where the sound atoms are placed in frequency and time.

Updates and Future Work

We continued to refine our collection of techniques for manipulating granular sources to identify where, in time, an infrapitch sound object exists. With the development of a self-regulating point-process pooling detector, we also explored the idea of a *proto-stream* that develops over time into a coherent *auditory stream*. In this context, we note that the self-regulating pooling detector has three states - wideband, adaptive/pre-detect, and adaptive/detect. We conjectured that infrapitch sound objects first exist as proto-streams, where a collection of

self-regulating detectors achieve the adaptive/pre-detect state for a brief period of time before they reinitialize. However, when the granular sound repeats, *the same collection of self-regulating detectors will again achieve the adaptive/pre-detect state.* A reasonable strategy in auditory processing is to store potentially interesting collections to see whether they occur again. With sufficient re-occurence of such collections over time, a *auditory stream* is created.

Our preliminary work appears promising. Figure 1.4 shows the output of a collection of self-regulating pooling detectors for two periods of an infrapitch noise. What is interesting about this figure is how generally "quiet" the wideband infrapitch noise is, despite the fact that one hears a fairly dense, stationary wideband noise. That said, there appears to be some activity, none of which is sufficiently energetic to be heard on first listening, but could evolve into a coherent source with enough repeats. In the present case, the most intense of these proto-stream regions (820 msec/channels 50-60) are heard to evolve into an infrapitch event.

From these observations, we have a number of research questions that fall beyond the scope of the present project but which may well be appropriate for study under future funding.

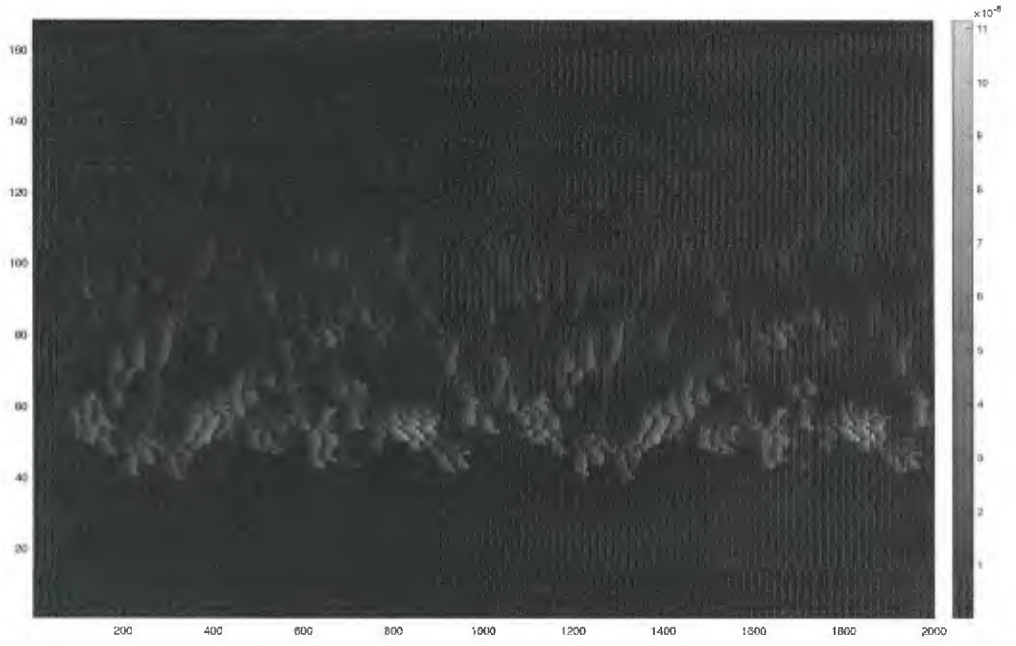

Figure 1.4. An example of a pooling detector representation of two 1-second periods of a wideband granular infrapitch sound. When listening to this sound, an organized infrapitch event is heard to evolve from the proto-stream in the neighborhood of 820 msec and neural channels 50-60.

Current work: Opening the talker-utterance black box

In our original formulation of the auditory module, a speech utterance was modeled as a sequence of words wherein each word was described by *content*, *pitch*, and *loudness*. The presence of two or more talkers who may be speaking simultaneously necessitates the need for perceptual rules to associate each of a collection of words to one of a collection of talkers. These

rules developed for the auditory module were embodied as a *tracking* algorithm that creates and maintains *auditory streams*, one for each talker.

Our work separated the linguistic content of each word from the more primitive psychoacoustic properties of pitch and loudness, and built the tracking algorithm based on the pitch and loudness of individual words[1]. As observed in the human data, the greatest source of error in the CRM task reflects intrusions of the masker word in the target response. These are handled in the EPIC model by errors in the tracking algorithm whereby a masker word is mistakenly assigned to the target stream. In combination with an appropriate sub-optimal task strategy, the remaining errors are handled in the EPIC model by content-detection functions that determine whether or not the content of a call sign, color, or digit word of the *talker stream* is masked by that of the *masker stream*. After determining that the two-stream approach could account for the data using average pitch and loudness values for each utterance, the tracking algorithm was further extended to handle instantaneous values of pitch and loudness through the use of a Kullback-Leibler classifier.

Black boxing of the auditory module occurs at the level of pitch, loudness, and linguistic content. Our approach this year was the following:

- We retained pitch as a primitive that is accessible directly through perception based on the abundance of evidence that listeners have no problem following multiple pitch-bearing sources such as musical instruments and voices;

- We eliminated loudness as a feature that is directly accessible through perception - to model it otherwise would require modeling partial loudness, which remains mostly an art rather than a science - but we retained a feature that reflects some measure of relative intensity;

- With these two "streaming variables" fixed, we investigated the mapping of the acoustic features of speech to linguistic content, particularly for a mixture of two or more speech utterances, in light of the CRM data reported in Thompson et al (2015).

Modifying the stream states: Voiced (pitch) + Voiceless + Level

One of the conveniences of modeling the talker stream at the word level was that we didn't need to address those segments of a word in which voicing did not occur. Stream state at the word level consisted of a single pitch (or collection of pitches in the updated model) and a single level (or collection of levels in the updated model), in which the latter acted as a surrogate for the mix of voiced and voiceless segments within a word.[2]

The acoustics of the speech signal, of course, are far more complicated. The panels shown in Figures 1.5-7 portray the occurrence of voiceless events over the CRM corpus (indicated by

[1] This is an important simplification that is worthy of exploration in the future, inasmuch as there is some experimental data that suggests linguistic content may also be fed into a tracker.

[2] Recall that the voiced segments are far more energetic than voiceless, so that a word with a relatively high proportion of voiceless-to-voiced (e.g., six) will be encoded as softer than one with a low proportion (e.g., one).

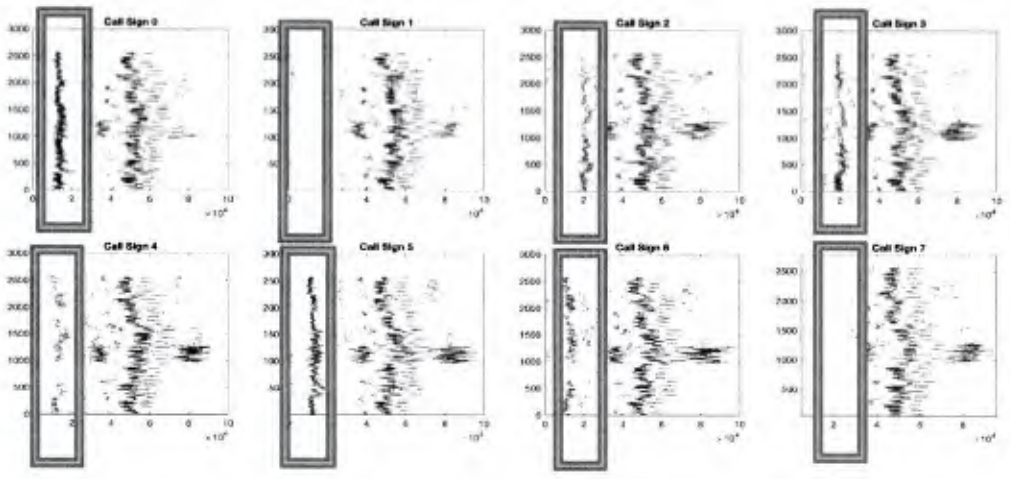

Figure 1.5. Occurrence of voiceless events for each utterance in the CRM corpus as a function of Call Sign. Boxes highlight the segments during which the call sign is spoken. Call Signs are indexed from 0 to 7, corresponding to Charlie, Ringo, Laker, Hopper, Arrow, Tiger, Eagle, and Baron.

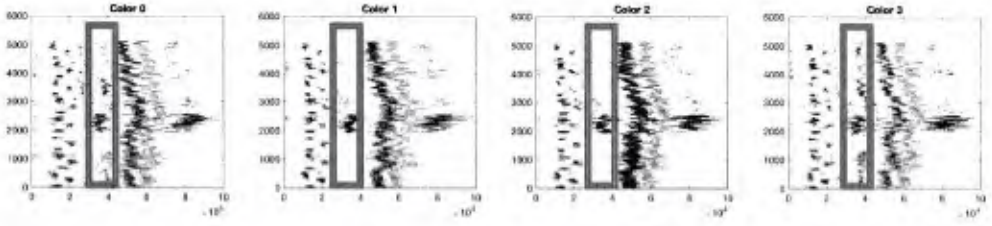

Figure 1.6. Occurrence of voiceless events for each utterance in the CRM corpus as a function of Color. Boxes highlight the segments during which the color is spoken. Colors are indexed from 0 to 3, corresponding to Blue, Red, White and Green.

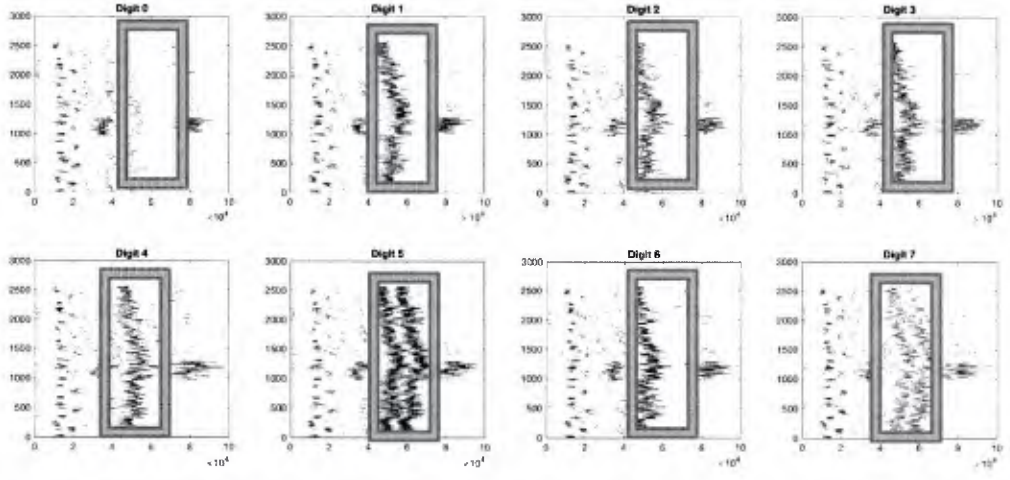

Figure 1.7. Occurrence of voiceless events for each utterance in the CRM corpus as a function of Digit. Boxes highlight the segments during which the digit is spoken. Digits are indexed from 0 to7, corresponding to One, Two, Three, Four, Five, Six, Seven, and Eight.

index along the ordinate) for Call Sign, Color, and Digit words. In all cases, the default parameters of the Praat algorithm (2001) was used to measure the pitch of each utterance in isolation. As can be seen, the relative occurrence of voicing differs from word to word, but, more importantly, *across* word groups. In particular, the Baron and Ringo call signs are almost entirely comprised of voicing, the Hopper, Charlie, and Tiger call signs have substantial voiceless segments, while the remaining call signs lie somewhere in-between. Within word groups, Colors tend to be the most homogenous with the exception of one talker who tends to produce fairly noisy pitches and thus appears to be "more voiceless" than the rest of the talkers. Both the timing and preponderance of voiceless events among the Digit words appears as a discriminating feature of this group, when compared with the other two.

From these observations, we propose to modify the state of a stream to include *voiceless*. Once fully integrated into the EPIC auditory module, we expect that this put less pressure on the detection functions to vary as a function of Call, Color, or Digit. Recall that the best fitting content-detection functions in the auditory module differ in mean, depending on whether the word is drawn from the call signs, colors, or digits. These differences, while necessitated by the data, are more likely to reflect differences in the acoustics of the associated words than the content-detection process itself. By adding the *voiceless* state, we can consider a mixture of detection functions: Voiced-on-Voiced (when both the target and masker are voiced), Voiced-on-Unvoiced/Unvoiced-on-Voiced (when one source is voiced and the other is not), and Unvoiced-on-Unvoiced (when both the target and masker are voiceless). Of these, the most easily modeled cases are those involving the voiceless state where the distinctions between voiced and voiceless are likely to be the most detectable when compared with Voiced-on-Voiced. As shown in Figure 1.8, if we assume for discussion, that the original EPIC model's best-fitting detection function for Color reflects primarily the underlying detection function for Voiced-on-Voiced, then we would expect the Digits to be, on average, much better detected than Color and that performance with respect to Call signs would lie somewhere in-between. This is in qualitative agreement with the best-fitting detection functions shown in the figure.

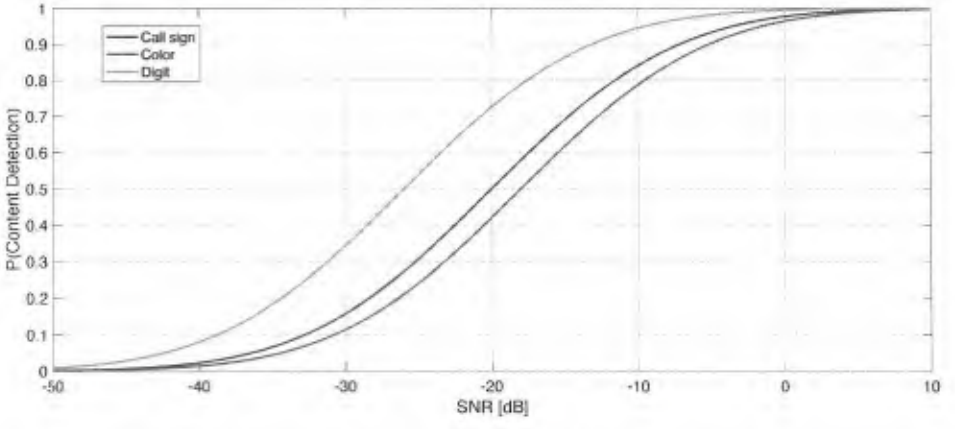

Figure 1.8. The best-fitting detection functions of the black-box auditory model share a common variance but differ in their means. From a first-principles perspective, this is an artificial distinction inasmuch as "masking" (energetic or otherwise) should reflect a single process that is independent of word type.

Gender Asymmetries

Human performance in CRM experiments depends not only on the relative intensity of the target utterance to that of the masker, but on whether both target and masker are drawn from the same talker (the TT condition), same gender (the TS condition), or different genders (the TD condition). Of the three, our work has always concentrated on the TT and TS conditions. Since these show the greatest degradation in performance as well as the highest number of masker intrusions, fitting them exercises the key components of the EPIC model: content-detection functions, the tracking algorithm, and the task strategy. However, as we looked further into the speech acoustics, we have become very interested in the TD condition. In this condition, masker intrusion errors due to the tracking algorithm are almost nonexistent, suggesting that the TD condition presents the purest data for modeling content detection.

As shown in Figure 1.9, our analysis revealed an interesting asymmetry. In this figure, the CRM results are displayed for all responses (solid symbols), cases where the target talker is male (open squares), and cases where the target talker is female (open diamonds). We see that the average difference in performance across these two conditions is 10%, which is a sizable effect when one considers that average performance in this task ranges from 65% to 90%. Given that the levels of the talker utterances are normalized, this difference is not likely due to differences in the signal-to-noise ratios of the partials. Since there are roughly two partials for each male utterance to one partial for each female, in the case of isolated partials, the male masker should be *less effective* than the female masker, whereas for the case of unresolved partials, the male and female maskers should be equally effective. Therefore, it is more likely that these differences in masking have something to do with differences in the lability of vowels to interference. Since the fundamentals of the female utterances in the corpus are roughly twice as high as those of the male utterances, the spectral sampling of the F1 and F2 formants for the female utterances will be sparser and more likely to degrade in the presence of interfering partials. This consequence of

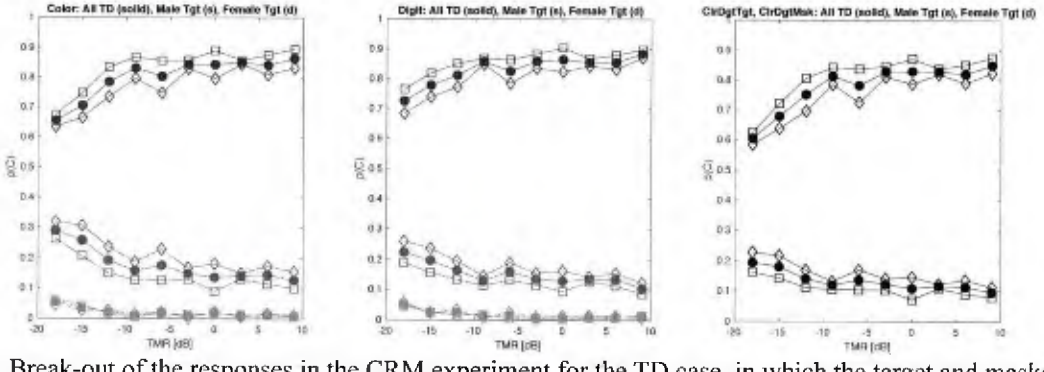

Figure 1.9. Break-out of the responses in the CRM experiment for the TD case, in which the target and masker on each trial is drawn from talkers of different gender. The left panel shows the results for Colors, the middle shows the results for Digits, and the right panel shows the results for Joint Color and Digit. Solid symbols show performance aggregated over all the data, where blue denotes a correct target response, red denotes a masker intrusion error, and green denotes a non-intrusion-error response. The open symbols show the results when either the Target talker is male (squares) or the Target talker is female (diamonds). The results show that a male masker is more effective at masking a female target than vice versa, and that this is a consistent effect regardless of the target-to-masker ratio.

higher-pitched fundamentals is also a well-known issue for automatic speech recognition algorithms.

Voiced-on-Voiced Interactions

A standard approach in computational auditory scene analysis (CASA) is to emulate the auditory front-end with a band of bandpass filters (typically gammatone filters) and to aggregate the outputs of these filters to extract amplitude modulation, common onsets/offsets, spectral centroids (brightness), and the like. With respect to the CRM task, such aggregation is problematic inasmuch as the speech of each talker will factor high along pitch, modulation, spectral centroids, etc. so that the representations are likely to be poorly separated in this feature space. Similarly, a standard approach in speech processing is to extract formants either from the time-varying spectrum of the signal (e.g., a spectrogram), or from a physical model of the vocal tract (e.g., all-pole modeling). In the case of speech, either approach works "in principle" as long as the signal of interest is that of a single talker. When more than one talker is present, as is the case with CASA, these nonlinear procedures tend to degrade rapidly; at best, one source tends to capture the parameter estimates and leaves the second source as unobservable.

Our work over the past year was motivated by the phenomenology of the CRM task. When listening to two utterances from the same talker at a TMR of 0 dB, one *hears* two talkers and one *hears* two call signs, two colors, and two digits. This is consistent with the results in which masker intrusion error alone accounts for almost all of the errors in this condition. Applying simple CASA feature extraction or speech models fails to truly reflect what was is heard since the mathematics of either technique fight against source separation when both sources share similar parametric attributes.

Therefore, we argue that some process, opposite to that of aggregation, must be at work prior to formant estimation. This process must involve decomposing the combined signal following the auditory front end into two separate representations. To model this process, we drew from a different signal processing tradition that deals with sparse signal representations under noisy conditions. These approaches assume that the input signal of interest is truly sparse in the time-frequency domain (e.g., a transient click - dense in frequency but sparse in time, or a harmonic complex - sparse in frequency but dense in time) and utilize this sparseness property to enhance the quality of the extracted signal. From our modeling work, the importance of pitch as a stream attribute is clear. Accordingly, we have developed a process that models each voiced utterance as a harmonic complex, estimates the amplitudes of each harmonic, and uses this "de-cluttered" representation of each utterance to extract information about the speech formants.

Computational model

The block diagram in Figure 1.10 shows the general steps of the extraction process. The first block emulates the auditory front end with a bank of bandpass filters followed by some form of envelope extraction. From this front end, two additional sources of information are extracted - the fundamental frequency and level of each source. The sparse-harmonics block uses the

fundamental frequencies to separate the two sources. These separate sources are then processed by a vowel classifier that draws upon both the pitch and level information for each source.

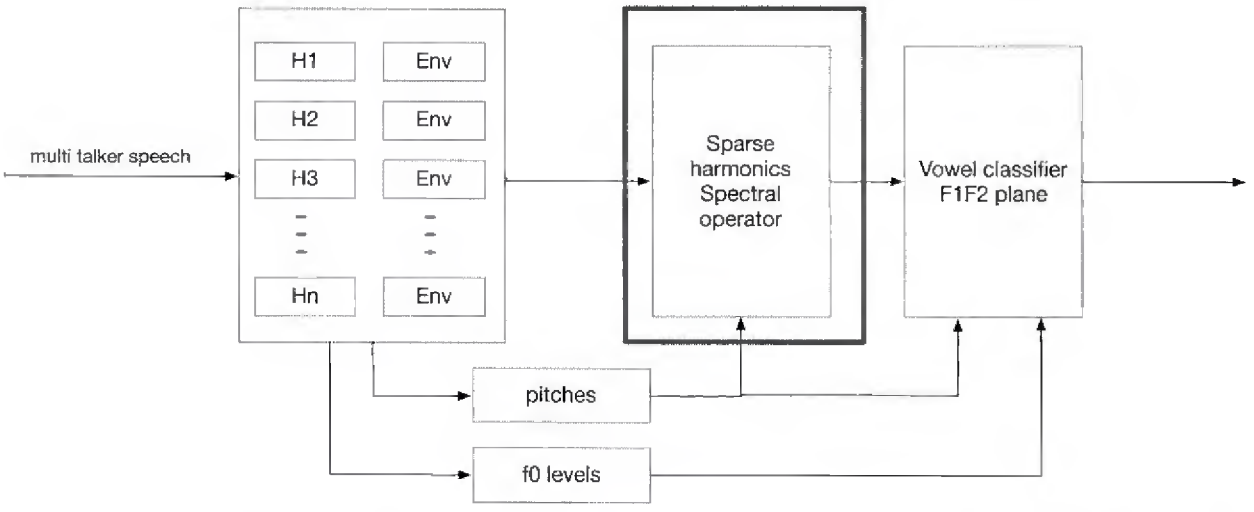

Fig. 1.10. A block-diagram of the source-separation process. The first block emulates the auditory front end with a bank of bandpass filters followed by some form of envelope extraction. From this front end, two additional sources of information are extracted - the fundamental frequency and level of each source. The sparse-harmonics block uses the fundamental frequencies to separate the two sources. These separate sources are then processed by a vowel classifier that draws upon both the pitch and level information for each source.

Implementation

In our implementation, we utilized 126 IIR Butterworth filters (80 Hz to 3 kHz) spaced logarithmically with Q10 values consistent with psychophysical tuning. The envelope detector consisted of half-wave rectification followed by a lowpass filter (60 Hz cutoff), which is approximates the temporal modulation transfer function (TMTF) reported by Viemeister (1979). Consistent with our previous approaches, we assume that pitch separation is handled through an "oracle" detector, and apply the Praat algorithm to each utterance alone. The spectral operator projects the output of the 126 envelopes onto the (much smaller dimensional - 10-20 dimensions) spaces defined by each separate harmonic model.

Qualitative evaluation

As demonstrated during the ONR program review on May 28th, the result of this process yields a perceptually compelling separation of the two talkers in the 0 dB TMR condition. Harmonic amplitudes were extracted every 10 msec and used to drive a time-varying harmonic synthesis algorithm. When compared with each original utterance, the primary degradation in quality is the absence of the unvoiced segments rather than distortions of the vowel stream. Indeed, when listening to synthesized versions of the processed *isolated* utterances, the separated sources sound similar.

Consistent with the limitations of the mathematics of sparse signals, as the TMR begins to favor one utterance over the other, the overall quality of the extracted softer utterance degrades.

Informal listening suggests that the source separation works well up to TMRs ranging from -6 to +6 dB. Beyond this range, one typically hears two voices and may actually hear the dominant color or digit spoken in the voice of the target!

Relative insensitivity to implementation choices

While we haven't exhaustively evaluated the differences, we have substituted gammatone filters for the Butterworth filters and a Hilbert envelope operator for the half-wave LP filter envelope detector without affecting the results. There are interesting peculiarities that appear to be related to degree of filter overlap, filter spacing, and the like, which will likely require fine-tuning should the gammatone option be used. Nevertheless, this design choice is not an important one if applied to the problem of modeling voiced-on-voiced speech interactions. Our own preference is to take advantage of the very clean phase alignments that are possible within Mathworks (*filtfilt*) using computationally efficient IIR filters.

Additionally, we deliberately chose not to utilize any of the neural-based auditory front ends, as these, we believe are overkill. Such models are appropriate for handling tasks involving large dynamic ranges, very fine discriminations, or the detection of very weak signals in noisy, cluttered environments. In the case of two-talker speech, the operating levels are sufficiently small to ignore level-dependent nonlinear tuning at the auditory periphery. Furthermore, the frame rate necessary for recovering the (slowly-varying) partial amplitudes is on the order of 10 msec, rather than less than 1 msec. Given that these auditory front ends require substantial computational power, we opted for a simple, off-the-shelf implementation instead to at least validate the approach.

We do believe that further research into the sparse-signal spectral operator may improve performance. Our technique has not been refined beyond the standard textbook implementation of pseudo-inverses and singular value decomposition. It is well known that more computationally intense methods (e.g., robust PCA using L1-norm minimization as opposed to SVD L2-norm minimization) are much better at eliminating leakage. We have explored several alternating projection techniques on the L2-norm solutions without substantial improvement, but we expect additional gains may be realized using L1-norm minimization and alternating convex projections.

Vowel classifier

Given our qualitative evaluations, the information extracted using our proposed source-separation algorithm appears to preserve the linguistic content of each source. With this in mind, we developed a vowel classifier to quantify such content. Because we would like the classifier to degrade gracefully as the louder source begins to leak into the weaker separated signal, we chose not to use standard spectral modeling of formants. Additionally, we wanted a classifier that would still work reasonably well when the quality of the original speech production degrades. Many of the utterances in the CRM corpus "violate" the intended phonemes of the "Ready <call sign> go to <color> <digit> now" script. There is considerable vowel neutralization, diphthongizing, elision, and voicing of nominally unvoiced consonants which challenges standard methods of quantifying.

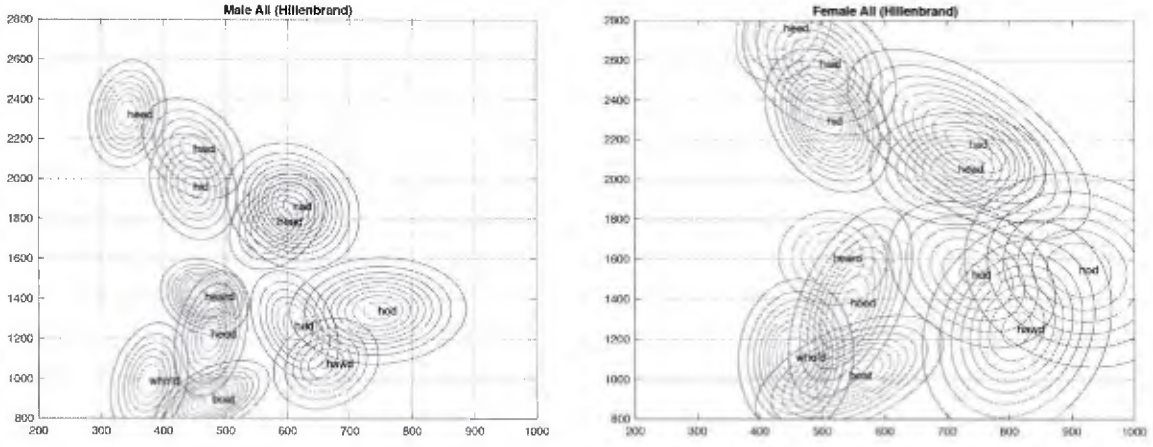

Fig 1.11. Bivariate Gaussian fits to Hillenbrand's data from midwest talkers for Male (left panel) and Female (right panel) talkers. F1 is represented in Hz along each abscissa and F2 is represented in Hz along each ordinate.

Our first step was to build a mathematical representation of the vowel space. We followed one of the standard approaches in the literature by using the online data provided by Hillenbrand (Hillenbrand, Getty, Clark, and Wheeler, 1995; Hillenbrand, 2019) to model the vowel space as a set of bivariate Gaussian distributions on the F1F2 plane. To capture the more general quality of the running speech found in the CRM corpus, we used Hillenbrand's scoring of F1 and F2 over the onset, steady-state and offset segments of each vowel, rather than the "cleaner" steady-state values. The result of this analysis are the probability distribution functions shown in Figure 1.11. What is noteworthy is that the vowel space is hardly sparse; it exhibits a substantial degree of overlap among classes. We subsequently use this property in creating our performance measures.

The next step was to represent the relevant spectral information from the sparse-signal spectral operator. To do this, we form the outer product of each recovered harmonic amplitudes. As a first approximation, we did not adjust for relative differences in level between the F1 and F2 regimes. An example of this is displayed in Figure 1.12 where the outer product is overlayed on top of the centroids of the vowel space.

Finally, a vector of vowel likelihoods is formed by weighting the outer-product matrix by each Gaussian distribution. In this calculation, we considered several choices of confidence regions and found that a bounding region of 80% worked well. Figure 1.13 displays these vectors as a time series (top panel) with the most likely option shown in the lower panel. Note that in some cases, the best-fitting vowel is a clear "winner", whereas for many other cases, there are several options within a few dB of each other.

Classifier performance as a function of TMR and word type

As a first pass at characterizing the source separation and classifier, we measured the percentage agreement between the vowel classification vectors for the extracted and isolated utterances and report these as probabilities. A soft threshold is used to accept all vowel IDs above

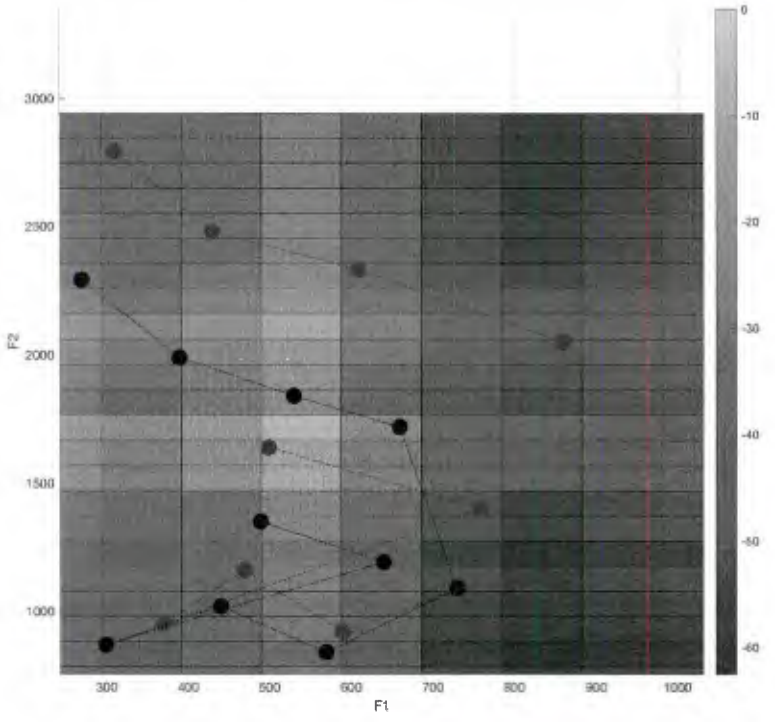

Figure 1.12. Outer product of a frame of harmonic amplitudes is overlayed with the centroids of the vowel space.The hot spots indicate the most likely location of the formants.

a particular value. For example, when the threshold is set to -5 dB, we gather the most likely vowel ID and all others within 5 dB of that weighted value for the two-talker observation and compare these to those extracted for each utterance in isolation. The results are shown in the panels of Figure 1.14 for three different acceptance thresholds: -5, -10 and -15 dB. Each figure displays 95% confidence intervals for the probability of agreement as a function of TMR and word type. Results are shown only for target agreement; masker agreement is the basically the mirror image of the that for the target.

The two most important findings are that performance degrades monotonically with TMR (as expected) and performance is independent of word type (as desired). Thus, for targets at +9 dB TMR, the percentage agreement between the vowel content of the separated target source and the isolated target source is 90% or better. As the TMR decreases, percentage agreement drops the most for the most conservative threshold (-5 dB), so that agreement is no better than approximately 55% for a TMR of -18 dB. In contrast, for the most liberal threshold (-15 dB), the agreement percentage is approximately 80%. The latter makes sense in that the vowel features associated with the target in the -18 dB TMR condition are likely to be more than -18 dB below those of the masker. What source separation is achieved in the -18 dB TMR case attenuates but doesn't eliminate the features associated with the louder masker.

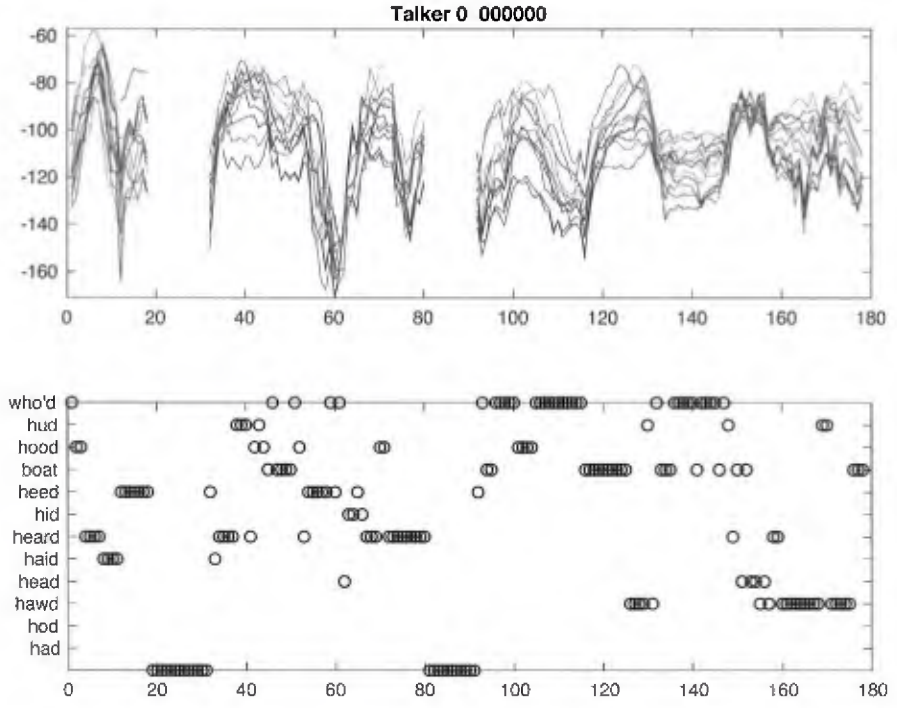

Figure 1.13. Time-series display of the weighted vowels over the time-course of one of the utterances from the CRM corpus (top panel). The bottom panel displays the best-fitting vowel.

With respect to word type, as we had hoped, whether the vowel is drawn from a call sign, color or digit word has no effect of the recovery of vowel content. These effects are somewhat vowel-dependent, as suggested by the range of the 95% confidence intervals. Nevertheless, with respect to the content-detection black box, we conclude that the voiced-on-voiced content detection function is independent of word type and the actual source of difference in the original content-detection functions reflects the relative degree of unvoiced segments in each word.

These results are based on the performance of the algorithm for the TT condition. Similar behavior is shown for the TS condition, while the trends for the TD condition follow the gender asymmetries noted above.

Level as an important source of side information

The above analysis of the performance results suggests that knowledge about the relative levels of the target and masker can be useful in identifying the voiced content of the softer source. For positive TMRs, a conservative threshold prevents the possibility of accepting the masker vowel identity along with that of the target. However, for TMRs well below 0 dB, setting a more liberal criterion for the target in conjunction with a conservative criterion for the masker allows one to discount that vowel information in the target that is likely due to the masker. This will be part of the work to be performed over the summer, and will be included in the final report at the end of funding.

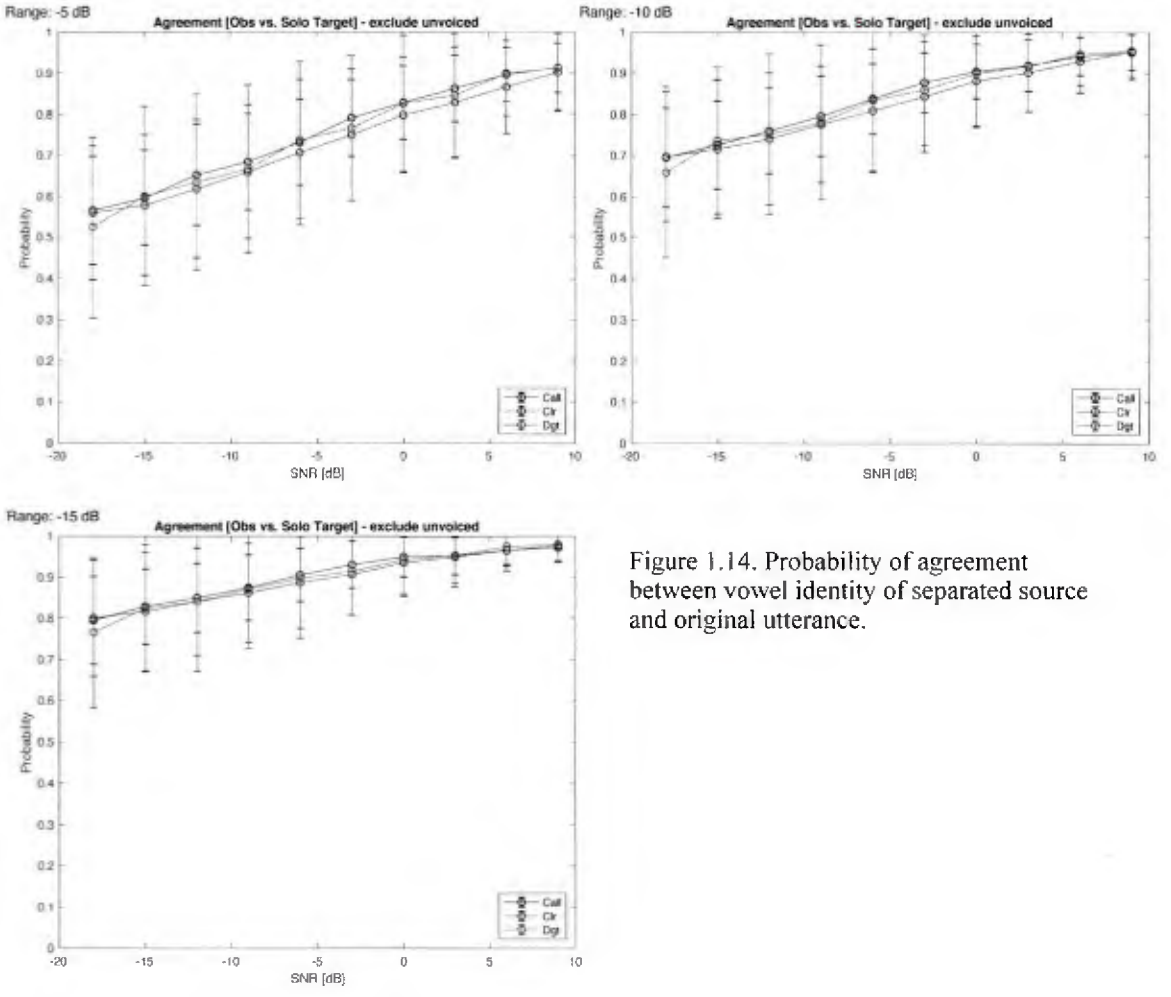

Figure 1.14. Probability of agreement between vowel identity of separated source and original utterance.

Speech-driven Classifier

Guided by the acoustically-driven modeling of the vowel classier above, our remaining project work adapted a more traditional pitch-formant representation of speech to account for the TMR, pitch, word-specific, utterance-specific, and gender effects in the two-talker CRM data, which we have extracted from the Thompson et al. (2015) data. Among the various speech toolboxes, we chose to work within the Praat software environment both for its open-source architecture and its reasonable support for scripting. The results bring us very close to the levels of performance of the Standard EPIC model, that was based entirely on aggregated "word" features.

Conclusions

Over the course of the current funding period, we have opened two of the critical black boxes of the EPIC auditory module. We have expanded the basic components of the auditory stream to include voiceless segments along with voiced ones. We have also developed a model based on the approach of source separation that emulates the general trends observed in the CRM task with respect to target-to-masker ratio.

COVID-19 delay. A detailed accounting of this work is being written separately in a technical report that summarizes the entire scope of the EPIC account of the CRM task. A final version of this technical report, however, is not available at the time of submitting the Final Report, due to the impact of COVID-19 on teaching and administrative responsibilities beginning in early March. As these activities will be reduced substantially at the end of the Winter semester, we will return to completing the technical report and will send the final document along to ONR. This technical report is being written to accommodate several journal papers for targeted audiences, e.g., cognitive scientists, psychoacousticians, and experimental psychologists.

Goal 2: Modeling the Effects of Spatial Location - Multi-talker Tasks

Background

When the"cocktail-party effect" was first described, the fact that the different talkers had different spatial locations would have been an obvious factor. Clearly, our ability to segregate the streams for individual talkers would be considerably better if they are spatially separated. Oddly enough, however, very few of the early studies of multichannel speech processing actually used spatially separated sources (see Yost, 1997). Most of the studies were done with *dichotic* presentation using headphones; the target message was provided to one ear, and the masker to the other. This results in a very strong stream segregation cue because the two ears each get a different signal, and low-level energetic masking effects between the two messages should be minimized because different inner-ear systems are involved.

However, dichotic presentation is not the same as *free field* presentation, in which the message sources are actually at different locations that are some distance from the listener, as would be the case with actual human talkers in a cocktail party, or if separate loudspeakers are used to present the messages. In the free field, each ear gets a version of a single sound that differs in both level and timing from the other ear - the stimuli are binaural- both ears are involved. Both this *interaural level difference* (ILD) and *interaural time difference* (ITD) are cues to location; the detailed effects are very complex and irregular due to how sounds of different frequencies interact with the shape of the human head, the outer ear, and ear canal to determine the ILD and ITD (Shaw, 1974). For example, at high frequencies, the head produces an *acoustic shadow* that attenuates the sound reaching the ear on the opposite side, but there is little effect at low frequencies, where the size of the head is much smaller than the wavelengths of the sounds. Since speech messages are broad-band, the ILD and ITD effects for speech will have a complex relationship to those for simple sounds. Consequently, if two speech messages from different locations are present, then the two ears get versions of both messages that differ in level and timing, producing a very complex signal for the auditory system to analyze and segregate. Studies often use synthetic localized sounds presented over headphones and modified by head-related transfer functions (HRTFs) both as a cheaper and more flexible way to provide localized sound, and to explore how localized sound could be used in real-world environments where headphones are practical, but a large loudspeaker array is not.

Like many other studies, the original Brungart (2001) study and the Thompson et al. (2015) replication simplify the multichannel speech situation by minimizing spatial cues with *diotic*

headphone presentation, in which the exact same signal is provided to both ears simultaneously, which produces an apparent location in the center of the head for both messages. This allows a detailed analysis of how the listener can segregate the streams using attributes such as loudness and pitch, which our current model can account for very well, while sharing the same perceived location (or lateral position) in space.

Complicating the analysis of whether spatial location can serve to help segregate streams is the fact that depending on the locations of the target and masker sources, the ILDs can produce the effect of a change in SNR - that is, depending on the specific locations, the target might be easier to perceive simply because it is louder than the masker at one or both ears. Various studies (see Arbogast, Mason, & Kidd, 2002) have shown that the effect is present, but when ILD contributions to SNR are properly controlled for, speech messages can in fact be segregated by a difference in actual perceived spatial location.

Status

The work accomplished during the 2016-17 funding period and reported in the 2017 Annual Report exhausted the available experimental results provided by our colleagues at AFRL. We did obtain some data from a more constrained version of the spatial study from Thompson at AFRL. However, after initial analysis, it was clear that we needed a better understanding of the content detection functions before proceeding to model the Thompson study as the experiment involved both two-talker and three-talker conditions. No further work has occurred since this time.

Goal 3. Extending the Visual Architecture

Introduction

Early in the development of EPIC's models for multi-task performance, Kieras and Meyer recognized that the visual information available at a point in time depends on the position of the eyes, and a critical determinant of multi-task performance is how quickly the eyes can be moved from one locus of task information to another (Meyer & Kieras, 1997; 1999). It thus became important to represent these performance constraints in the architecture, starting with an extremely simple retina model that divided the visual field into zones of fovea, parafovea, and periphery (Kieras & Meyer, 1995, 1997; Kieras, Meyer, & Ballas, 2001).

The purpose of extending the visual architecture is to arrive at components ready to use in building models that examine a display and act on what is seen, such as the typical radar console displays in Navy CICs. Visual search of displays of many icon-like objects, with labels or other coding, called *complex visual search* in what follows, is especially relevant to this application; however, most research in visual search has used much smaller and simpler displays, termed *simple visual search* in what follows.

Complex visual search. The main focus of the EPIC architecture work on visual search has been limited to tasks in which the number of objects in the display is very large and constant, and the objects are fairly uniform in density across the display. The search task consists of locating a target object that is always present and designating it with e.g. a mouse click. The response time (RT) and eye movements are the major measures. Eye movements are especially useful in understanding visual search, and the EPIC models for such tasks demonstrate the scope of the architecture (Kieras, 2009, 2010; Kieras & Hornof, 2014).

Simple visual search. The factors held constant in the available complex visual search studies work have obscured some important processes in visual search that have been extensively studied in simple search tasks. In these experiments, subjects view displays of relatively few objects, whose number is varied, and the subject must decide whether a specified target object is *present* or *absent*. The objects themselves are usually very simple, such as colored shapes, and the main measure involved is response time (RT), which is much less informative than eye movements, but is much simpler for both experimentation and modeling. During the previous project period, some selected simple search tasks were modeled which demonstrated the generality of the architecture but also identified a shortcoming of the visual architecture.

Goal of this work. The current models for complex visual search are accurate enough on some basic performance measures to be useful for simulating human performance in a design evaluation context or simulating humans in training simulations. (Kieras, 2009, 2010; Kieras & Hornof, 2014). The goal this project is to improve the models, so that they will be more scientifically valid and more generally applicable, and move them in the direction of a broader class of displays.

The specific question addressed in the project proposal was whether the current set of components was adequate to explain the details of complex visual search. If not, then the

question would be whether the shortcomings could only be addressed by moving down into lower-level details of vision, working directly with the actual images, and possibly adopting more complex concepts of how the visual scene is represented. The work conducted in this period on simple visual search shows some shortcomings in the original architecture that are fairly easily addressed, though they may not be critical in modeling complex visual search.

Basic Approach. The visual search modeling work with EPIC has relied on two important simplifications: First, the domain of images was limited to displays of icon-like objects that had relatively simple geometric shapes, text labels, and simple color schemes. This simplification includes the kinds of displays used in many military systems, and so is face-valid for many applications.

Second, the low-level visual processes were "black boxed" — the model was given as input a pre-parsed image, that is, a description of the image in the form of objects at locations with size, color, shape and other features. EPIC's eye model (described more below) then describes which of these properties is detected for each object in the visual field as a function of the current eye position and the eccentricity (distance from the point of fixation), and the size of the object. EPIC's cognitive strategy can then choose what object to fixate next based on the available information about the objects and the task goal. As discussed more below, a key simplifying assumption has been that these processes are performed on each object independently of the other objects.

In general, as shown by many studies, visual search behavior is guided in this way — for example, if one is searching for a red square, most fixations are made to red objects, but fixations on square objects are only somewhat more likely than chance. However, the guidance effect of different visual properties depends heavily on the specifics of the display, such as the size of objects, and their homogeneity with regard to properties such as color, and their density. Some of these issues are clear in the complex visual search work, and others are addressed in the work on simple visual search models.

Why model simple visual search?

During earlier reporting periods, David Kieras and David Meyer, a collaborator on the development of EPIC, continued work on a theory and methodology paper on the importance of characterizing the task strategy in order to interpret experimental results, and how this approach works best in terms of a cognitive architecture that distinguishes task-independent perceptual, cognitive, and motor mechanisms from task-dependent strategies. In the course of this work, the question arose whether EPIC's current mechanisms and the strategy used in the complex visual search models could be used to account for key effects in the vast literature on simple visual search where effects probably due to task strategy have been observed empirically, but not taken into account theoretically.

Simple visual search tasks are very heavily studied in the experimental literature, comprising almost all of the studies of visual search published since the early 1980s. In the dominant form of these tasks, a display of visual objects is presented and the subject must make a response whether a specified target object is present or absent. Typically the target is actually present on half of the

trials (*Positive* trials), and absent on the other half (*Negative* trials). The primary independent variables are the specific properties of the objects and the definition of the target (e.g. color is red, orientation is vertical), and the number of objects on the display, usually called "set size", typically ranging from 1 to 30. The primary dependent variable is the reaction time (RT) to make the present/absent response. Error rate (ER) is often reported, but rarely controlled with explicit incentives, nor considered theoretically . There is a common methodological assumption in RT experiments that if the ER is low (e.g. a few percent), it does not need to be considered as theoretically important. Furthermore, because the causes of erroneous responses are considered to be ill-defined, only the RT from correct trials is analyzed and theorized about.

Because of the simplicity of the methodology, a huge number of simple visual search studies have been conducted and published; the stereotypical result is that the RT is linear with the set size (number of objects). The focus is the slope of the RT as a function of set size, and how this slope varies with the perceptual properties of the display and the nature of the target specification. For example, if the target is a red circle, and the distractors are green circles, a common result is that the RT slope is close to zero, considered an "efficient" search; In contrast, if the target is hard to discriminate from the distractors, such as an T amid a field of Ls, the slope will be on the order of 30ms/item or more, which is considered "inefficient."

The simple visual search task differs in several ways from the complex search tasks that previous EPIC modeling has focussed on:

- The subject simply responds with key presses for present/absent rather than designating (e.g. by pointing and clicking on) a target object which is always present.
- The number of objects in the display is variable and relatively small, rather than a constant high number.
- The distribution of the objects on the display is not controlled strictly. The display area is held constant, and the objects to be used in a trial are placed at random locations in that display area. The local density, or proximity of the objects to each other, is not controlled, even though the average density is confounded with set size. In contrast, complex visual search displays are typically very dense with a fairly consistent spacing between objects.
- The objects are typically fairly large in subtended visual angle (e.g. 3°) in a relatively small display area (e.g. 20° square), rather than the smaller sizes and larger displays often studied in complex visual search.

Because of these differences, it is important to know whether the EPIC architecture and modeling approach would apply to these tasks as well as they have to the complex tasks. In addition, some of the important effects appearing in the simple visual search literature should be addressed because the architecture and models may have trouble addressing them. Two examples:

How do subjects decide to make an absent response? This is important because typically

trials in which the absent response is correct (*Negative* trials) show much stronger effects of set size than trials in which the present response is correct (*Positive* trials). This is consistent with the concept of a serial self-terminating search: if the subject examines each object in sequence to determine that the target is not present, then the effect of set size will be larger for Negative than for Positive trials that require examining only half of the objects on average. However, another possibility is that a "time-out" process is used: if the target is not located within a period of time that the subject has learned to expect during practice, the absent response is made (see Hulleman & Olivers, 2017 for review). This would also explain errors to some extent. How would absent responses be produced in an EPIC model?

What determines the slope of the RT function? As noted above, in some conditions, target-present responses are often fast and independent of set size, suggesting some kind of parallel scanning process. In other conditions, the slope is substantial, suggesting a serial scanning process. The simple visual search literature contains various theoretical ideas, starting with the Treisman & Gelade (1980) Feature Integration Theory, and the popular Guided Search Theory developed by Wolfe (Wolfe, Cave, & Franzel, 1989; Wolfe, 2014). Remarkably, these theories are not based on the characteristics of the human eye and oculomotor system, and based on remarkably weak evidence, these authors even claim that eye movements play no role at all. Rather, they assume that the basic mechanism underlying the RT effects is a serial covert "deployment of attention" to each object representation. The basic rationale for this covert attention model seems to be that the slope of the RT function is too small to be consistent with a model in which the eyes are moved to each object. However, as pointed out by Hulleman & Olivers (2017) in a recent review, there is long-standing evidence that the visual system can process more than one item at a time in a fixation, and thus it is the number of fixations, not the number of objects on the display, that produces the slope in the RT function.

This approach is consistent with the Active Vision concept promoted by Findlay & Gilchrist (2003) which is closely related to the EPIC architecture approach originally developed by Kieras & Meyer. The Active Vision approach says that time taken to perform a visual search task will be based on what can be seen in extra-foveal (peripheral) vision that can be used to guide the choice of the next fixation, the speed and accuracy of the oculomotor system in positioning the eye, memory systems involved in retaining the visual information over saccades, and importantly, the cognitive strategy for performing the task. This approach works well in accounting for complex visual search. Can EPIC models account for the key effects in the simple visual search literature?

Preview

In the earlier reporting periods, these questions were addressed with preliminary models of two important simple visual search studies: First, the classic Treisman and Gelade (1980) Experiment 1 was modeled with interesting and promising results, but the many design defects and poor methodological description in that paper made it unsuitable for more thorough modeling. The second preliminary modeling dealt with a relatively well-reported experiment by Wolfe, Palmer, and Horowitz (2010). In the previous reporting period, extensive work was done with the publicly available Wolfe, et al. data. The modeling work described in this report is an updated and more complete version of the work in the last annual report. The major progress was

refining and clarifying the visual crowding mechanism, and clarifying in some detail exactly how the models could account for the data. A key development is the use of "explanatory sequences" to show how the architectural and strategy mechanisms provided by EPIC are required for a satisfactory explanatory model; the concept is *abductive reasoning as inference to the best explanation* (Douven, 2017).

COVID-19 Delay. A large portion of the new work on simple visual search just summarized was performed after project termination. It was intended to complete this body of work in the form of a complete technical report that could be distributed to other researchers and used as the basis for a significant publication prior to submission of this Final Report. This would have enabled a short presentation here simply by reference to the technical report. In the interest of completing the formal requirements for project termination, the simple visual search work for Goal 3 presented below is essentially a preliminary version of the technical report, which hopefully will be finished soon and submitted to ONR, and revised to submit as an archival publication.

The Visual Search Experiment

The data used for this modeling was collected by Wolfe, Palmer, and Horowitz (2010), who made it available for download at http://search.bwh.harvard.edu/new/data_set_files.html. They focussed on whether the distributions of RTs provided evidence about the nature of the visual search process, rather than accounting quantitatively for the RT and error effects. This dataset is exceptional because of the relatively well-specified stimuli and very large number of trials from very well-practiced subjects. For completeness, the experimental method is re-stated here in the context of how the experiment was simulated in the model.

Method

Tasks. There were three different present/absent search tasks; Figure 3.1 shows a sample target-present display produced by the model for each task condition.

The three tasks are as follows; the acronyms labels are used here for greater clarity.

- **Color Single Feature (CSF)**: The display contains vertical green or red bars. The distractors are always green. The target is always red.
- **Color-Orientation Conjunction (COC)**: The display contains bars that are vertical or horizontal, and red or green. The distractors are always half green verticals and half red horizontals. The target is always a red vertical bar.
- **Shape Single Feature (SHP)**: The objects are "digital 5" and "digital 2" shapes similar to these digits on traditional seven-segment displays, made up of vertical and horizontal line segments. The distractors are always 5s; the target is always a 2.

Stimuli. Wolfe, et al. provide a good level of detail about the stimulus properties, but unfortunately, the download data set does not contain information about the actual display configuration used in each trial, so for purposes of modeling the display had to be generated for each simulated trial.

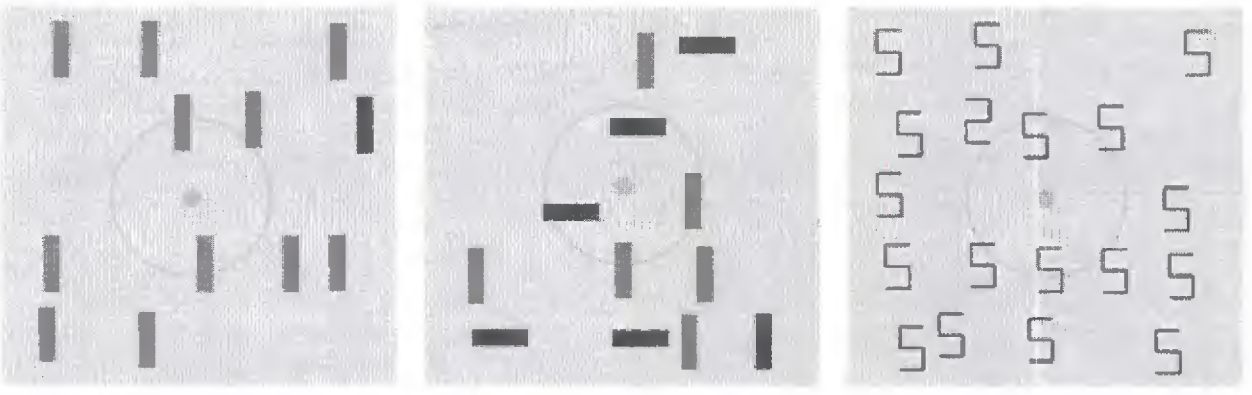

Figure 3.1. Example displays produced by the model for Positive trials in each task condition used in the Wolfe, et al. experiment. From left to right, Color Single Feature (CSF), Color-Orientation Conjunction (COC), and Shape Single Feature (SHP).

The search display was an area 22.5° × 22.5°, containing 25 invisible cells of 5° × 5°; Wolfe, et al. state that each object appeared in a random location within one of the cells, but did not state whether or how touching or overlapping objects were prevented. Assuming that such displays were not allowed, the random location within a cell was constrained to keep the horizontal or vertical edge of an object at least 0.25° away from the cell boundary, ensuring a minimum separation of 0.5° between edges of adjacent objects. Set sizes were 3, 6, 12, and 18. In the model, a display was generated for each trial as follows: the set size number of distractors were first placed in randomly chosen display cells. With probability of 0.5, the trial Polarity was then determined; if the trial was Positive (target present), a randomly chosen distractor was replaced with a target object.

In the CSF task, the objects were 1° × 3.5° vertical bars; the target bar was red, distractor bars were green. In the SHP task, the objects were 1.5° × 2.7° character-like shapes; the target was a *2* and the distractors were *5*s. In the COC task, the objects were 1° × 3.5° bars, red or green, oriented either horizontally or vertically. The target was a red vertical bar, and distractors were red horizontal and green vertical bars. Wolfe, et al. do not state exactly how COC distractors were chosen; in the model, half of the distractors were chosen to be of each type, with set size 3 special-cased so that at least one distractor of each type was present. Since a Positive trial display was produced by replacing a random distractor with a target, over trials, each type of distractor would appear equally often.

Design. There were 10 subjects in the COC task condition and 9 in the other two. One subject was in both COC and SHP, but the data set does not identify this subject, so the task condition was treated as a purely between-subject manipulation.

Procedure. Each trial began with a centered fixation cross. Subjects were instructed to "keep their eyes focussed on this cross" but because eye movements were not monitored, subjects could have moved their eyes, and based on other studies, it is likely that they did so. The search display was presented and remained visible until the subject pressed a key for target present or target absent. Subjects were instructed to respond "as quickly and accurately as possible." Correct/incorrect feedback was presented for 500 ms after each trial. Unlike many experiments,

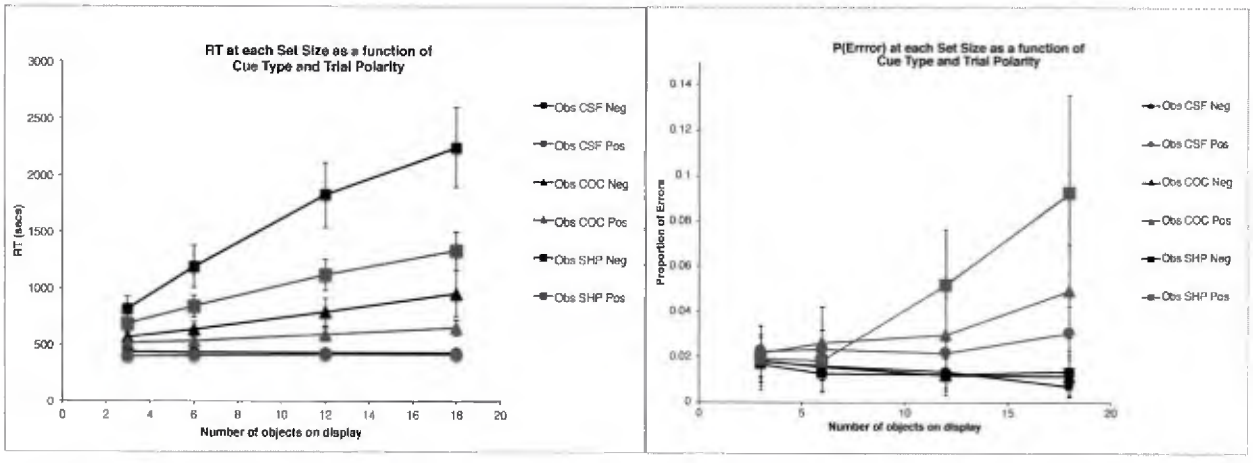

Figure 3.2. Observed RT (left panel) and ER (right panel) in each task condition. CSF: circles, COC: triangles, SHP: squares. Negative trials are plotted in black, Positive trials in red. The error bars are 95% confidence intervals based on the standard error of the individual subject means underlying each plotted point and thus reflect between-subject variability.

the subjects were very well practiced, with about 500 trials per subject for each combination of set size and Positive/Negative trial Polarity.

Results

The downloaded data consisted of the RT and correct/incorrect status for each subject for each trial at each set size and trial polarity. Following common practice in RT experiments, the data were reduced as follows: For each task condition, for each subject, the mean RT for correct trials and the proportion of errors (error rate, ER) for that subject was calculated for Positive and Negative trials at each set size, giving a total of 8 data points for RT and 8 data points for ER for each subject. These subject means were then averaged to produce the data points plotted in Figure 3.2. Through this report, Positive (target present) trials are shown as red points and lines, Negative (target absent) trials in black. CSF is plotted with circles, COC with triangles, and SHP with squares. The 95% confidence intervals around each data point were calculated by determining the standard error of that mean using the 9 or 10 individual subject means contributing to that point, thus reflecting between-subject variability, but not within-subject variability. Table 3.1 presents some summary statistics; given the great importance in the literature attached to the linearity of the RT functions, this table provides the intercept, slope, and r^2 of a linear fit to the mean RT data in each condition. Also shown is the mean ER and the maximum ER in each condition. Finally, the ratio of the Negative trial RT slope to the Positive trial RT slope is also shown. Since the CSF slopes are essentially zero, the slope ratio is meaningless in this condition. Note how the SHP slope ratio is roughly 2, the classic value for a self-terminating serial search, but the slope ratio is larger for the COC condition.

Table 3.1

	Negative				Positive					
Task Condition	Intercept	Slope	r^2	ER	Intercept	Slope	r^2	ER	ER Max	Slope ratio
CSF	436	-1	0.68	0.014	395	1	0.90	0.025	0.031	-0.69
COC	480	26	1.00	0.014	483	9	1.00	0.032	0.049	2.84
SHP	589	95	0.99	0.014	574	43	0.99	0.045	0.093	2.21

Wolfe, et al. did not report any overall statistical tests of these results. Therefore, unequal-*n* ANOVAs were performed using the **R ez** package on the reduced data. For RT, the main effects of Task Condition, Trial Polarity, Set Size, and all two- and three-way interactions were significant ($p < .05$). For ER, whose overall average was 2.4%, the Task Condition main effect was not significant ($p > .1$) but the Trial Polarity and Set Size main effects, and all two- and three-way interactions were significant ($p < .05$). For ER, an unequal-*n* ANOVA shows the Task Condition main effect was not significant ($p > .1$) but the Trial Polarity and Set Size main effects, and all two- and three-way interactions were significant ($p < .05$).

Examination of specific within-subject effects was done with Fisher Least Significant Difference values, which to avoid clutter are not shown on the graphs. For within-subject (within-condition) comparisons of the 24 mean values plotted in the graphs, the Fisher Least Significant Difference values are 68 for RT and 0.011 for ER. For RT, these values show that for CSF, of course the RTs are not different for either Trial Polarity or set size. For COC and SHP both the FLSD value and the between-subject confidence intervals indicate that the increasing trends for the RTs, and the tendency for Negative RT to be greater than Positive are reliable effects. For ER, these values indicate that the differences between the Negative trial ERs at set size 3 and 18 are significant. For Positive trial ER, for CSF, the set size 18 point is not quite reliably different from the smaller set size points; for COC, the ER for set size 12 is not reliably higher than for the smaller set sizes, but set size 18 ER is higher than all smaller set sizes ERs. Finally, for SHP Positive trials, ER for set size 12 is higher than the smaller set sizes ER, and set size 18 ER is higher than all smaller set sizes. Roughly speaking, for this means that most of the apparent within-condition effects in the graphs are reliably different even if the between-subject confidence intervals overlap.

Discussion

The RT results follow the classic pattern obtained in most visual search experiments. The RT functions are essentially flat in the CSF task (Positive trial slope is about 1 ms/item), this prominent effect with the color property in a single-feature search task is frequently described as "pop out". Otherwise, Positive and Negative trial RTs have a substantial slope, with the Negative trial slope roughly twice that of the Positive trials. Treisman & Gelade's (1980) Feature Integration Theory focussed on explaining how conjunctive searches had to be inefficient compared to single-feature searches. But Wolfe, Cave, and Franzel (1989) had already shown

that this central claim did not stand up to further testing: conjunctive searches can be very fast, and single-feature searches can be either very fast or very slow, depending on the specific visual properties involved. A fine point to note is that there is a hint of negative acceleration in the steepest RT function; much stronger curvilinear effects are commonly observed (see Wolfe et al., 1989) but have usually been ignored (but see Buetti et al., 2016).

Error rate (ER) effects are typically ignored in the simple visual search literature. In these results, the overall mean ER is only 2.4%, which most researchers would use to justify the conventional approach of ignoring the errors and focussing only on the RT for correct trials. Accordingly, Wolfe, et al. did not analyze or theorize about the ER results. But it is clear that the ER effects are highly systematic. In particular, there are very few errors on Negative trials (False Alarms), and their rate does not appear to depend at all on the task condition and very little on set size (apparently declining slightly with set size). In contrast, the errors on Positive trials (Misses) are more frequent and generally strongly increase with set size, and definitely depend on the task, being lowest in CSF and highest in SHP. The increase in both RT and ER with task difficulty rules out a simple overall speed-accuracy effect in the RT data, but because the ER effects are statistically reliable in spite of large individual differences, they deserve to be explained as well rather than ignored. Rather than postulate that ER depends on "task difficulty" and represent it by error rate parameters that increase with "difficulty," it would be better to explain this effect in terms of the visual and strategy mechanisms at work in the different tasks.

Mechanisms provided by EPIC and used in the Models

Visual Mechanisms

Figure 1 in the beginning introductory section of this document shows the overall structure of the EPIC architecture, but in simplified form; each of the processors in that diagram is fairly complex. Figure 3.3 shows the detailed architecture of the visual system. The *physical store* represents the current visual environment, e.g. what is on the screen of a display. Changes in the state of the physical visual environment are sent to the *eye processor*, which represents the retinal system and how the visual properties of objects in the physical store are differentially available depending on their physical properties, such as color and size, and their *eccentricity* — the distance in degrees of visual angle from the center of the fovea (Findlay & Gilchrist, 2003). The resulting "filtered" information is sent to the *sensory store*, where it persists for a fairly short time and comprises the input to the *perceptual processor*, which performs the processes of recognition and encoding. The output of the perceptual processor is stored in the *perceptual store*, which is the *visual working memory* and whose contents make up the visual modality-specific partition of the production system working memory. Thus, production rules can test visual information by condition items that match the current contents of the perceptual store. The visual stores are slaved to the visual environment as filtered by the current eye position. If the eye moves or the physical objects appear, disappear, change location, color, or size, the visual perceptual store will be updated to reflect the current visual scene. Because production rules can test the visual perceptual store contents, they can respond to this constantly updated representation of the current visual environment.

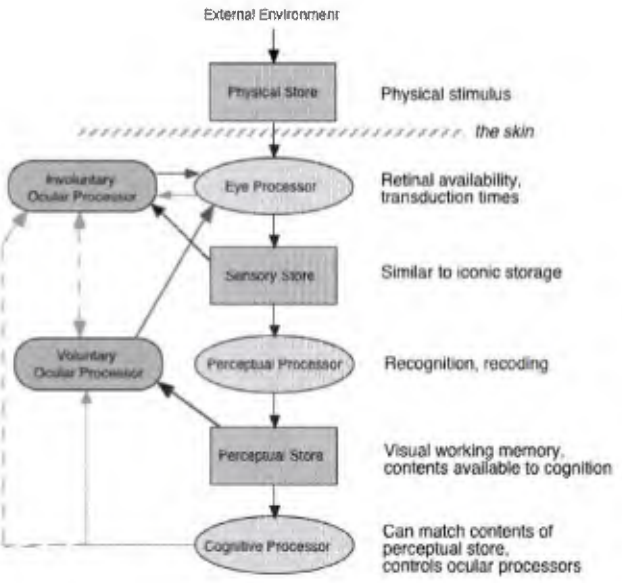

Figure 3.3. The detailed structure of the visual architecture in EPIC.

The appearance or disappearance of an object, or changes to its properties (such as its location), will be quickly updated in visual perceptual store, but if the information is no longer supported by visual input due to eye movements away from the object, the information persists for some time, on the order of seconds (see Henderson & Castelhano, 2005). In this way the visual perceptual store integrates over eye movements and maintains a cohesive representation of the current visual situation—corresponding to our subjective experience of a continuously present and integral visual surround.

The key property of EPIC's visual system is that there is no built-in limit to the number of objects or their properties that can be held in the perceptual visual store. However, the position of the eye and the properties of the early-vision system determine which objects are actually present in the visual perceptual store and the amount of information about them.

Retinal availability functions. A common hidden assumption about vision in cognitive psychology seems to be that only foveal vision needs to be considered (cf. Findlay & Gilchrist, 2003). However, classic psychophysical measurements make it clear that considerable information is available outside the fovea, and even into peripheral vision. More specifically, what can be perceived of an object depends not only on the eccentricity of the object but also on its size. Anstis (1974) provides some example measurements and comparisons that show that a single letter can be identified in the periphery if it is large enough. For example, the threshold size of a single letter was only about 0.2° at eccentricity of 5°, and about 1.3° at about 30° eccentricity. Moreover, different visual properties are differentially available in peripheral vision. A long-proposed neural mechanism for this relationship between eccentricity and size is *cortical magnification*: a constant amount of visual cortex (presumably supporting a certain number of receptive fields) is required for performing discrimination at a certain level, and since anatomically, the density of cortical representation declines with distance from the fovea, the size

of the stimulus must increase with eccentricity to involve the same amount of cortex. Such cortical magnification functions have been measured in psychophysical experiments; a typical result (e.g. Virsu & Rovamo, 1979) is that to maintain discriminability, the required size increases linearly up to a moderate eccentricity (e.g. 30° in Anstis, 1974) and then quite sharply in the further periphery, with a cubic function providing a good fit (cf Kieras & Hornof, 2014). Visual search studies such as Carrasco & Frieder (1996) using short exposure duration show that if object size is constant, then targets at greater eccentricity are located more slowly, but if peripheral objects are magnified in size according to the empirical magnification functions, search time becomes flat with eccentricity.

Unfortunately, the available psychophysical literature does not use a standard set of stimulus properties, so it is impossible to to combine empirical results into a well-parameterized set of functions for describing the detectability of object properties as a function of size and eccentricity. Thus EPIC's retinal processor uses availability functions of a certain form based on the psychophysical literature, and the parameters of the functions have to be estimated to fit the data being modeled. However, the psychophysical results do set some constraints — for example, the color of an object can be expected to be much more available than its shape.

In the models reported here, the maximum eccentricities are less than 30°, which suggests that a simple linear relationship between threshold size and eccentricity can used. The availability function is thus a Gaussian detection function that gives the probability that a specific property will be available (or detected) for an object with size s at eccentricity e:

$$P(detection) = P(s > N(\mu, \sigma)),\ \mu = \theta e,\ \sigma = 0.5$$

The value μ, the mean of the Gaussian function, can be interpreted as the 50% threshold for object size. It depends linearly on the eccentricity proportional to θ which is the *availability threshold coefficient*. Thus small values of θ correspond to the property being more available (more eccentricity possible for a given size), and large values of θ mean that the object is less available (e.g. must be closer for a given size). The standard deviation σ determines the "steepness" of the detection function for a given value of μ, and was held constant at 0.5. Thus only the θ parameter was adjusted to fit the observed data.

The value of θ differs depending on the visual property involved, and as noted above, will have to be estimated for the stimulus displays used in a particular experiment. Furthermore, for simplicity, the specific values for each property are assumed to have the same availability; e.g. a Color property value of Red is assumed to have the same θ as a Green value. Figure 3.5 shows the availability functions for some representative values for θ used in models for the Wolfe et al. (2010) experiment presented below, and illustrated in Figure 3.1; namely $\theta_C = 0.1$ for Color, $\theta_O = 0.2$ for Orientation, and $\theta_S = 0.5$ for Shape. Figure 3.5 also shows a couple of useful eccentricity metrics relevant to these models. The average initial eccentricity corresponds to the eyes being on the initial fixation point when the display appears. The average pairwise eccentricity corresponds to the average distance between display objects, corresponding to what can be seen

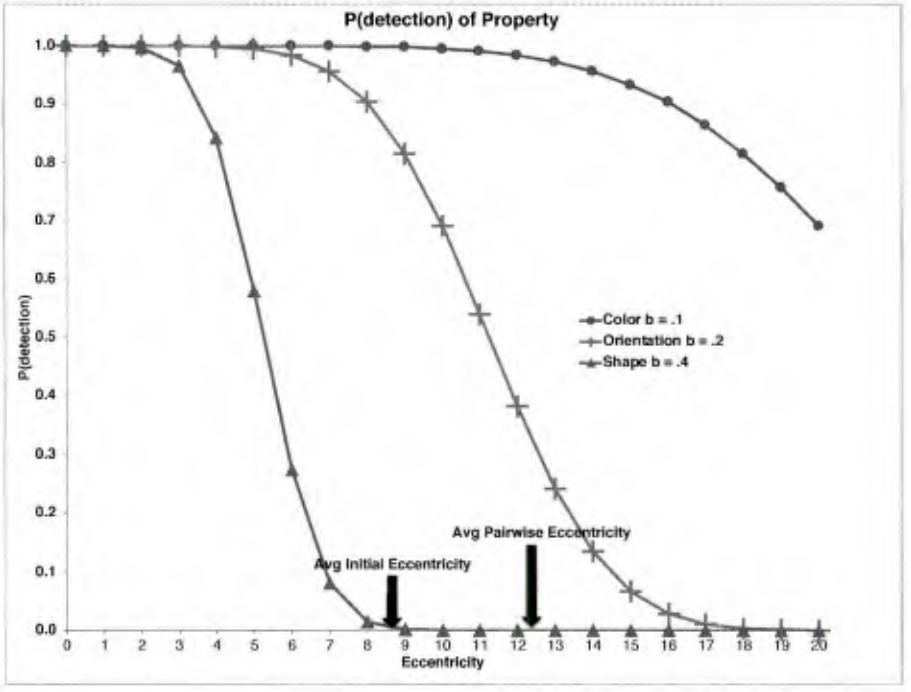

Figure 3.5. Example model availability functions for the different object properties used in the simple visual search task.

when the eyes are fixated on one of the objects. Color is very available; its detection probability is high throughout the eccentricity range. Orientation is significantly less available; at the average pairwise eccentricity, the probability of detection is only about 0.3. Finally, Shape is not very available at all; even at the average initial eccentricity, the probability of detection is almost zero. Thus very close fixations within a few degrees will be required to detect the Shape.

No bottleneck on availability for multiple objects. In this model of visual availability, the properties of more than one object can be available for a given eye location. No "attentional limit" is assumed to operate at this level. If the availability is high enough (a low θ threshold), then the properties of multiple objects, even at large eccentricities, will be available.

The availability for each property is independently resampled for all objects whenever the eye is moved. As the eye moves around, the available properties of a particular object can fluctuate, and will not be reliably available from one fixation to the next. However, the properties, once acquired, will remain for some time in the perceptual store.

This property of EPICs visual system is similar to the *area of conspicuity* (Engel, 1977) or *functional viewing field* (FVF, see reviews Findlay & Gilchrist, 2003; Hulleman & Olivers, 2017) in that the FVF also allow more than one object to be processed in a single fixation. However, the FVF concept assumes that there might be other factors that govern the number of objects covered by a fixation, including possible attentional limitations, or purely visual factors, such as *crowding effects*, to be discussed next.

Visual crowding effects. *Crowding*, also known as the *flanker effect,* refers to the phenomenon in which the perception of an object is impaired if it is surrounded by *flanking* objects that are spaced closely enough, but the same object is perceived accurately if the spacing is larger or there are no flanking objects (for reviews, see e.g. Levi, 2008; Pelli & Tillman, 2008; Rosenholtz, 2016). Crowding was first described for the recognition of characters in reading by Bouma (1970), and Anstis (1974) provides examples for letter recognition. Pelli & Tillman (2008) provide many other demonstrations of the effect. Crowding effects appear if the center-to-center spacing is less than the *critical spacing*, which for a wide variety of visual properties turns out to be approximately half the eccentricity of the object in question, a relationship first reported by Bouma (1970). Thus if a set of objects was being viewed at a large eccentricity, crowding could impair perception of them unless their spacing exceeds the critical spacing, but by moving the point of fixation closer, the critical spacing becomes smaller, and the crowded objects could be perceived correctly.

The reason why crowding might be relevant to simple visual search tasks is that *simple visual search experiments almost always confound the number of objects on the display with object spacing.* The usual experiment, e.g. Treisman & Gelade (1980), Wolfe, Cave, Franzel (1989), Wolfe et al. (2010) varies the set size while keeping the display area constant, placing the objects at random within the display area, and thereby producing higher object density at higher set sizes. The few studies attempting to separate crowding and set size effects suggest that at least most of the reported set size effects in visual search could in fact be due to crowding rather than simple numerosity of the objects (e.g. Motter & Simoni, 2008; Wertheim, Hooge, Krikke, & Johnson, 2006).

Rosenholtz (2016) argues that taking crowding into account is essential to understanding extra-foveal vision because it appears to be more responsible for the limitations on peripheral vision than simple loss of resolution. However, the specifics of crowding effects and their mechanisms remain unclear despite the extensive literature. There is a consensus that the visual system attempts to form visual objects by integrating information over integration fields, retinal areas whose size increases with eccentricity (cf the cortical magnification concept). If more than one physical object occupies a single such integration field, the integration process will be disrupted in some way. But if the point of fixation is closer, the smaller size of the integration fields will allow the same visual objects to be correctly formed. The problem is that the empirical work has not clarified, even in simple situations, the basic rules for the integration process and the results of crowding disruption and how it relates to the visual availability of the relevant properties.

To be more specific, many experiments measure the *amplitude* of the crowding effects in terms of an elevation of detection or discrimination threshold of the relevant property of a crowded target object as a function of crowder spacing. Different properties have differing threshold elevations; the available literature suggests that color has a low crowding amplitude and is thus relatively resistant to crowding effects, while the detailed shape of characters has a high crowding amplitude. In addition, the greater the similarity of flankers and targets, the larger the crowding amplitude. Other experiments simply ask the subject to identify the property (e.g.

color) of a central target object surrounded flankers, and vary the eccentricity or spacing, and report differences in accuracy.

What is unclear is what these experiments are actually measuring — if performance decreases in the presence of crowding, is it due a simple loss of availability of the target object property (i.e. an increased θ), or is it that a flanker property tends to be confused with the target property? This second possibility is a popular hypothesis in the literature: In the presence of crowding, the existence of the crowded object is still detected, and its basic perceptual features also are detected (θ is unchanged), but the disrupted integration process associates those features with the wrong object, such as a flanking object, and vice-versa — the features are essentially *scrambled* between the objects that crowd each other. Such scrambling seems apparent in demonstration examples of crowding (e.g. Pelli & Tillman, 2008), and might account for degraded perception of the target object in detection or discrimination tasks. Clear evidence is found in the relatively few studies that examine the characteristic of the errors when the actual target property has to be reported (e.g. Põder & Wagemans, 2007). In fact, more recent work (Yashar, Xiuyun, Jiageng, & Carrasco, 2019) shows that misreporting a flanker feature as a target feature accounts for most of the errors. Thus the feature scrambling concept was adopted as the crowding mechanism in this work.

A simple crowding mechanism. The visual architecture determines the properties of the visual objects in two steps that applied whenever the visual situation changes, that is, when the display appears or the eyes are moved. The visual architecture first applies the availability functions to determine which properties are available for each object from the current eye position. Once all the available property values been determined for all of the objects, a crowding mechanism is applied that that randomly scrambles the values of each property between objects that are within the critical spacing of each other, including any blank property values. To parameterize the magnitude of the crowding effect, scrambling for each property type and each object is performed with a certain *crowding probability,* ϕ.

The algorithm can be summarized as follows: First, using the availability functions, the available perceptual property values of an object are determined for the current eccentricity. If a property is unavailable for an object, the property is assigned a *blank* value. Second, each property of each object is examined, and with probability ϕ, the property values of the objects that crowd each other are randomly scrambled. The value of ϕ can differ depending on the property involved (but not the values of that property). Typical estimated values are $\phi_C = 0.025$ for Color and also $\phi_O = .025$ for Orientation, consistent with the dissimilarity of their two values, and $\phi_S = 0.1$ for Shape.

More specifically, in the actual visual system, it is reasonable to suppose the scrambling will happen simultaneously in some way for all of the objects, but the simple scrambling algorithm works sequentially as follows: It first determines all of the groups of objects that crowd each other. These are the *crowding groups*, one group for each object. The crowding relationship can be asymmetric: Suppose object A has a small eccentricity and object B has a larger eccentricity.

Object A might not have B in its crowding group, but object B might have A in its crowding group. Then, for each property, in order of increasing eccentricity, a "coin flip" is performed for each object, and with that probability ϕ the property values of all of the objects in that crowding group are collected (including blank property values), randomly shuffled, and then assigned back to the objects in the crowding group. Thus the actual number of available and unavailable property values are not changed; rather those values are scrambled for the objects in the crowding group. However, if an object has no crowders (i.e. it is closely fixated or relatively far from other objects), and its properties are available, these properties then become "sticky" in the visual perceptual store, and will not be lost or replaced by a blank property, but could still be scrambled in the future with available properties of crowding objects.

Note that the crowding probability ϕ applies at the level of each object. For example, a crowding group of four objects will have the crowding probability "coin flip" and scrambling applied a total of four times, once for each of the objects. Thus the probability of at least one scrambling operation being applied increases with the number of objects in the crowding group.

Illusory distractors, blanks, and targets. As stated above, the availability model and then the crowding model is applied when objects first appear, and then again after each eye movement. During the course of a visual search trial, as the eyes are moved around, the sticky perceptual store properties accumulate, and more objects acquire properties, either from becoming available due to nearby fixations, or from scrambling from nearby objects. During this process, the property value for an object might get replaced by some other object's property value, One case is that a distractor object might get a target property and thus become an *illusory target*. A more subtle case is that the target object might get a non-target value, becoming an *illusory distractor,* or the target object might get a blank property, even though its actual property value would be available, becoming an *illusory blank*. Illusory distractors are more likely than illusory blanks, because if the point of fixation is close enough for the target property to be available, there is a good chance that a nearby distractor object also has its property available for swapping.

Motor Mechanisms

Oculomotor accuracy and movement time. Production rules can send commands to the *involuntary* and *voluntary ocular processors* that control the position of the eyes. The voluntary ocular processor is directly controlled by the cognitive processor. A production rule action can command an eye movement to a designated object whose representation is in the perceptual store. The involuntary ocular processor generates "automatic" eye movements, such as reflex saccades to a new or moving object or smooth movements to keep the eye foveated on a slowly moving object. This processor is not involved in the models presented here and so will not be further described. More detail can be found in Meyer and Kieras (1997*a*), Kieras and Meyer (1997), and Kieras (2004).

The voluntary oculomotor processor includes motor "noise" that affects saccade accuracy. A variety of studies have shown that saccades tend to fall short of the actual fixation target, and the standard deviation of the saccade length tends to be proportional to the length (see Abrams, Meyer, & Kornblum, 1989, and the review in Harris, 1995). Thus, the oculomotor processor

samples the length for a saccade to an object at eccentricity e from a Gaussian distribution,

$$saccade\ length = N(\mu, \sigma)),\ \mu = g \cdot e,\ \sigma = s \cdot \mu$$

Typical empirical values for g (*gain*) range from 0.85 - 0.95, and s (*spread*) is typically around 10%. Harris (1995) did some modeling work that showed that given the variability in saccade length, and the resulting need to make multiple saccades to ensure fixation on an object, optimum total eye movement times to a target were obtained with $g = 0.95$, $s = 10\%$, which are consistent with observed values of these parameters. These values are used in EPIC as the "standard" default values for oculomotor noise along the line of flight of the saccade. Angular error might also be present; a saccade might not only fall short, but it might also miss to one side. Unfortunately, there are very few studies on angular error; a simplified model inspired by van Opstal and Gisbergen (1989) samples the saccade polar angle θ from a Gaussian distribution whose mean is the actual angle and whose standard deviation is a constant, current defaulting to simply 1°.

The time duration of a saccade was determined using the classic linear function described by Carpenter (1988):

$$saccade\ duration(ms) = 21 + 2.2 \cdot saccade\ length(degrees\ of\ visual\ angle)$$

The fixed intercept in this function was taken to represent the time to initiate the saccade; therefore the movement initiation time parameter in the original EPIC oculomotor processor was set to zero.

Manual Motor Mechanisms

The time to make a keypress response was represented by an architectural mechanism first proposed in EPIC for use in models of the psychological refectory period task (Meyer & Kieras, 1997). In these high-speed tasks, the subject has a finger poised over each response key so only a rapid finger flexion is required. The manual motor processor made use of a motor programming concept that producing a movement of this type requires selecting motor features that specify the hand and finger, then initiating the actual movement, which takes a certain amount of time to close the key switch (see Kieras, 2009, for additional discussion of this concept). Each feature was estimated as requiring 50 ms to select; the initiation time was also estimated at 50 ms, and the movement time at 25 ms. The motor features can be reused if identical, making the next movement faster.

Wolfe et al. do not provide any details about the response keys actually used, making it necessary to assume the parameters. Since present and absent responses were approximately equally probable, it is reasonable to suppose that the motor time on the average for Present and Absent responses would be the same in terms of the average number of feature reused. Thus the time contributed by the manual motor processor was set at 125 ms for both Present and Absent responses in all task conditions and set sizes.

Cognitive Task Strategy Mechanism

Task strategies. The models described below are implemented in terms of functions with

parameters that govern the "black box" visual and motor processors in the architecture. However, the activity of the cognitive processor is described in terms of production rules that specify a strategy for doing the task. In a visual search model, these rules will examine the contents of the visual perceptual store and make decisions, such as choosing an action such as moving the eyes or making a keypress response. A visual search strategy is necessarily specific to the task of visual search, but the basic features of strategies for simple visual search can be described in a general way; these are based on the previous EPIC models for more complex visual search.

EPIC's cognitive processor applies production rules in parallel in a 50 ms cycle. The total number of cycles required to produce a response contributes to the time to complete the task. The ability of the rules to apply in parallel is an important feature; it allows EPIC models to be developed for high-performance immediate tasks such as visual search and multitasking (c.f. Meyer & Kieras, 1997, 1999; Kieras, Meyer, Ballas, & Lauber, 2000).

To be effective in doing a task, the strategy must take perceptual and motor capabilities and limitations into account in a way that maximizes task performance. A fundamental assumption of EPIC and similar production-rule cognitive architectures is that subjects in experiments create a set of production rules when first instructed in the task, and then refine those rules as they gain experience in the task; clearly, feedback or incentives will play a role in how the strategy is refined. EPIC does not attempt to represent the underlying learning mechanisms, but can represent a strategy assumed to be operative after adequate amounts of practice. In the models described here, it is assumed that the amount of practice is extensive enough that subjects have developed a stable strategy that can be chosen so as to fit the data within the constraints of the mechanisms that EPIC provides.

A strategy can be easily summarized in pseudo-code, making it unnecessary to provide the technical details of the production rule syntax and the specific production rules used in the models. In this section, the task strategies used in the models will be described.

The production rules in the visual search models presented here are a variation of a basic strategy used in previous EPIC visual search models; this *Basic Search* strategy is shown as pseudocode in Box 3.1. Once the display objects appear on the screen, after a delay time held constant at 100 ms in these models, the strategy production rules alternate between a Step 3 *nomination phase*, in which *nomination rules* nominate objects (including those in peripheral vision) that are either *the target*, or *possible targets* because a relevant target property matches or is unknown, and a Step 4 *choice phase*, in which rules fire to take an action. If a target object has

1. Start the trial with the eyes on the fixation point and wait for object display to appear.
2. Delay for some time.
3. Nominate possible target objects using their available properties.
4. Choose an action:
a. If a nominee matches the target, respond present, trial is done.
b. If there are no nominees, respond absent, trial is done.
c. Otherwise, choose the closest candidate object from the nominees, and move the eyes to it, and go to step 3.

Box 3.1. Pseudocode for the Basic Search strategy.

been nominated, a target-present response is immediately made via a manual motor processor keystroke command. If there are no nominations, meaning that all objects appear (even in peripheral vision) to be distractors, then a target-absent response is made. Otherwise, there are only possible-target nominations, so oculomotor processor eye movement command is issued to move the eyes to the *closest* nominated object. Once the eye movement is complete, the nomination phase starts again. The implementation is such that each numbered step in Box 3.1 corresponds to a single production rule cycle, and thus the strategy is implemented in the minimum number of production rule cycles.

Basic Search as an optimal strategy. The Basic Search strategy is essentially the "fastest reasonable" way to perform the task because it takes extra-foveal vision into account: it may not be necessary to fixate each object to know whether it is the target or a distractor. Thus a target-present response is produced as soon as either a target is detected, even if it is not fixated, and a target-absent response is produced as soon as all objects appear to be distractors, regardless of whether they have all been fixated. This strategy is "reasonable" as well as fast because the response choice will be accurate if the perceptual information and motor action are accurate. The Basic Search strategy is the core strategy for all of the models; there are some useful variations on it to be discussed next.

Thus, as fixations are made, information about the objects accumulates until either the target object becomes known, or the known properties of all objects show that none of them could be the target — all of the objects look like distractors. The major determinant of RT is how many eye movements are made in this process. Unlike many models which do some form of time-out for making target-absent responses (cf. review in Hulleman & Olivers, 2017), the Basic Search strategy states simply that if there are no possible-target nominations (everything looks like a distractor), then make a target-absent response.

Strategy variations. In general, the choice of strategy has a large effect on whether the model can fit the data, and a satisfactory fit can only be obtained by a combination of parameter values and strategy choice. As argued above, the Basic Search strategy is a "fastest reasonable" strategy that produces fast responses which will be accurate if the perceptual and motor systems are accurate. As will be shown, some variations on this strategy will be needed for good fits to the data.

The extremely simple *Fixed-Eye* strategy is shown in Box 3.2. The eyes are left on the central fixation point and the target-present or target-absent response is chosen after a single nomination phase. This strategy corresponds to the instructions often given to subjects in visual search experiments where the display remains present until the response, namely to keep the eyes

1. Start the trial with the eyes on the fixation point and wait for object display to appear.
2. Delay for some time.
3. Nominate possible target objects using their available properties.
4. Respond present if target nominated, respond absent if not, trial is done.

Box 3.2. Pseudocode for the Fixed-Eye visual search strategy.

on the fixation point. Whether subjects can or do follow this instruction and apply the Fixed-Eye strategy is another question.

The *Basis Search with Confirm-Positive* strategy, shown in Box 3.3, provides protection against erroneous target-present responses to illusory targets produced by crowding. The strategy confirms that an apparent target is an actual target by moving the eyes to it, which would mitigate any crowding, and respond present if the apparent target appears to be a target, or continues the search if not. This confirmation eye movement is skipped if the apparent target object is already fixated, defined as an eccentricity within 1°.

1. Start the trial and wait for object display to appear.
2. Delay for some time.
3. Nominate possible target objects using their available properties.
4. Choose an action:

a. If a nominee matches the target, and it is already fixated, respond present, trial is done. If not already fixated, confirm by moving the eyes to the nominee: If the object is the target, respond present, trial is done; otherwise, go to step 3.
b. If there are no nominees, respond absent, trial is done.
c. Otherwise, choose a candidate object from the nominees, and move the eyes to it, and go to step 3.

Box 3.3. Pseudocode for the Basic Search with Confirm-Positive strategy.

The *Limited-Fixations Strategy Option* shown in Box 3.4 provides a way to speed up the core Basic Search strategy, which simply moves the eyes as many times as necessary to produce a response is made. To respond faster, the strategy could "time out" and respond absent if some time has elapsed without finding the target, as often proposed as an error mechanism in other models (cf Hulleman & Olivers, 2017). Since Basic Search moves the eyes at a roughly constant pace, this "time out" option was implemented more simply as a limit, *NFix*, on the number of eye movements, including any Confirmation eye movements. This option is applied to the Basic Search strategies at each step prior to initiating an eye movement. Note that the initial fixation on the fixation point at the beginning of the trial is not counted. Thus the Fixed-Eye strategy can be considered as $NFix = 0$; the Basic Search strategy, either with or without Confirmation, corresponds to *NFix* set to an extremely large value (e.g. 99) that is never exceeded in the model fits.

Before making an eye movement, check:

a. If the number of fixations made thus far is greater than *NFix*, respond absent, trial is done.
b. Otherwise, proceed.

Box 3.4 Pseudocode for the Limited-Fixations Strategy option.

Applying the architectural mechanisms to model the experiment

Crowding effects

The reason why crowding might be relevant to simple visual search tasks is that these experiments almost always confound the number of objects on the display with object spacing. The usual experiment, e.g. Treisman & Gelade (1980), Wolfe, Cave, Franzel (1989), Wolfe et al. (2010) varies the set size while keeping the display area constant, placing the objects at random within the display area, and thereby producing higher object density at higher set sizes. The few studies attempting to separate crowding and set size effects suggest that at least most of the reported set size effects in visual search could in fact be due to crowding rather than simple numerosity of the objects (e.g. Motter & Simoni, 2008; Wertheim, Hooge, Krikke, & Johnson, 2006).

In contrast, Wolfe et al. (2010) simply asserted that the stimuli "can be easily identified outside of the fovea" in these tasks (p. 1305) but report no measurements about whether this is true in their higher-density displays; this seems unlikely at least in the SHP condition, and even the CSF condition produces more Miss Errors than False Alarms, which implies that there is some source of systematic difficulty in what would seem to be a trivial task.

To assess whether the Wolfe, et al. displays confound crowding with set size, two measures of crowding were computed for a large number of displays generated by the model. The first assuming that the eyes were at the initial fixation point, and the second a pairwise measure assuming that the eyes were fixated on each individual object in the display. The eccentricity from the assumed fixation point was calculated for every other object on the display; the average eccentricities for these two assumptions were shown in Figure 3.5 above. Then for every display object, the number of surrounding objects within the critical spacing (defined as half of the eccentricity) was counted. Figure 3.6 shows the average number of crowding objects for each

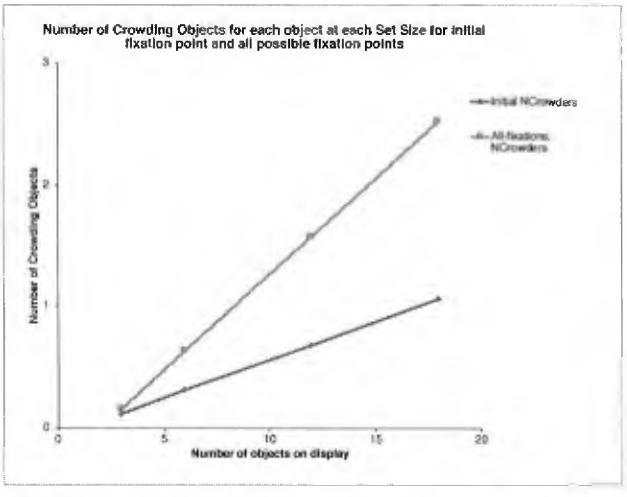

Figure 3.6. Crowding in the WPH displays. The upper curve shows the average number of objects that crowd another object assuming a fixation on each possible object. The bottom curve shows the average number of objects that crowd another object assuming the initial fixation point.

measure. When the eye was assumed to be at the initial fixation point, the mean eccentricity from the central initial fixation point was 8.8° and the number of crowding objects increased from an average of 0.1 at set size 3 up to 1.1 for set size 18. When the eyes were assumed to be on each object, the average eccentricity was 12.3°, and the number of crowding objects increased from .15 at set size 3 to 2.5 at set size 18. Thus it is clear that crowding effects are present in this simple visual search task — at small set sizes, almost no crowding is present, but substantial crowding appears at larger set sizes once the eyes are moved away from the initial fixation point. Simple visual search experiments have indeed confounded their basic manipulation with an unacknowledged, but powerful, factor in visual perception.

Accordingly, models of these tasks should incorporate crowding effects.

Visual Properties and the Basic Search Strategy

The above presentation of availability functions and the crowding algorithm implied a choice of the relevant visual properties for the Wolfe et al. tasks. It seems obvious that for the CSF and COC tasks, the relevant properties are the traditional Color and Orientation features, that if available, have one of two values. The nomination and choice rules are thus very simple for the CSF task because only a single object property is involved: An object is nominated as the target if it has a Red Color, or as a possible target to be fixated if it has an unknown Color. If all objects have a Green Color, then no objects can be a target, so there are no nominations, leading to an immediate absent response.

In contrast, for COC there are four possible nominations: a target nomination for a Red Vertical bar, and three possible nominations for possible target objects: Red Color & blank (unknown) Orientation, blank Color & Vertical Orientation, and blank Color & blank Orientation. If a target has been nominated, a present response is made; if one or more possible-target nominations are present, the strategy should choose one to fixate in the descending priority order as just listed. Finally if there are no nominations because all objects appear to be either Horizontal Red or Vertical Green, an immediate absent response is made.

What are the relevant properties for the SHP task? The description of the availability functions above imply that Shape is treated as a unitary property just like Color and Orientation, only less available. In fact this approach is justified by the Basic Search strategy; this can be demonstrated by assuming that an object's Shape property consists of subfeatures. For example, a common idea is that characters are made up of line segment subfeatures, each of which must be detected and then subject to crowding scrambling. So perhaps the partially available Shape for the "digital 2" and "digital 5" should consist of a subset of the seven possible line segments where each is either horizontal or vertical and in a particular location within the object, producing 2^7 possible shapes. Alternatively, perhaps the subfeatures are a leftward- or rightward-facing "digital c" each with a top or bottom location. Availability and crowding scrambling would produce a percept of a partial shape resembling a *c,* reversed *c*, a *3*, or *E*. Unfortunately the vision and crowding literature does not provide much guidance on the relevant properties of characters — defining the hypothetical features of visual objects has always been fairly speculative; any assumptions along these lines will arbitrary and lead to complexity in the model.

However, the Basic Search strategy justifies a great simplification: Since these partial and jumbled shapes match neither a target nor a distractor, the strategy arguably should treat them as *possible targets* to be fixated to determine what they actually are, just as if they had a "blank" shape. This means that the Shape property can be treated as a unitary property: each object has a Shape property with a value that is either *'2'*, *'5'*, or *blank*. Availability of Shape can be represented with a detection function whose threshold θ_S is higher than that putatively involved with detecting the hypothetical individual subfeatures. Crowding will scramble these property values for Shape according to the same algorithms as for Color and Orientation.

The SHP task is thus a single-feature task with nomination and choice rules for SHP that are as simple as those for CSF. If a *'2'* is visible, a target is nominated and a target-present response is made; if a blank is visible, it is a possible-target nomination to be considered for fixation. If all objects appear to have a *'5'* shape, then a target-absent response is made.

How the visual search models makes errors

As alluded to in the discussion of the Wolfe et al. results, there are clearly systematic effects in the ER data as well as the RT data that should be accounted for. A model that attempts to account for both RT and ER is rare in visual search work, and not common in other areas of human performance: it is common theoretical practice that if the ER is "low enough," then the RT results are taken to be trustworthy indicators of mental processes, and nothing more is said about the errors. Since this effort attempts to use both RT and ER as equal-status indicators of visual search processes, it is important to be clear on how the models make errors.

Errors have three sources under the strategies used in the models:

Action slips. First, note that the rational strategy will not "deliberately" respond Present on a Negative trial, which would be a False Alarm Error. The constancy of the False Alarm ER in the observed data suggests a basic error mechanism that is often postulated in human performance research. Namely, with some probability, a person will make an *action slip* — the intention is correct, but at random, an incorrect motor action will be triggered. Correspondingly, this most basic of error sources was incorporated into these models: When the strategy calls for a Present or Absent response, the *opposite* response is made with an "oops" error probability *OopsER*. This will produce both False Alarm and Miss Errors, but with a constant probability across search tasks, trial polarity, and set size.

Stopping too soon. The key feature of the Miss ER in the data is that it increases with set size and apparent task difficulty, so accounting for these effects is an important challenge to the model. Miss Errors could be produced by a strategy that limits the number of eye movements. The Eyes-Fixed strategy is an extreme in that the eyes are not moved at all, but the same argument applies if one or more fixations are allowed. Note that both Eyes-Fixed and the Basic Search strategy always undertake to respond present immediately if the target is visible. Thus if the trial is terminated with an absent response before the target has become available, a Miss Error would result. This is more likely to happen if there are more objects on the display, so the Miss ER would increase with set size.

Illusory distractors from crowding. Another source of Miss Errors is that the strategy rule that detects the absence of nominations fires when the target is in fact present on the display. This would happen if all of the objects appear to be distractors. This will be exactly the situation if crowding scrambling turned the target into an illusory distractor and at the same time, all of the other objects appear to be distractors.

Thus the simple scrambling model of crowding effects in combination with action slips and the logic of the Basic Search strategy predicts the asymmetry of False Alarm and Miss errors and the increase of Miss errors with set size, and search task apparent difficulty due to differences in availability of the relevant properties.

Generating and evaluating model predictions

Implementation details. The scrambling model of crowding requires that each visual object be processed in the context of the other objects in its crowding group. Since the current EPIC visual processors are based on the concept of processing each object independently, it would require restructuring the architecture code to implement the scrambling model; it is a better tactic to evaluate the prospects of the crowding model before undertaking a careful reprogramming of the EPIC code base. Accordingly, the specific processes used in implementing the simple visual search model in EPIC were reproduced in a stripped-down body of native C++ code. By setting the crowding probability to zero, it was possible to check that the C++ model produced identical RT predictions to the full EPIC version. This was interestingly non-trivial; because the EPIC software directly supports fully parallel processing between and within architectural components, programming an EPIC production-rule strategy is much easier than coding the corresponding behavior in plain C++ code.

Model fitting issues. The best-fitting strategy and parameter values were determined by informal iteration. Since the ER data can involve very small values (< 0.01), it was necessary to run a very large number of simulated trials to get stable model predictions for small ERs. This was achieved by using the highly simplified and efficient C++ "clone" of the actual full EPIC production rule model described above. Because of the very high speed of the clone compared to the full EPIC system (which is coded in C++ as well), it was convenient to obtain *one million* simulated trials for each combination of task condition, set size and trial polarity; all of the reported simulation results use this sample size.

Throughout this section, the goodness-of-fit of a model will be reported in terms of three metrics that between them, usefully describe the relationship between the predicted and observed values. The first metric is the common correlational measure used in modeling work, r^2, the proportion of variance in the observed values accounted by the predicted values; this basically measures the extent to which predicted and observed values parallel each other. However, it is misleading if the predicted and observed points show parallel trends but differ in magnitude, or when there is no observed trend, e.g., flat RTs, means there is no variance to account for even if the predicted and observed values are the same. A good measure of how well the values actually match is the average absolute relative error, *aare,* expressed as a percentage, between the predicted and observed values

$$aare = (\sum_i(|o_i - p_i| / o_i)) / n$$

Note how in this measure, under-predictions do not cancel out over-predictions. This gives an easily understood measure of how close in actual values the predictions are to the data, interpretable in terms of customary rules of thumb in engineering practice. However, in this data, some of the ER observed values are extremely small; their presence in the denominator of the relative error tends to "inflate" this metric, sometimes dramatically. This is suggests a third metric, the simple average absolute error, *aae*, between the predicted and observed values

$$aae = (\sum_i|o_i - p_i|) / n$$

This measure avoids the small-denominator problem but at the expense of being in the same units (ms or error proportion) as the data.

Note that there is an asymmetry in the reliability of the RT and ER observed data. There were about 500 trials per subject per task, set size, and polarity condition. For RT measurements, this sample size can produce relatively "tight" confidence intervals, even when, as shown in the graphs, they are computed between-subjects for 9 or 10 subjects. However, the variability of proportions, such as ER, is intrinsically higher, and does not constrain the underlying estimate of the population proportion very much. This shows up in the relatively large confidence intervals on the ER data in Figure 3.2 when computed in the same way as for the RT data. Thus while the extent to which the predicted values fall within the confidence intervals around the observed data points is an attractive indicator of goodness of fit, the large size of some of these can make it overly generous.

Throughout all of the model fits presented, preference was given to minimizing the number of different strategies invoked to fit the data, and trying to hold as many parameters at the same and simple values as possible. Since accounting for ER was a novel undertaking in visual search models, some preference was given to a close fit for ER at the expense of the RT fit.

The Concept of Explanatory Sequences

It is common in modeling work to simply show the final good-fitting model with the predicted and observed data, the corresponding parameter values, and the goodness-of-fit metrics, and declare victory. However, modeling work is very much more informative if the mechanisms in the model, the strategy, and the parameter values, can be shown to be actually necessary to obtain the good fit. This can be shown with an *explanatory sequence* of models, that shows the effects of adding mechanisms with their corresponding parameters. In cognitive architecture models, we start with a set of "hardware" mechanisms and their parameters and use them in conjunction with the "software" of task strategies. In EPIC, the "hardware" mechanisms are the given "black-boxed" and parameterized sensory-perceptual and motor systems, and the well-defined and limited production-system engine of the cognitive processor, while the "software" is the cognitive task strategy represented as the specific production rules applied by the cognitive processor to choose what motor actions to perform based on the perceptual data in order to meet the task requirements. In this approach, the "hardware" is assumed to be fixed, but the parameters have to be estimated, and the strategy, although it is constrained by the task requirements, is free to vary due to both procedural details of the experiment and the preferences of individual subjects. So an explanatory sequence of EPIC models will show the effects of changing both the parameters that govern the given architecture components, and the strategy used in the task. As pointed out above, parameter estimates and strategy choice can interact; however it is often very informative to hold one constant while varying the other. But the "EPIC philosophy" for explaining effects is to give first priority to including the known effects of the fixed perceptual and motor mechanisms with their estimated parameters, followed by varying the strategy as needed to match the data. If such a model suffices to account for the effects, there is no need to propose additional architectural components, such as cognitive mechanisms with relatively ill-defined properties such as covert attention or limited central capacity. In fact, it is the claim of this work that despite its traditional popularity, the cognitive mechanism of covert attention is not required to explain these visual search effects once known visual mechanisms, eye movements, motor errors, and the task strategy are represented in the model.

Explanatory Sequence for the SHP Task

The first explanatory sequences in this report start with accounting for the SHP task. The reasons for this choice are that of the three tasks in the Wolfe, et al. dataset, this task is the most prototypical visual search task — visual search definitely happens when detecting the presence or location of an object is non-trivial, requiring some kind of examination of the display for a few seconds, and there is a large increase in the time required as the number of objects increases. So the key features of this data to be explained are: RTs that steeply increase with Set Size, almost linearly, with Negative trial RT having a greater slope and slightly more curvilinear, than for Positive trial RT. The ER for Negative trials (False Alarms) is small and almost constant (as it is in all task conditions), while the ER for Positive trials (Misses) increases significantly with Set Size. The following presents the explanatory sequence to account for these features as a series of numbered steps.

1. RT slopes are due to Availability and the Basic Search strategy. We start with the basic

mechanisms of visual availability, visual-perceptual store retention, eye movement times, and the Basic Search strategy, and will attempt to fit the RT effects first. The model in this first step makes no errors, but ER will be addressed in next step. Given these mechanisms and the Basic Search strategy, as Set Size increases, object density increases, and more objects will be covered by each fixation. Thus availability will determine how many eye movements are required before the strategy chooses a response, which then determines the RT. Accordingly, Figure 3.7 shows the effects of increasing the threshold coefficient parameter θ_S from 0.0, where the Shape of all objects on the display are available from the initial fixation point, to 0.6 where very few will be available for any given fixation point. By bracketing the observed RTs, these fits show the effects of availability with the Basic Search strategy.

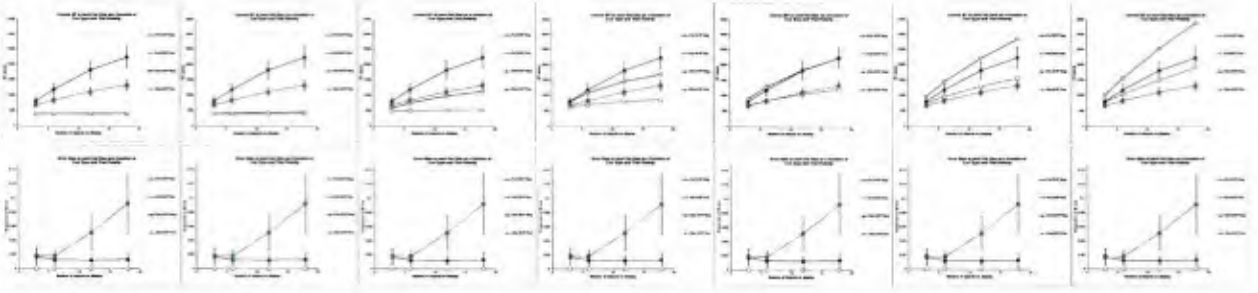

Figure 3.7. θ_S = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6
Slope ratios: 0/0, 594, 9.8, 3.6, 2.7, 2.3, 2.1
Fixations: {- 0 - 0, + 0 - 0} {- 0.0 - 0.3, + .0 - .0} {- 0.9 - 3.3, + 0.4 0.7} {- 1.9 - 6.5, + 1.1 - 2.3} {-2.4 - 9.1, + 1.5 - 4.0} {- 2.7 - 12.1, + 1.7 - 5.8} {- 2.9 -15.0, + 1.9 - 7.6}
Sources: SHPAll_VM2dS9b_*_0_0_99_200116
Note: RT Y-axis is from 0 to 3500

If the Shape property is very widely available (leftmost panel), the RTs are flat and very fast, about 500 ms, because no eye movements need to be made. As Shape becomes less available, the Basic Search strategy will make more eye movements, leading to increasing RTs with set size. In the rightmost panel where Shape is narrowly available, the RTs are very linear and very steep, especially for negative trial RTs, and the ratio of the negative RT slope to Positive RT slope is 2.1. In this case, a fixation almost always covers only a single object. In the rightmost panel, over the range of Set Size, the number of eye movements made by the model goes from 1.9 to 7.6 for Positive trials, and 2.9 to 15.0 for Negative trials. At intermediate availabilities, the RTs are somewhat negatively accelerated, reflecting how Shape becomes available for more objects in a single fixation because the objects are closer together with greater Set Size. A fairly good fit to the RTs appears in the third panel from the right, where the Shape availability parameter is set to 0.4; $r^2 = 0.98$, but the predicted RTs are somewhat more negatively accelerated than the observed. Here the range of eye movements is 1.5 - 4.0 for Positive trials, and 2.4 - 9.1 for Negative trials.

In the visual search literature, much is made of the linearity of the RTs with Set Size and how the ratio between Negative and Positive RT slopes is often in the vicinity of 2:1, which is consistent with a serial self-terminating search, which in fact is what the Basic Search strategy implements. However, over the availability parameter range shown in Figure 3.7, the predicted linear slope ratios are 0/0, 594, 9.8, 3.6, 2.7, 2.3, 2.1. Thus, the RTs are linear and slope ratio is close to 2:1 only to the extent that fixations tend to cover single objects. As pointed out by earlier

work (see Hulleman & Olivers, 2017 for a review), if the property is available over a wide enough area that more than one object can be recognized at time, the search will be faster because fewer eye movements are needed, and the slope ratio will be larger, mostly because the slope for positive RT becomes flatter. This set of fits makes it clear that a very simple model of visual search, involving availability, eye movement times, and the Basic Search strategy can account for this family of RT effects: Depending on availability, RTs can range from flat to fairly linear steeply sloping RTs, and negative RTs are generally larger than positive RTs by a ratio that converges on 2:1 as availability decreases. That is, serial processing, as indicated by the 2:1 linear slope ratio, appears when objects have to be separately fixated; when more than one object can be covered in a fixation, negatively accelerated RTs with linear slope ratios greater than 2:1 appear. This range of effects is governed simply by the availability of the relevant object properties and the need for eye movements.

2. False Alarm errors are due to action slips ("Oops!" errors). These last set of fits show zero predicted errors, which is clearly a serious misfit. The model never makes an error because in no case does the visual system produce an incorrect representation — the Shape is either veridically available for an object, or it is not, and the search continues until all objects have a

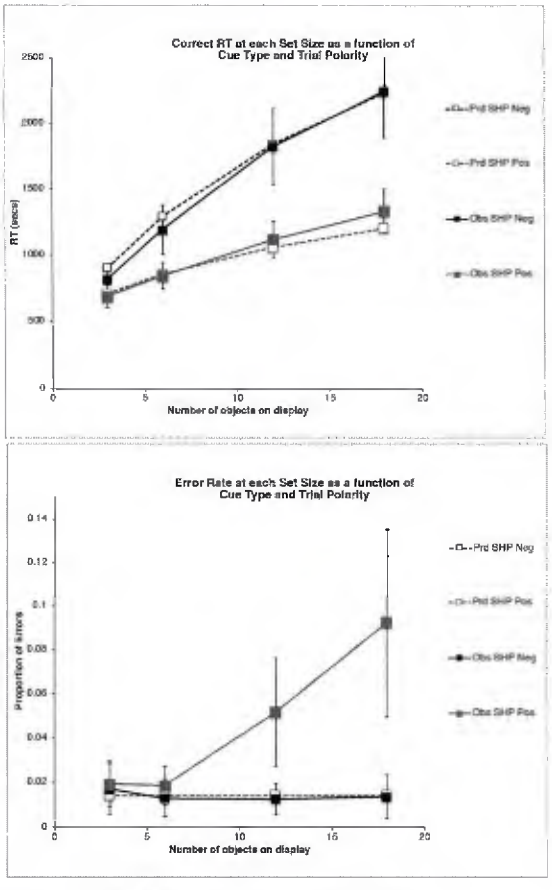

Figure 3.8. θ_S = 0.4, OopsER = .014

	r^2	aare	aae
RT	0.98	5%	54.5
ER	0.00	32%	?0.017

known shape, whereupon the choice rules always apply reliably, so no errors are made. As pointed out previously, a remarkable feature of these data is that errors on Negative trials (False Alarms) are produced at a very low and almost constant rate: the average over all three conditions is 0.014. In contrast, errors on Positive trials (Misses) increase with Set Size, especially for the SHP condition.

The model was modified to include the action slip mechanism with the *OopsER* parameter set simply to the ER for False Alarms. After the strategy determines the response, with probability *OopsER*, the *opposite* response is made. Figure 3.8 shows the fit of this model with the same availability parameter of 0.4. While this improves the ER fit substantially (*aare* goes from 100% to 32%), there is no predicted trend of increasing Miss ER, and so the r^2 for ER is 0.0. Note that the average correct trial RT is unaffected by Oops Errors because they occur independently of the visual processing and strategy execution, and incorrect trial results are not averaged into the correct trial RTs .

3. Are frequent Miss errors due to "time-out" from limiting fixations? One way to get Miss errors would be to simply stop the search and respond absent after some number of fixations with the Limited Fixations option. Figure 3.9 shows the results using *NFix* = 8 with the same availability and *OopsER* parameters as in Figure 3.8 . Note how the RTs are strongly negatively accelerated due to the early termination of searches. The predicted mean ER at Set Size18 is a good fit, but not the other set sizes. The fit could be improved by estimating a *NFix* value for each set size, but such a model has two fatal conceptual problems: First, it assumes that subjects rapidly perceive the Set Size and choose an appropriate *NFix*, which would require a more complex strategy and additional visual architecture mechanisms and their associated free parameters that consume some of the degrees of freedom in the data. Second, according to the "EPIC Philosophy," if there is a known perceptual mechanism that could account for the effect, it should be given precedence over a strategy explanation, especially one whose main justification is an ad-hoc way to fit the data. Such a perceptual mechanism known to be at work

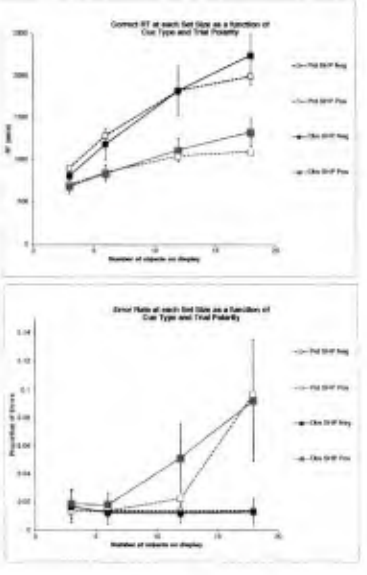

Figure 3.9. θ_S = 0.4, ϕ_S = 0.0, OopsER = 0.014, NFix = 8.

in this task is visual crowding. Thus this attempt to account for the increasing Miss ER is a poor choice; however, limited fixations will prove important in other cases in this report.

4. Crowding effects account for Miss errors sharply increasing with set size. As described above, the object Shape is assumed to be a unitary property. The crowding mechanism scrambles the Shape property of objects that crowd each other, with a "blank" or unknown property value participating in the scrambling. The "strength" of the crowding effect is specified by the crowding probability parameter ϕ_S as described above.

The effect of crowding on RT and ER is somewhat subtle and requires careful explanation. First, note that the crowding scrambling will not produce an illusory target on a Negative trial, because there is no unitary target Shape on the display. But, the scrambling can result in the distractor Shape property being moved to an object whose Shape was not actually available — a form of illusory distractor. However, on Positive trials the situation is more complex. The scrambling might move the target Shape property to a distractor object, producing an illusory target. But because the task requires simply making a Present response rather than identifying the correct target object, this illusory target has little effect on RT. But as more fixations are made, it is possible that *all* objects receive a scrambled distractor Shape, including the target object (which becomes an illusory distractor), before all of the objects have been covered by a fixation close enough to make their actual Shape property available. In this case, the Basic Search strategy will immediately halt the trial with an Absent response, which is correct on a Negative trial, but a Miss error on a Positive trial. The likelihood that an object's property will be overwritten by a distractor property value depends on both the prevalence of the distractor properties, which depends on the availability, and the crowding probability parameter. These two parameters can be varied to bracket the effects on RT and ER.

Accordingly, Figure 3.10 expands upon Figure 3.7 to show the effects on RTs and ERs of varying both the Availability θ_S and Crowding ϕ_S parameters, with *OopsER* remaining at 0.014 and with Unlimited Fixations in the Basic Search Strategy. The three rows of RT and ER graphs from top to bottom correspond to decreasing availability, with the θ_S going from 0.2 (very available) to 0.4, then to 0.6 (very unavailable). The columns of graphs from left to right correspond to crowding probability ϕ_S of 0.025, 0.075. 0.1, 0.2.

As before, decreasing availability makes the RTs longer, but the extent to which they become more linear also depends on the crowding probability parameter. But note how the Miss ER systematically increases with Set Size as ϕ_S increases, but the magnitude of the Set Size effect on ER also depends on θ_S. Because Miss errors result from an early termination of the search when all objects have the distractor property, a higher Miss ER corresponds to more negative acceleration in the RT functions. Thus the availability and crowding mechanisms jointly affect both the RT and ER. A fairly good fit is obtained with $\theta_S = 0.4$ (middle row) and $\phi_S = 0.075$ (center left).

Figure 3.11 shows a refined fit based on the above, with $\theta_S = 0.425$ and $\phi_S = 0.075$. The fit to

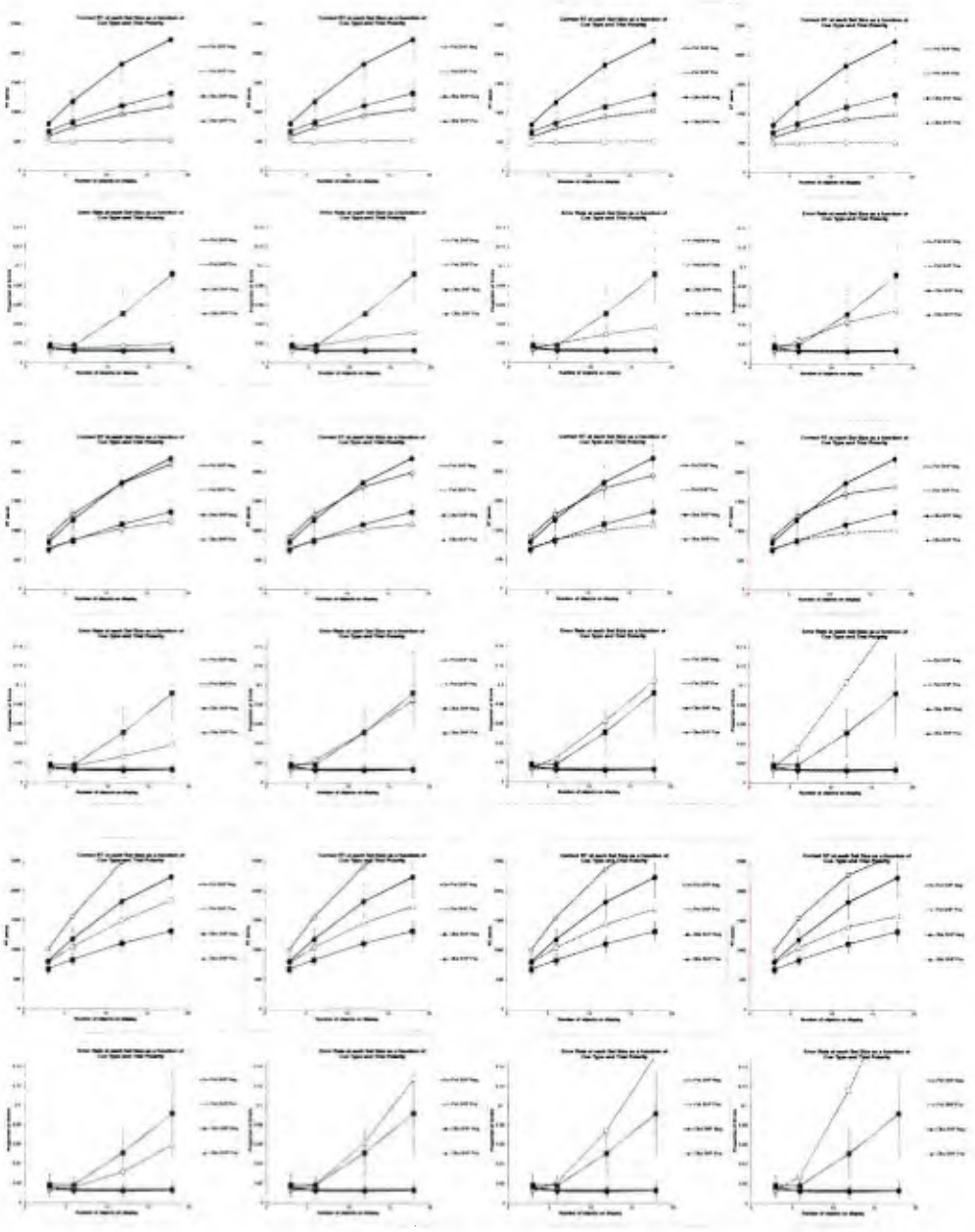

Figure 3.10. The effect of varying availability and crowding probability on RT and ER. Availability parameter: Top row 0.2, middle row 0.4, bottom row 0.6. Crowding probability parameter: Left column 0.025, center left 0.075, center right 0.1, rightmost is 0.2.

both RT and ER is extremely good, capturing the ER effect with only two adjustable parameters, whose effect depends only on the fact that greater set size produces more objects that crowd each other. Because the crowding can cause the strategy to terminate early more frequently with increasing set size, the RTs are also more curvilinear than shown in Figure 3.8 which uses the same strategy, θs and OopsER parameters.

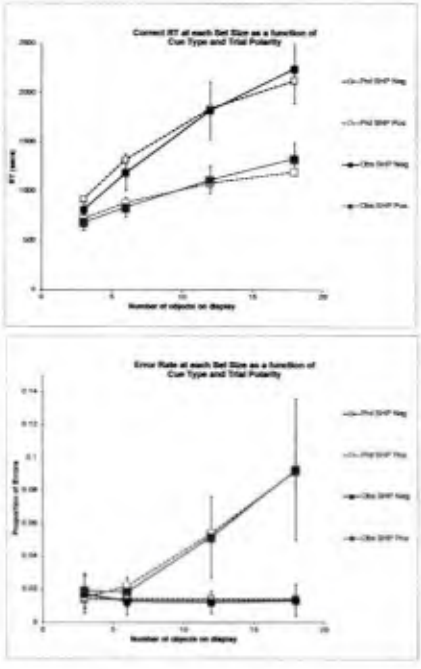

Figure 3.11. Basic Search Strategy, Unlimited Fixations, $\theta_S = 0.425$, $\phi_S = 0.075$
Source: SHPAll_VM2eLS9c_425_075_014_99_200304

	r^2	aare	aae
RT	0.97	7%	78
ER	0.99	11%	0.002

Account of SHP Task. Thus the SHP aggregate RT and ER data can be well accounted for by a model which combines the known visual mechanisms of availability and crowding, with the motor mechanisms of eye movements and manual responses, both with simple inherent error mechanisms. These mechanisms are tied together by a simple cognitive strategy that attempts to produce fast and reasonably accurate responses based on the output of the visual system. The effects in the data reflect only the perceptual-motor limitations; no concept of a central "bottleneck" or covert attention is required to account for this prototypical visual search task.

Explanatory Sequence for the CSF Task

The next explanatory sequence is for CSF task aggregate data. CSF is a good next choice for explanation because like SHP, it involves a single property of the objects, but the color property is at the opposite extreme of availability from SHP: There are many psychophysical results that suggest that object color is very widely available in the visual field, so widely that it is often described as "popping out" in a visual search task. The explanatory sequence here shows that "pop out" is not a special mechanism, but simply a result of a property being very available over the visual field. Similar to the SHP sequence, the RT effects will be accounted for first, and then the ER effects.

1. Fast and flat RTs are due to high Color availability, and False Alarms due to Oops Errors. The starting point in the sequence is based on the SHP account: The model will include Availability, Eye Movements, Oops Errors, and the Basic Search strategy with Unlimited Fixations. The explanatory sequence starts with Figure 3.12 which shows the effects of changing the Availability θ_C of the Color property. The model fits shown use the Basic Search Strategy with Unlimited Fixations, an OopsER of 0.014, and θ_C ranging from 0.0 (almost all objects have available color), 0.1, 0.15, and 0.2, and no crowding effect ($\phi_C = 0.0$). First, note that for all of the fits, the Miss ER is identical to the False Alarm ER, which is turn is identical to the Oops ER. In other words, only Oops errors are predicted.

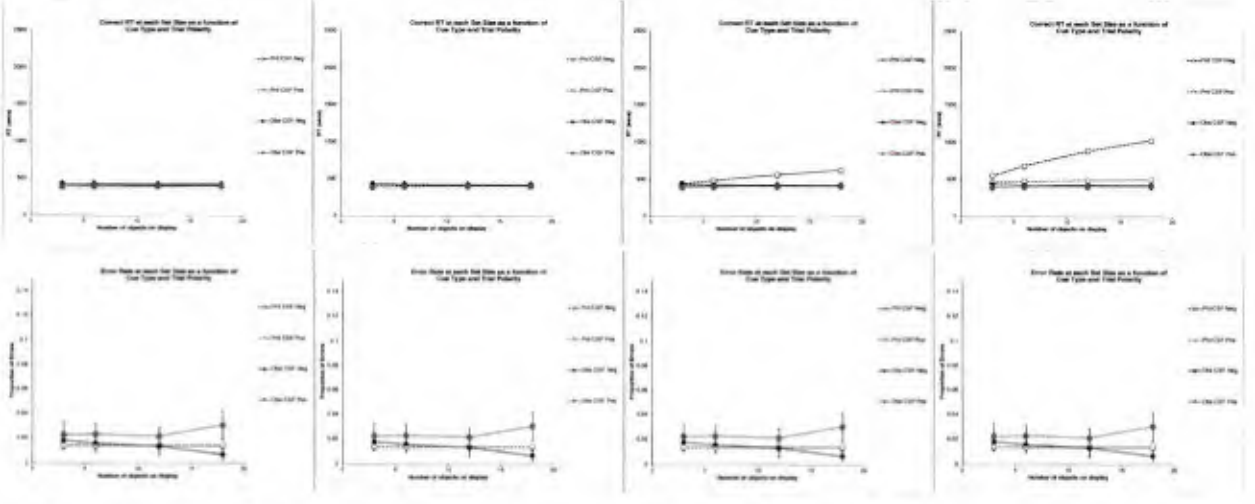

Figure 3.12. Basic Search strategy, OopsER = .014, Color availability left-to-right: 0.0, 0.1, 0.15, 0.2. No crowding effect.

For the first two fits ($\theta_C = 0.0$ and 0.1), the predicted RTs are very fast and flat and very close to the observed RTs. Underlying these RTs are the number of fixations/trial produced by the model. As expected, for $\theta_C = 0.0$ the model makes no fixations other than on the initial fixation point. For $\theta_C = 0.1$ and Set Size 18, the model makes only 0.022 fixations/trial for Positive trials, and 0.128 for Negative trials, which produces RTs that are almost flat and almost identical for

positive and negative trials. However for $\theta_C = 0.15$ there is significant slope in the predicted Negative RT which is very different from the observed RT, resulting from 1.068 fixations/trial at Set Size 18. But the predicted Positive RT is still almost flat and very close to the observed RT at a tiny .094 fixations/trial at Set Size 18. In this case, the predicted slope ratio is 48:1! Finally, for $\theta_C = 0.2$, the Positive RT is now visibly greater than the observed, but still fairly flat with 0.483 fixations/trial at Set Size 18, while the Negative RT is much more steeply increasing with 2.852 fixations/trial at Set Size 18. The slope ratio is now only 13. From these results it is very clear that even a single fixation on the average at the largest Set Size will produce a definitely sloped RT, meaning that θ_C needs to be close to 0.1 or below to be a plausible fit to the fast and flat RTs. But again, this leaves the ER effect of more Misses than False Alarms unexplained. Rather than an ad-hoc explanation that the OopsER is larger for Present responses than for Absent responses, it would be better to find an explanation in terms of the "EPIC Philosophy" for the combination of flat RTs but more Misses than False Alarms.

2. Fast and flat RTs and Miss Errors are due to a Fixed-Eye Strategy. Rather than making the Color property extremely available, another way to get a flat RT is the Fixed-Eye strategy (Box 3.2), which limits the fixations to only the initial fixation, allowing no subsequent eye movements at all, with the response to be made based only on what is available during this initial fixation. If the Color property is not always Available, some Miss responses will result. This gives an extremely simple account: on a Positive Trial, if the target Color is available, the response will be Present; if the target Color is not available, the response will be Absent; on a Negative trial, all responses will be Absent. All of these responses will be inverted at the constant *OopsER*. Thus the RTs are constant, determined only by the constant time required for perception, decision, and action, while the ERs are determined only by *OopsER* and θ_C. Figure 3.13 shows the Fixed-Eye strategy model predictions for the same set of values for θ_C and OopsER as Figure 3.12. Note that the rightmost ER graph for $\theta_C = 0.2$ has a different scale since

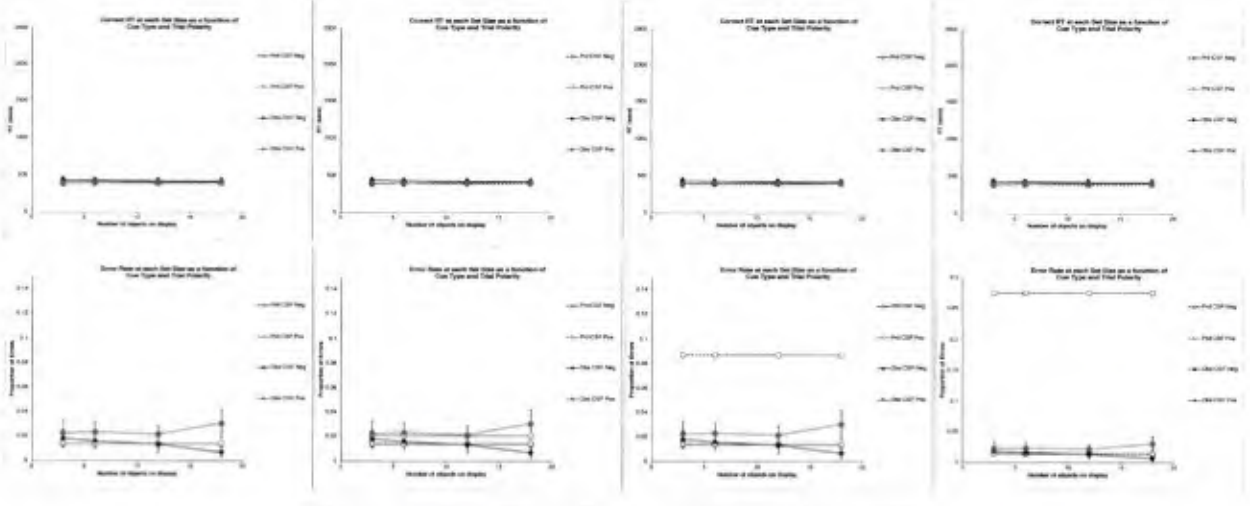

Figure 3.13. Eyes fixed (NFix = 0), Oops ER = 0.014. Color availability left-to-right: 0.0, 0.1, 0.15, 0.2. No crowding effect.

the predicted Miss ER is very high (0.276). A fairly good fit is obtained for $\theta_C = 0.1$. Figure 3.14 shows a refined fit obtained with $\theta_C = 0.11$.

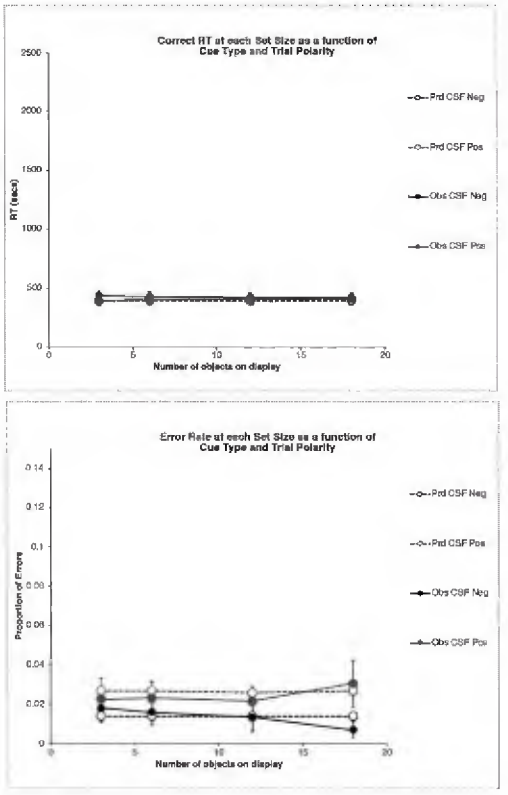

Figure 3.14. Fixed-Eye strategy, $\theta_C = 0.11$, OopsER = 0.014.

	r^2	*aare*	*aae*
RT	0.01	4%	18.2
ER	0.68	25%	0.004

Source:CSFAll_VM2eS9c_11_0_014_0_200225

3. Crowding has a negligible effect on Miss ER in the Fixed Eye strategy. Crowding explained the Miss ER effects in SHP. What about for CSF? A series of Fixed-Eye model fits (not shown) with a very wide range of values of θ_C and ϕ_C showed detected, but negligible, effects of crowding probability ϕ_C on Miss ER. In the most extreme test cases, crowding added at most 0.001 to the Miss ER.

This result is explained as follows: On a Positive trial, θ_C determines whether the target object Color will be available when the eye is at the initial fixation point. Crowding might swap which object has the target Color by producing a pair where the actual target becomes an illusory distractor and an actual distractor becomes an illusory target. This has no effect on response correctness. It is also possible that scrambling involving multiple objects could result in the target object Color property being lost from overwriting with a blank property, turning the target object into an *illusory blank*. The result would be a Miss Error, even if the target Color was originally available.

But crowding scrambling is relatively unlikely under the Fixed-Eye strategy because in the initial fixation, the average eccentricity is only 8.8, and as shown in Figure 3.6, the mean number of crowding flankers is only 1.0 at Set Size 18, and only one scrambling operation will be performed since no eye movements after the initial fixation are made. Thus, given the high availability of Color, the probability is extremely low that an available target property will get overwritten and turned into an illusory blank; rather, almost always any crowding will simply move the target object Color property (either available or blank) from one object to another. Thus the correctness of the response depends almost completely on whether the target property was available during the initial fixation, which depends only on the Availability parameter.

Thus, although crowding is certainly at work in the CSF task, the Fixed-eye strategy and the high availability of Color makes the effect on ER negligible. In contrast, the next explanatory sequence shows how crowding plays a powerful role in the COC task, even though both Color and Orientation are highly available.

Account of the CSF Task. The CSF task is at the opposite extreme from the SHP task, and the models capture the difference. The CSF effects can be accounted for by the extremely simple Fixed-Eye strategy and the availability mechanism. No special "pop-up" mechanism is required — Color is simply visible enough that eye movements would be rarely required, and can simply be eliminated if some Miss Errors are acceptable. Interestingly, the effect of crowding, while detectable in the model, was not needed to adequately fit the data in this situation of no eye movements and high availability. But a value could be assigned to ϕ_C (e.g. for use in the COC task) without affecting the accuracy of the CSF Model.

Explanatory Sequence for the COC Task

1. Shallow COC slopes are due to eye movements. Even though Color is involved in the COC task, the Eyes-Fixed strategy can be immediately ruled out because (1) the COC RTs are sloped rather than flat; (2) Orientation is almost certainly less available than Color; (3) the task is more difficult because a conjunction target must be located, and so should be more like SHP than CSF. Thus the Basic Search strategy is a good starting point for the explanatory sequence.

A reasonable simplifying assumption is that the availability parameter θ_C has the same value in COC as in CSF, namely $\theta_C = 0.11$. A rough conclusion from available data is that Orientation is less available than Color, but there are no well-parameterized data on the relative availability of Orientation compared to Color, and this data set did not include an Orientation Single Feature version of the task that could be used to estimate θ_O separately.

Accordingly, Figure 3.15 shows a set of bracketing fits with the Basic Search strategy with Unlimited Fixations, using the best-fit value for Color availability $\theta_C = 0.11$, and with a range of Orientation availability θ_O values from 0.11 (same as Color) to 0.25. The RTs become definitely sloped for values of θ_O greater than 0.11, and are fit fairly well with $\theta_O = .225$ or .25, justifying that Orientation is less available than Color.

This makes an important point: *Eye movements are compatible with shallow slopes.* It suffices that the properties are available enough that only a few eye movements are required. For example at Set Size 18 and $\theta_O = 0.25$, the number of eye movements average only 1.0 for Positive trials, and 3.0 for Negative trials. However, because the search continues until the target is found, as in the corresponding SHP case, there are no deliberate Absent responses, and so no Miss Errors except for Oops Errors.

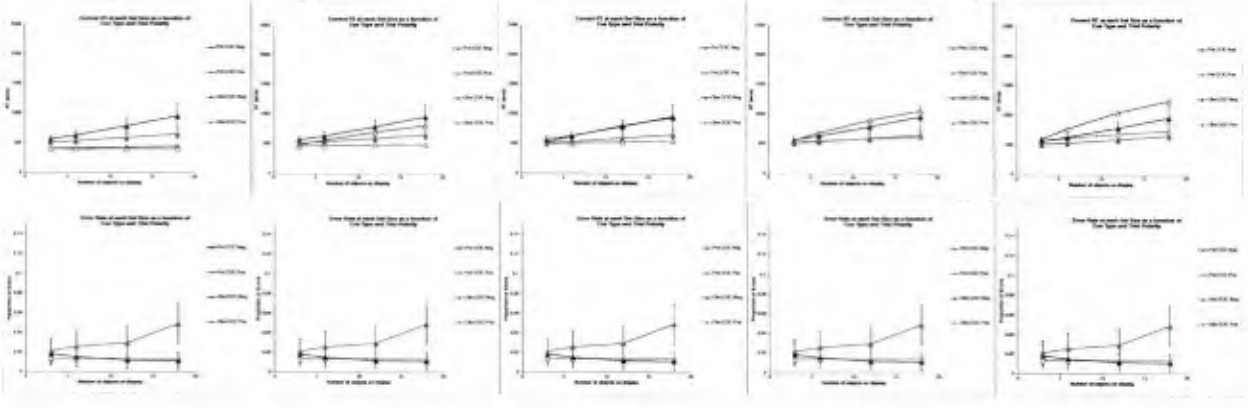

Figure 3.15. COC Task, No Crowding, Unlimited Fixations, OopsER = 0.014, θ_C = 0.11, θ_O = 0.11, 0.2, 0.225, 0.25, 0.3.

2. Crowding causes massive False Alarm errors in COC. Can crowding account for the Miss ER effects in COC as was the case with SHP? Figure 3.16 shows the results for Basic Search/Unlimited Fixations and availability parameters based on the best fit from the previous results ($\theta_C = .11$, $\theta_O = .25$) and with a small crowding effect added. The crowding probability is

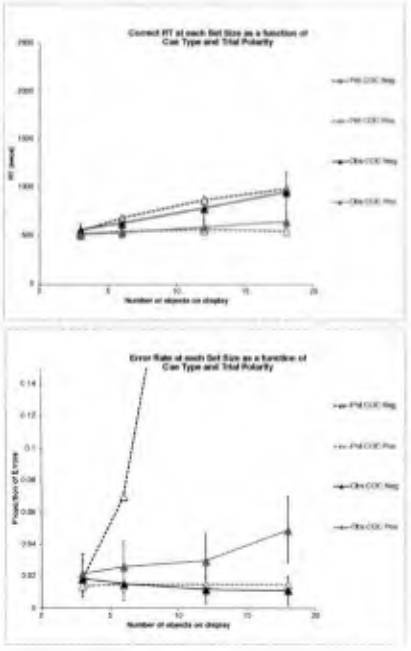

Figure 3.16. $\theta_C = .11$, $\theta_O = .25$, ϕ_C, $\phi_O = 0.025$, *OopsER = 0.014*

assumed to be equal for Color and Orientation, namely $\phi_C = \phi_O = 0.025$. The small value makes intuitive sense because Color is reported to be relatively insensitive to crowding. Similarly, the very distinct values of Orientation might also be resistant to crowding. The predicted RTs are fit fairly well, but note the massive and rapidly increasing predicted False Alarm ER shown in the graph! Any non-zero value for ϕ_C and ϕ_O produces a similar effect.

This surprisingly gross misfit is important to understand. The COC task has an almost equal number of Red and Green property values on the display, and also an almost equal number of Horizontal and Vertical property values on the display. The Target object is the only object that is both Red and Vertical. Since Color is very available, and Orientation fairly available, there are many instances of these values available at a time. However, crowding causes the Color and Orientation property values to get scrambled between the objects, and if a Vertical replaces a Horizontal on a distractor, or a Red replaces a Green on a distractor, the result is an illusory target. The Basic Search strategy will terminate as soon as an object appears to be a target regardless of whether it has been fixated. So even at this low crowding probability, on Negative trials, illusory targets are common enough to cause the strategy to frequently terminate early with a False Alarm error, and at a rate that increases with Set Size, as more objects crowd each other on the display.

But if the objects happen to be widely spaced enough that crowding doesn't create an illusory target, rather than a False Alarm, a correct Absent response will be made, but a few fixations might be required to make the properties all available, resulting in a sloping Negative RT. But on Positive trials, if the actual target is not visible at first, the frequent illusory targets will cause the strategy to terminate quickly with a Present response, which is still the correct response. The result is an almost flat Positive RT. Even if illusory targets happen not to be present, the

unlimited search will eventually find the actual target, so the only Miss errors are Oops errors.

Thus this model is remarkably incorrect even though it incorporates all of the previously considered visual and motor mechanisms, including crowding. The implication is that something additional is involved in the the COC task, and a change to the strategy is the next step in the explanatory sequence.

4. The COC strategy must confirm that a target is actually present. The False Alarms can be prevented with the Confirm-Positive version of the Basic Search strategy shown above in Box 3.3. Figure 3.17 shows the results for Basic Search with Confirm-Positive strategy using the same availability values as previous, and holding $\phi_C = 0.025$, and with three different values of the Orientation Crowding probability ϕ_O, increasing from left to right from 0.025, then 0.1, to 0.2. Starting at the left panel and going to the right, the Positive RTs have a slope similar to the observed slope, but are larger than the observed values, while the Negative RTs start much steeper than the observed values, and become extremely steep with increasing ϕ_O. The problem is that the value of ϕ_O that produces RTs most like the observed (left-most panel, $\phi_O = 0.025$) produces very few Miss Errors, but increasing ϕ_O to a high enough value to produce enough Miss Errors results in huge Negative RTs, as well as Positive RTs that are substantially larger than the observed.

What is happening on Negative trials is that the combination of Crowding, Confirm-Positive, and Unlimited Fixations means the search continues until all of the objects appear to be distractors, whereupon an Absent response is immediately made; many fixations to resolve

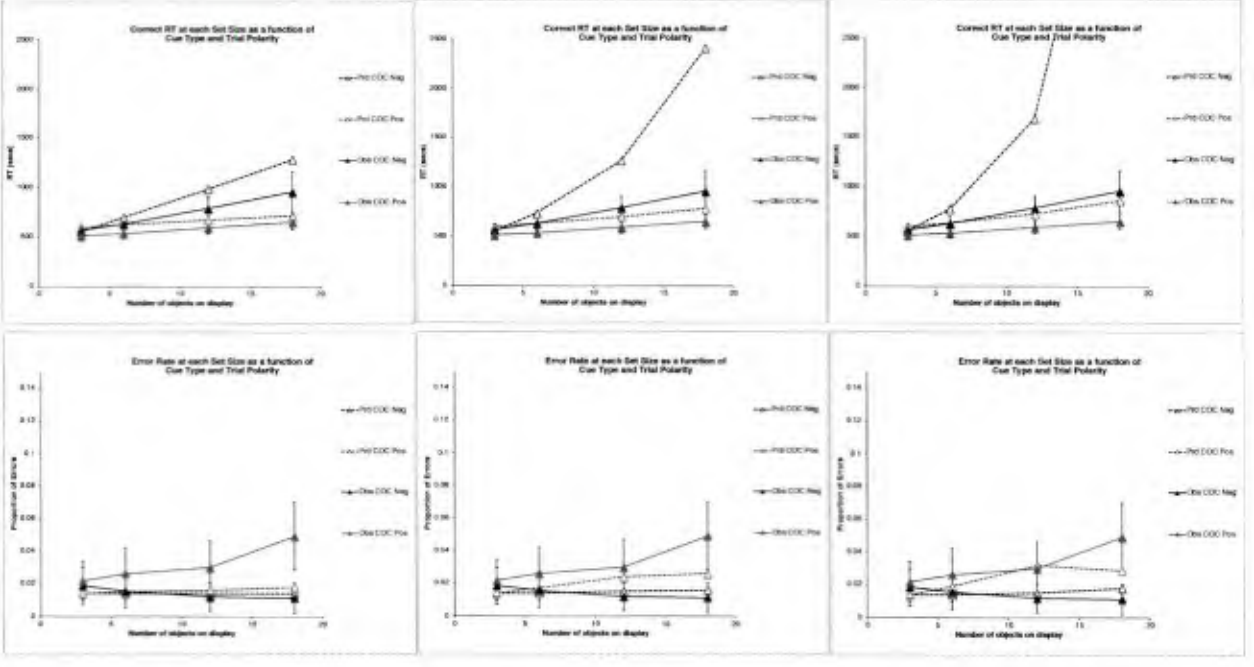

Figure 3.17. Basic Search with Confirm-Positive and Unlimited Fixations. $\theta_C = 0.11$, $\theta_O = 0.25$, $\phi_C = 0.025$, $\phi_O = 0.025, 0.100, 0.200.$, OopsER = 0.014.

illusory targets may be needed before the Absent response is triggered, but no deliberate (non-Oops) False Alarm Errors are produced, thanks to the Confirm-Positive step. On a Positive trial, the only way to make a deliberate Miss Error is if the actual target object appears to be an illusory distractor at the same time that all the other objects appear to be distractors. This will not happen very often, resulting in very few Miss Errors. Thus, even though Crowding is incorporated into the model, the problem appears to be that Unlimited Fixations produces either extremely long RTs or under-predicted Miss Errors. While Crowding is a problem, intuitively, it seems like subjects should be able to produce much faster RTs with an acceptable Miss ER.

5. COC with Limited Fixations produces fast RTs with acceptable ER. While limited fixations was a poor choice in fitting SHP, in COC it plays a critical role in achieving fast RTs without excessive errors. The Confirm-Positive strategy was modified to include the Limited-Fixations option (Box 3.4) of responding Absent after a fixed number of fixations *NFix*. This combination of strategy options in the presence of crowding eliminates frequent False Alarm errors, produces fast sloping RTs, and an acceptably low Miss ER that increases with Set Size. Figure 3.18 shows a good fit with $NFix = 3$, $\theta_C = 0.11$, $\phi_C = .025$, $\theta_O = 0.2$, $\phi_O = 0.025$.

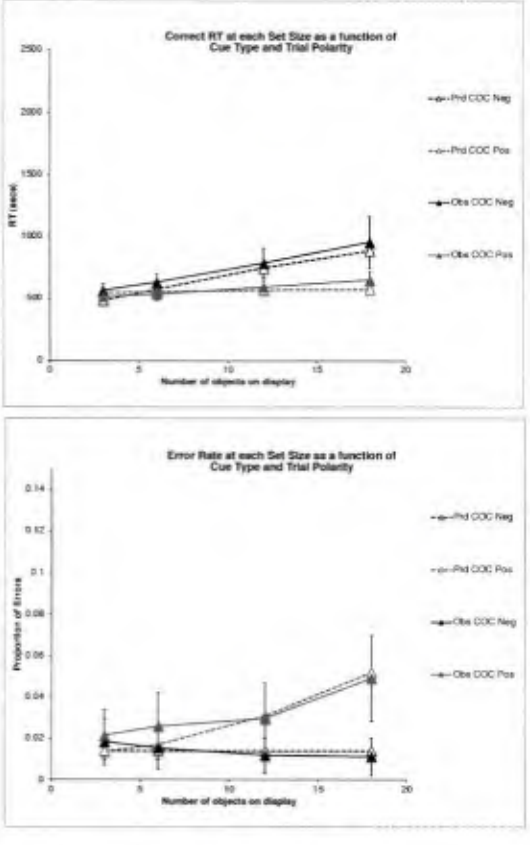

Figure 3.18. Confirm-Positive with Fixation-Limited. $\theta_C = 0.11$, $\phi_C = 0.025$, $\theta_O = 0.2$, $\phi_O = 0.025$, OopsER = 0.014, NFix = 3.
Source: COCAll_VM2eCS9c_110_025_200_025_014_3_CP_200208

	r^2	aare	aae
RT	0.92	8%	49.5
ER	0.88	19%	0.004

Account of the COC Task. The final model shows that the COC task is very different from the SHP and CSF tasks. The problem is that even with a very low probability of crowding, the scrambling of the Color and Orientation properties between objects produces a plethora of illusory targets, which requires a form of double-checking to confirm the presence of the actual target to avoid massive False Alarm errors. This was not at all an issue in the other tasks, where a single and unitary target property means that crowding scrambling could not cause an illusory target to appear on a Negative display, and at most would change the apparent, but irrelevant, location of the target in a Positive display. However, so prevalent are the illusory targets in COC that the confirm-positive double-checking takes a long time on Negative trials. The Confirm-Positive Basic Search with Limited Fixations thus becomes a reasonable strategy: the RTs are shortened without incurring more than an acceptable number of Miss Errors. Thus, like the other search tasks, the conjunction task can be explained simply in terms of the known perceptual mechanisms, but requires a task strategy appropriate to the inherent ambiguity produced by crowding of the COC stimuli.

Model results for all three tasks

Figure 3.19 shows the predicted RT and ER compared to the data, and Table 3.2 shows the strategy and parameter values for the model in each task condition, and the goodness-of-fit metrics for RT and ER in each task condition, and for the whole set of 24 RT and 24 ER data points. The graphs show a very close correspondence between predicted and observed values, indicated quantitatively by the high r^2 values and low *aare* or *aae* values. As pointed out above, when the RTs are flat and identical, produced by the Eyes-Fixed strategy in CSF, the r^2 values will be zero. There is some tendency for the RTs for the SHP condition to be more negatively accelerated than the data.

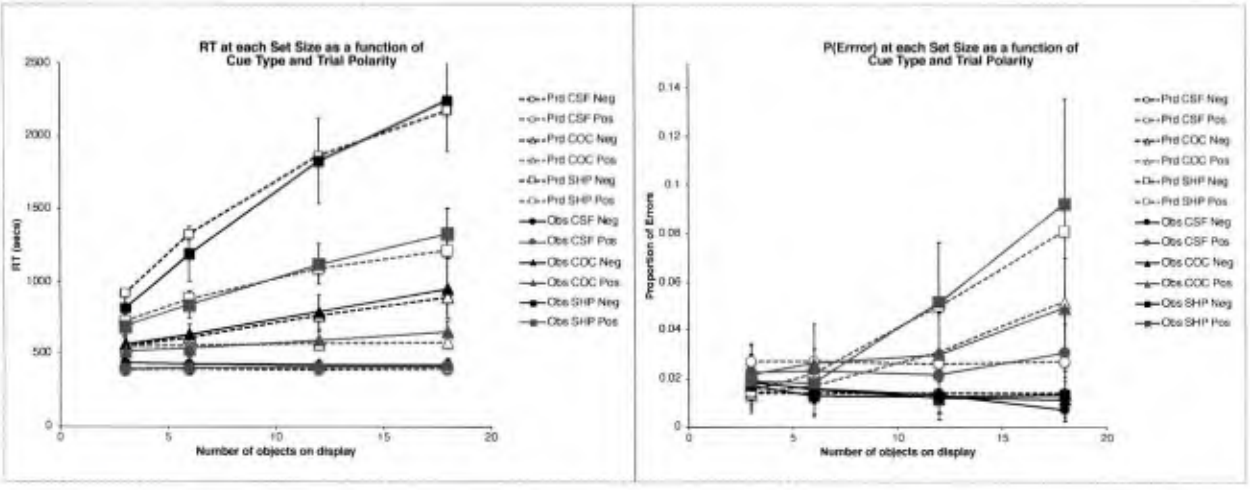

Figure 3.19. Predicted (open points, doted lines) and observed (solid points and lines) results for the All Subjects Observed (solid points and lines) and predicted (open points and dotted lines) for correct trial RT in each task condition. CSF: circles, COC: triangles, SHP: squares. Positive trials: red, Negative trials: black. The 95% confidence intervals are based on the standard error of the mean of the subjects' mean values underlying each data point and thus reflect between-subject variability.
Source: AllSubs_VM2eLS9c_11_0_025_2_025_425_075_0_3_99_CP_200318.

Table 3.2

		NFix	Color/Shape		Orientation		OopsER	GoF: RT			ER		
Task	Strategy		θ	ϕ	θ	ϕ		r^2	*aare*	*aae*	r^2	*aare*	*aae*
CSF	Eyes Fixed	0	0.11	0.0			0.014	0.01	4%	18	0.68	25%	0.004
SHP	Basic Unlimited	U	0.425	0.075			0.014	0.97	7%	79	0.99	11%	0.002
COC	Limited with Confirm-Positive	3	0.11	0.025	0.20	0.025	0.014	0.92	8%	50	0.88	19%	0.004
All Tasks								0.99	6%	49	0.95	19%	0.003

However, notice that the confidence intervals around the means are quite large for some of the RT data, and especially large for the Positive trial (Miss) ER data. The ER data for individual subjects were inspected; there were substantial individual differences, with some subjects

achieving almost perfect performance, while one subject produced 23% errors in the most difficult condition! Further inspection also showed different patterns in the RT effects for individual subjects in the same condition.

Representing Individual Differences in EPIC Models

The EPIC architecture provides a theoretical framework for characterizing individual differences: individual people might have different architecture parameters — e.g. clearly different people have different visual acuity. Unfortunately, the Wolfe, et al. experiment was a between-subjects design, so there is no direct way to isolate this source of individual differences. Also according to the architecture, and earlier applications of EPIC to human performance data (e.g. Schumacher, Seymour, Glass, Fencsik, Lauber, Kieras, & Meyer, 2001) individual subjects might devise different strategies for performing the same task. Like too many experimental psychologists, Wolfe, et al. provided ambiguous instructions on speed vs. accuracy — respond "as quickly and accurately as possible" — and did not use any payoff or incentive scheme to influence subjects' choice in any particular way. But they did provide trial-by-trial accuracy feedback and extensive practice, both of which might induce subjects to adopt a stable strategy, even if it was different for individual subjects. Clearly averaging over subjects who have different architectural parameters and different task strategies could produce average data that actually represents none of the subjects. While this is a long-standing insight in cognitive modeling, it is unusual for modelers to collect and model data reflecting this insight, and published data rarely allows the modeler to act on it. Fortunately, the Wolfe, et al. dataset includes all of the individual trials for each subject, allowing a subject-by-subject data analysis.

Cluster analysis of Wolfe, et al. subjects. To evaluate whether there were individual differences that distorted the pattern of effects in the mean data, the RT and ER functions were plotted for individual subjects; on inspection, they seemed to fall into a small set of patterns in each task condition. This was confirmed more formally by doing cluster analysis of the subjects in each task condition, using as clustering variables the intercept, slope, and best-fit quadratic coefficient for RT, and mean Miss ER and Miss ER at set size 18. Cluster analysis is usually done on very large data sets, so its application here is more to formalize what would otherwise be an intuitive hand-clustering process. However, there were about 500 trials per data point per subject, so the individual subject data was reliable enough to mitigate to some extent the fact that only 9 or 10 subjects were present in each task condition. The clustering method was k-medoid, and the clusters for k=1 through 5 were determined. The analysis was separate for each task condition, and the variables used in the clustering were made as few as possible consistent with a clear clustering result for k=3. For example, the quadratic coefficient for RT was used as a clustering variable only in the SHP condition, where some of the subjects had strongly negatively accelerated RT functions. In addition, the final clusters were chosen so that averaging the RT and ER data within a cluster preserved the basic trend patterns present for the individual subjects in the cluster. This required moving a total of two subjects out of the computed clusters into a cluster of one subject each, as will be described. The following sections will present, for each task condition, the individual subjects in their clusters, followed by the mean data for each cluster, followed by the results of fitting the model to each cluster. The model fits are based on

the modeling results for the explanatory sequences for the aggregate data, and so only the final best-fitting model will be presented for each cluster. The results will be presented for each condition in the same order as the explanatory sequences, starting with SHP, then CSF, and finally COC.

A preview of the results is that the clusters can be accounted for well, with individual models using the same visual and strategy mechanisms that could account for the the aggregated results; in particular, no new strategies were needed.

SHP Clusters

Figure 3.20 shows the four groups of subjects resulting from the cluster analysis of the SHP subjects; each column is a cluster with RT on the top and ER on the bottom; red curves are Positive trials, black curves are Negative trials. The curves for different subjects are shown by different plotting points (circles, triangles, squares, diamonds). The clusters are shown in order of increasing ER from left to right. Each cluster is labeled in terms of the subject ID numbers in that cluster and a short description. There is a single-subject cluster at the far right that was not included in the modeling. Note that the plotting scales cover the very large RTs and ERs produced in these clusters.

The leftmost cluster 129:Slow/LowER, has three subjects with fairly linear RTs with large slopes, but low ER with a small increase for Positive ER with set size. The next cluster 37:VerySlow/MedER, has two subjects with much more steeply sloped RTs and a much greater increase in Positive ER. The next cluster 458:NegAcc/HighER, has relatively fast but negatively accelerated RTs with very much larger ERs that for Positive trials greatly increase with set size.

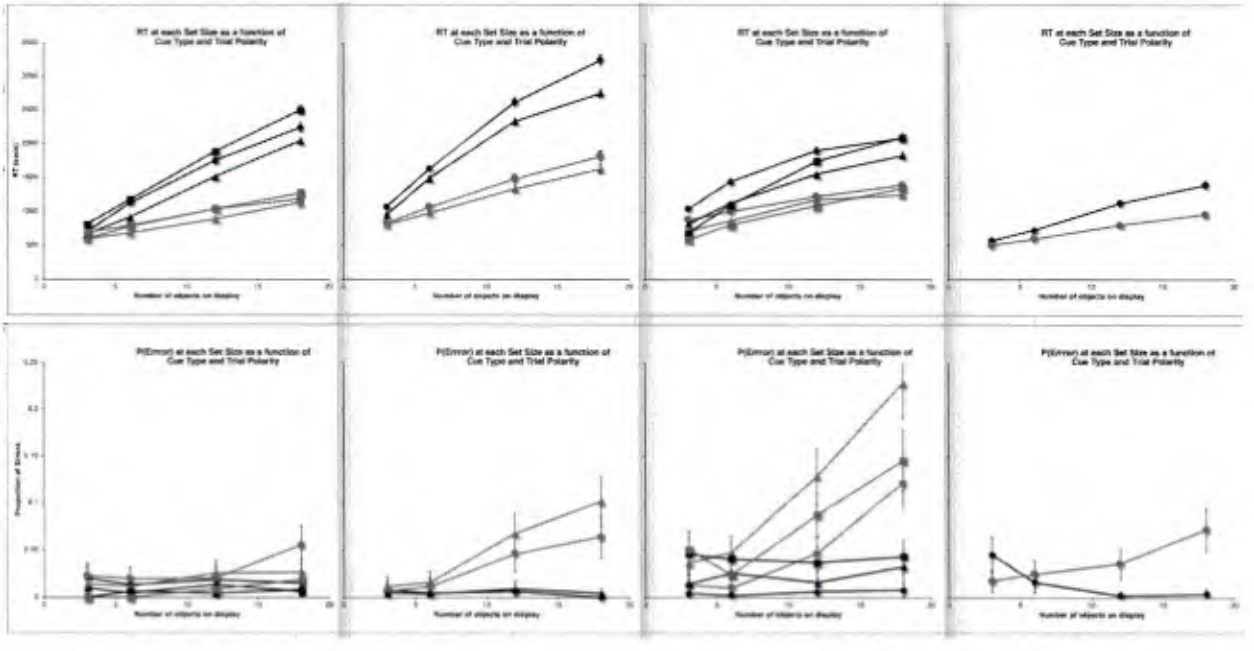

Figure 3.20. Individual subject RT and ER in each cluster of the SHP subjects. The 95% confidence intervals are based on the approximately 500 trials for each individual subject's RT and errors underlying each plotted data point.

The right-most subject was originally grouped with the leftmost cluster, but because the RT slopes were clearly much lower than the other three subjects and the ER higher, this subject was moved into a single-subject cluster which was not modeled in this report. Figure 3.21 shows the average RT and ER the three modeled clusters, and Table 3.5 provides their statistics. Note how the basic pattern of the effects for the individual subjects in the cluster is reflected in the means for that cluster, and how the confidence intervals for the data points in a cluster are fairly tight. Thus even though there are only a few subjects being averaged together, the subjects in a cluster

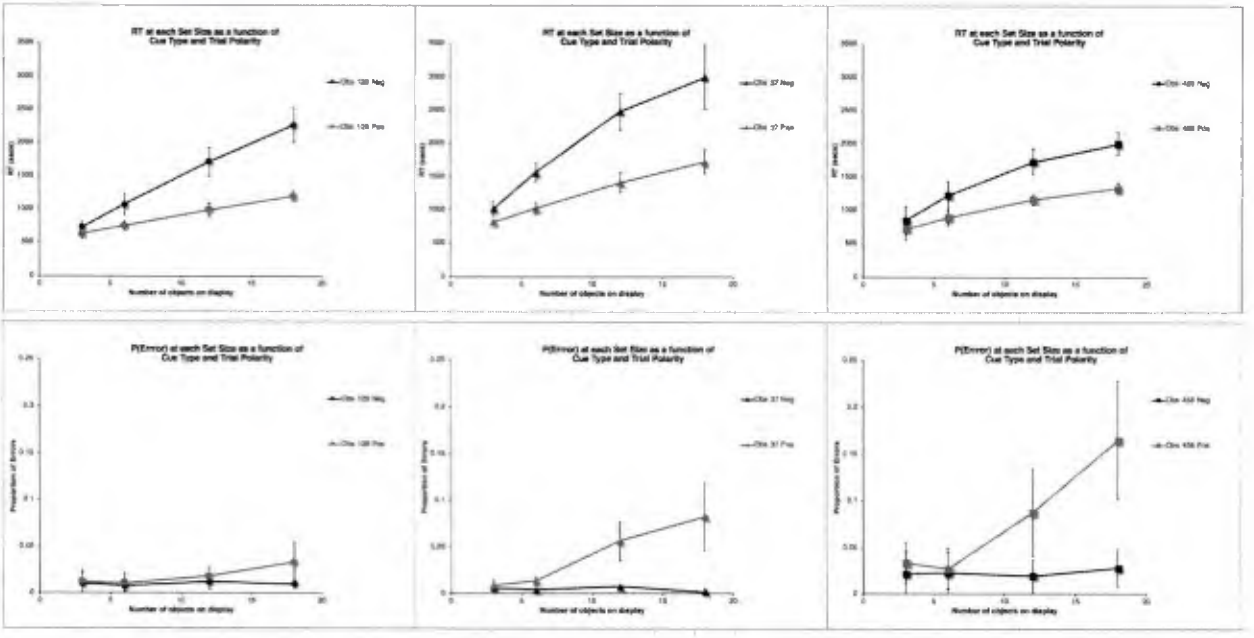

Figure 3.21. Mean RT and ER in each cluster of the SHP subjects. The 95% confidence intervals on each plotted point are based on the mean RT and ER for that data point for each of the 2-3 subjects included in the cluster.

are similar enough to each other that their average data is less noisy than in the all-subjects data.

Table 3.5

	Negative				Positive					Slope ratio
SHP Cluster	Intercept	Slope	r^2	ER	Intercept	Slope	r^2	ER	ER Max	
129:Slow/LowER	446	102	1.00	0.010	521	38	1.00	0.019	0.033	2.70
37:VcrySlow/MedER	718	133	0.98	0.005	651	61	1.00	0.041	0.083	2.19
458:NegAcc/HighER	707	76	0.96	0.023	628	41	0.98	0.078	0.165	1.87

The model was fit to each cluster following the same approach described for the overall average data. In all clusters, the Oops ER was set to the False Alarm ER. The model fits for the SHP clusters are shown in Figure 3.22, and the parameters and goodness-of-fit statistics are listed in Table 3.6.

These fits demonstrate strong differences in the visual parameters and the search strategy. The subjects in the leftmost cluster 129:Slow/LowER have steeply sloped RTs and low ER with slightly increasing Positive ER. This fit was obtained with the Basic Search strategy with unlimited fixations (shown as U for Nfix) and somewhat poor availability of the Shape property, but cautiously low Oops ER. These subjects could be described as needing many fixations because they could not see the object shapes very well in the periphery, but they took their time and achieved fairly high accuracy. Since the fixations are unlimited, the increased Miss ER at greater set sizes is due to crowding effects changing the target into an illusory distractor. As in the overall average data, the model predicts somewhat negatively accelerated RTs on Negative

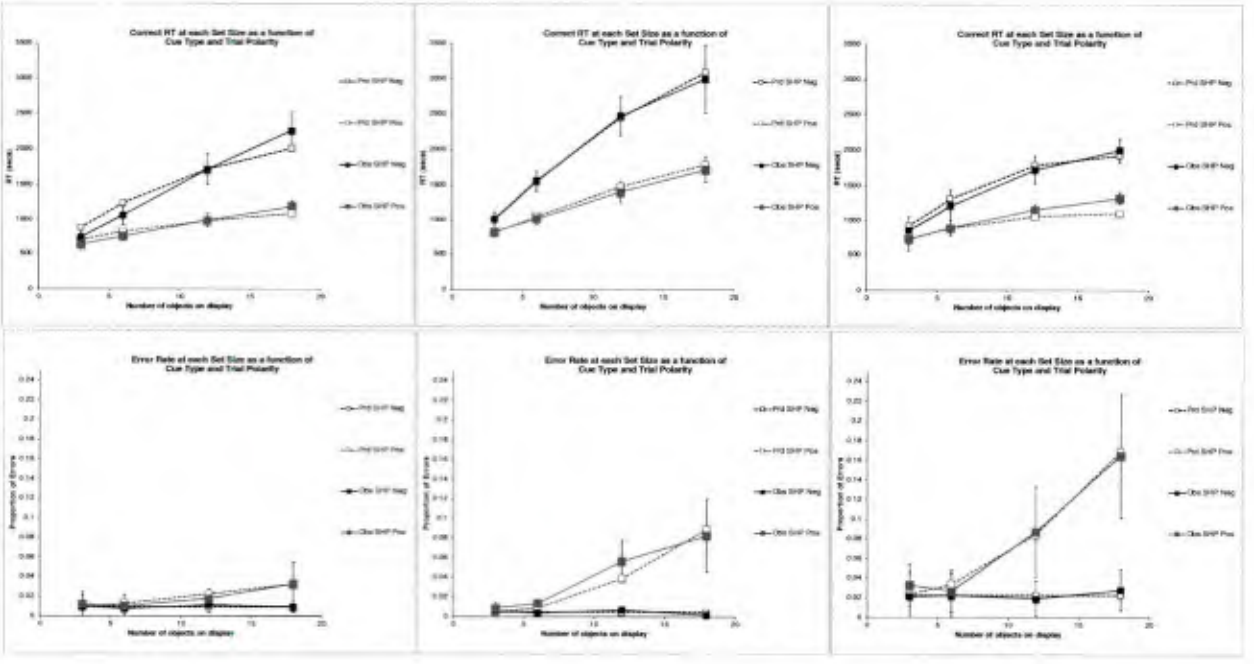

Figure 3.22. Observed (solid lines and points) and Predicted (dotted lines, open points) for the SHP Cluster Means. Upper panels: RT; Lower panels: ER. Left panel: 129:Slow/LowER. Middle panel: 37:VerySlow/MedER. Right panel: 458:NegAcc/HighER.
Sources: SHP129_VM2eLS9c_375_025_01_99_NC_200309, SHP37_VM2eL9c_6_05_005_99_NC_200309, SHP458_VM2eLS9c_425_1_023_8_NC_200309.

trials because more than one object can be perceived in a single fixation as the object density increases with set size. This is a systematic misfit of the model, but as shown in Table 3.7, the r^2 for RT is very high nonetheless.

The middle panel of Figure 3.22 shows the fit for cluster 37:VerySlow/MedER. The RTs are very steeply sloped and the Miss ER is medium-high and increases substantially with set size compared to the first cluster. This fit was also obtained with the Basic Search strategy with unlimited fixations, but as shown in Table 3.7, with larger availability and crowding parameters than the first cluster. Since the Shape property is quite a bit less available than in the first cluster, the predicted RTs are more linear and match the observed RTs quite well. The larger crowding parameter also produces the higher Miss ER. The goodness of fit metrics are extremely good. These subjects apparently had a great deal of trouble detecting the Shape property and so had to make more fixations than in the first cluster, but also were more subject to Miss errors from crowding effects. The large value for ER *aare* is a good example of the difficulty posed by some extremely low ERs in the observed data; since these values appear in the denominator of the relative error calculation, even a small absolute deviation of the predicted from the observed value will produce a large relative error. In contrast, the r^2 and *aae* shows fairly close fit.

The rightmost panel of Figure 3.22 shows the fit for cluster 458:NegAcc/HighER. These subjects produced RTs that are fairly fast, but strongly negatively accelerated, and the ERs are much higher than in the first two SHP clusters and the Miss ER is much higher and increases with set size more than any other cluster in any other condition. The fit capture these effects with

both limited fixations as well as substantial crowding effects. As shown in Table 3.6, this fit was obtained with moderate poor availability, a large crowding probability, high Oops ER, and a limit of 8 fixations. This combination produces the powerful trend in Miss ER, as well as the negatively accelerated RTs. These subjects can be described as willing to accept a lot of errors in return for shorter RTs simply by quitting with an absent response after several fixations.

In summary, as shown by the average fit metrics in Table 3.6, on the whole, the model goodness-of-fit metrics are extremely good for the SHP clusters. The model accounts for the clusters in a straightforward way: The first two clusters differ in terms of visual parameters and OopsER but use the same strategy; the third cluster reflects some parameter differences, but notably uses a strategy the trades higher ER for faster RTs.

Table 3.6

	Shape			NFix	OopsER	GoF: RT			ER		
SHP Cluster	Avail	CrPr				r^2	*aare*	*aae*	r^2	*aare*	*aae*
129:Slow/LowER	0.375	0.025		U	0.01	0.96	9%	101	0.92	15%	0.002
37:VerySlow/MedER	0.6	0.05		U	0.005	1.00	3%	41	0.94	43%	0.005
458:NegAcc/HighER	0.425	0.100		8	0.023	0.94	6%	80	0.99	13%	0.004
Average fit metrics						0.97	6%	74	0.95	24%	0.004

CSF Clusters

Figure 3.23 shows the three groups of subjects resulting from the cluster analysis of the CSF subjects. As before, each column is a cluster with RT on the top and ER on the bottom; red curves are Positive trials, black curves or Negative trials. The curves for different subjects are shown by different plotting points (circles, triangles, squares, diamonds). The clusters are shown in order of increasing ER from left to right. Each cluster is labeled in terms of the subject ID numbers in that cluster and a short description. The plotting scales cover the smaller range for RT and ER used for most graphs in this report.

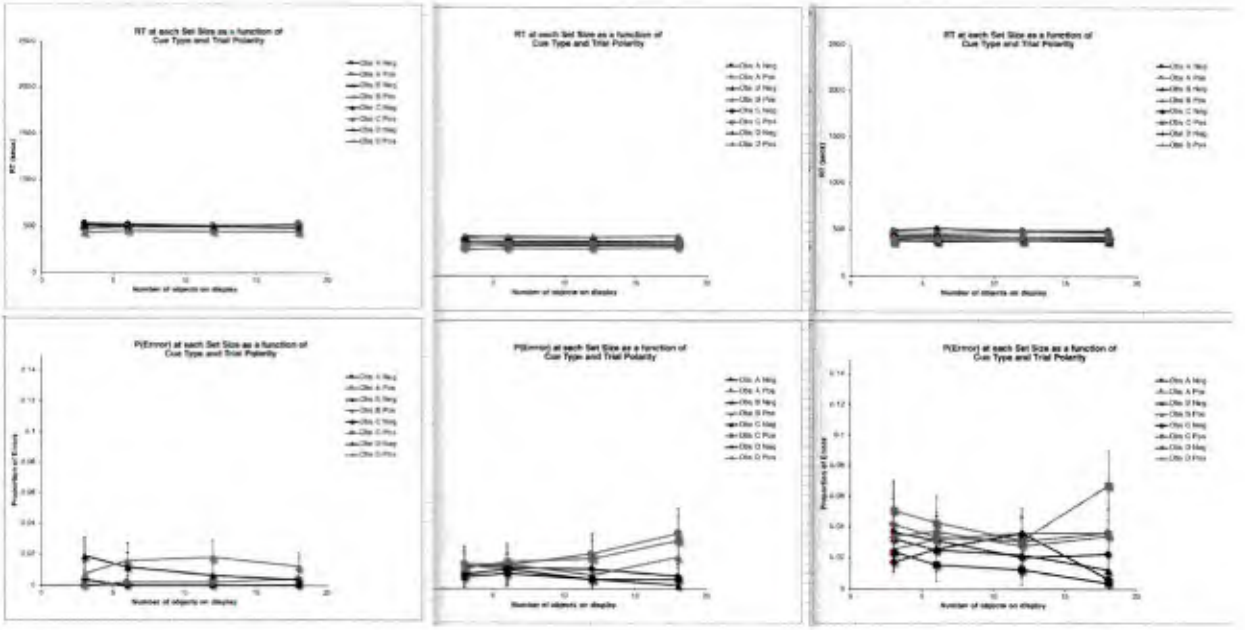

Figure 3.23. Individual subject RT and ER in each cluster of the CSF subjects. The 95% confidence intervals are based on the approximately 500 trials for each individual subject's RT and errors underlying each plotted data point.

All of the clusters have fast RTs (about 500 ms) that are quite flat with Set Size and very similar for Positive and Negative trials. Reading from the left, the first cluster 48:FlatSlow/ LowER contains two subjects who have the slowest RT and very low ER that is essentially unaffected by set size. The second cluster 356:FlatFast/MedER of three subjects has fast RTs and low and flat Negative ER, and Positive ER that increases with set size. The rightmost cluster 1279:FlatFast/HighERof four subjects shows much higher and varied ERs but show little overall effect of set size.

Figure 3.24 shows the average RT and ER for each of these three clusters; and Table 3.3 shows the statistics for each cluster based on those means. The slope ratios are not meaningful because the Positive and Negative RT slopes are very small and noisy. As before, note how the basic pattern of the effects for the individual subjects in the cluster is reflected in the means for that cluster, and how the confidence intervals for the data points in a cluster are fairly tight.

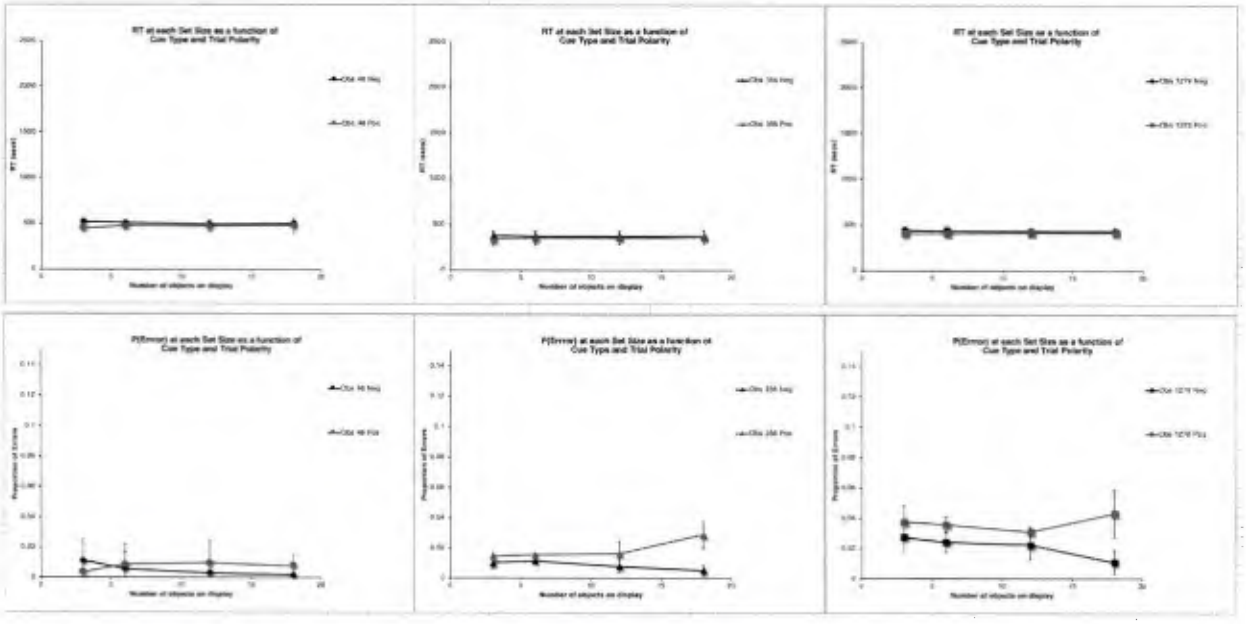

Figure 3.24. Mean RT and ER in each cluster of the CSF subjects. The 95% confidence intervals on each plotted point are based on the mean RT and ER for that data point for each of the 2-4 subjects included in the cluster.

Table 3.3

	Negative				Positive					Slope ratio
CSF Cluster	Intercept	Slope	r^2	ER	Intercept	Slope	r^2	ER	ER Max	
48:FlatSlow/LowER	520	-1.30	0.58	0.01	456	1.49	0.70	0.01	0.01	-0.87
356:FlatFast/MedER	375	-0.24	-0.32	0.01	341	1.08	0.89	0.02	0.03	-0.23
1279:FlatFast/HighER	440	-0.78	-0.86	0.01	341	1.08	0.89	0.02	0.043	-1.00

The model was fit to each cluster following the same approach described for the overall average data. As shown in Figure 3.25, for all three clusters, the Fixed-eye strategy provided a good fit, with only the Color availability parameter and the Oops ER adjusted to fit each cluster. Table 3.4 provides the parameter values, where 0 for *NFix* (the number of allowed fixations) represents the Eyes-Fixed strategy. As discussed above for the overall average data, the Color crowding probability parameter makes a negligible effect; it was set to a place-holder value of 0.025 to be compatible with the COC models. As in the overall CSF model, once Oops ER was set to the False Alarm ER, the Miss ERs were accounted for by adjusting the availability parameter. Note that because the predicted and observed RTs are flat and similar for Positive and Negative trials, the r^2 for RT fits is constrained to be very small — when there is no variability in the observed data, then there is no variance for the prediction to account for! The RT fit for 48:FlatSlow/LowER could be improved by setting the VDelay parameter to a higher value. But even without this adjustment, as shown in Table 3.5, the fits for CSF clusters are overall very good.

The account for the CSF clusters is thus very simple: All subjects followed the Fixed-Eye strategy, which made the RTs flat and fast, but differences in Oops ER and Color availability produced distinct levels of Miss ER. As with the average data model, the crowding probability parameter had a negligible effect.

Table 3.4

	Color				NFix	OopsER	GoF: RT			ER		
CSF Cluster	Avail	CrPr					r^2	*aare*	*aae*	r^2	*aare*	*aae*
48:FlatSlow/LowER	0.09	0.025			0	0.004	0.05	18%	89	0.10	48%	0.002
356:FlatFast/MedER	0.10	0.025			0	0.010	0.08	11%	38	0.63	25%	0.003
1279:FlatFast/HighER	0.115	0.025			0	0.020	0.10	5%	22	0.69	22%	0.005
Average fit metrics							0.08	11%	50	0.47	31%	0.003

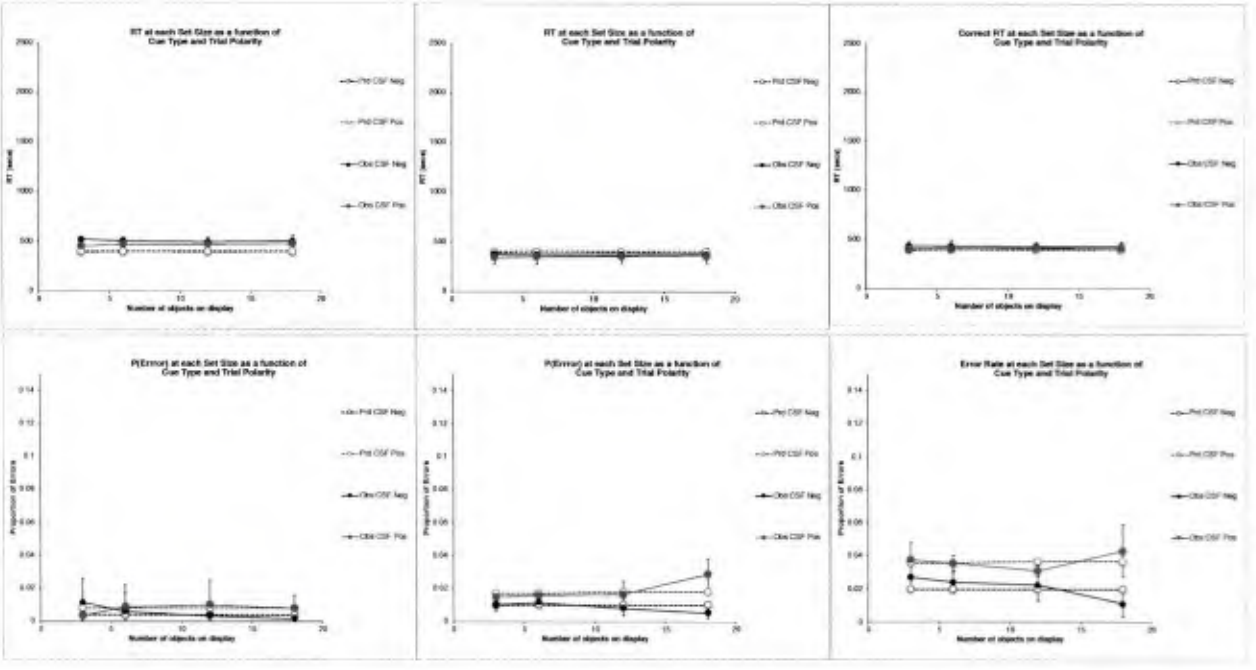

Figure 3.25. Observed (solid lines and points) and Predicted (dotted lines, open points) for the CSF Cluster Means. Upper panels: RT; Lower panels: ER. Left panel: 48:FlatSlow/LowER. Middle panel: 356:FlatFast/MedER. Right panel: 1279:FlatFast/HighER.
Sources: CSF_48_Vm2eLS9c_09_025_004_0_NC_200409, CCSF_356_VM2eLS9c_1_025_01_0_NC_200416, CSF_1279_VM2eLS9c_115_025_02_0_NC_200416

COC Clusters

Figure 3.26 shows the groups of subjects resulting from the cluster analysis of the COC subjects; again, each cluster is shown in a column with RT above and ER below, and are shown in order of increasing ER from left to right, with the exception of the single-subject cluster at the far right. The first cluster 310:Sloped/LowER contains two subjects as very low ER and RT that shows the classic linear RTs with greater slope for Negative than Positive. The second cluster, 1279:AlmostFlat/MedER, contains four subjects with almost flat RTs and Negative ER, but with Positive ER increasing with set size. The third cluster 458:AlmostFlat/HighER has three subjects who also have almost flat RTs but with very high ERs, both for Negative and Positive, with a tendency for Positive ER to increase with set size. The right-most single-subject cluster has RTs very similar to the first cluster and was grouped in that cluster by the analysis; however, the ERs are much higher and trended more with set size than the first cluster. To meet the criterion of not producing misleading mean values for a cluster, this subject was moved out of the first cluster to become a single-subject cluster of their own.

Figure 3.27 shows the average RT and ER for the first three clusters, and Table 3.7 shows the statistics for these average data. Notice that the two AlmostFlat clusters have very similar RT slopes, but very different ERs. However, it is important to note that the third cluster 458:AlmostFlat/HighER is much more heterogenous than the other clusters in in this data set, and accordingly the confidence intervals around this average data are fairly large.

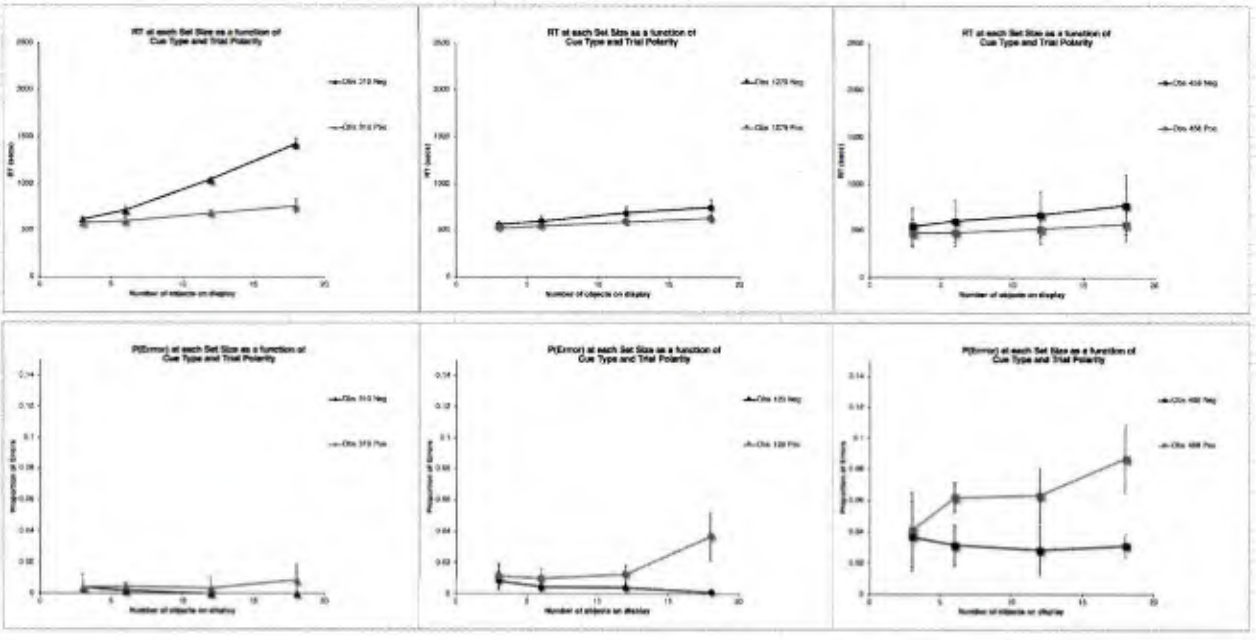

Figure 3.27. Mean RT and ER in each cluster of the COC subjects. The 95% confidence intervals on each plotted point are based on the mean RT and ER for that data point for each of the 2-4 subjects included in the cluster.

The model fits for the COC clusters are shown in Figure 3.28, and the parameters and goodness-of-fit statistics are listed in Table 3.8. The plotting scales are the same as those used for

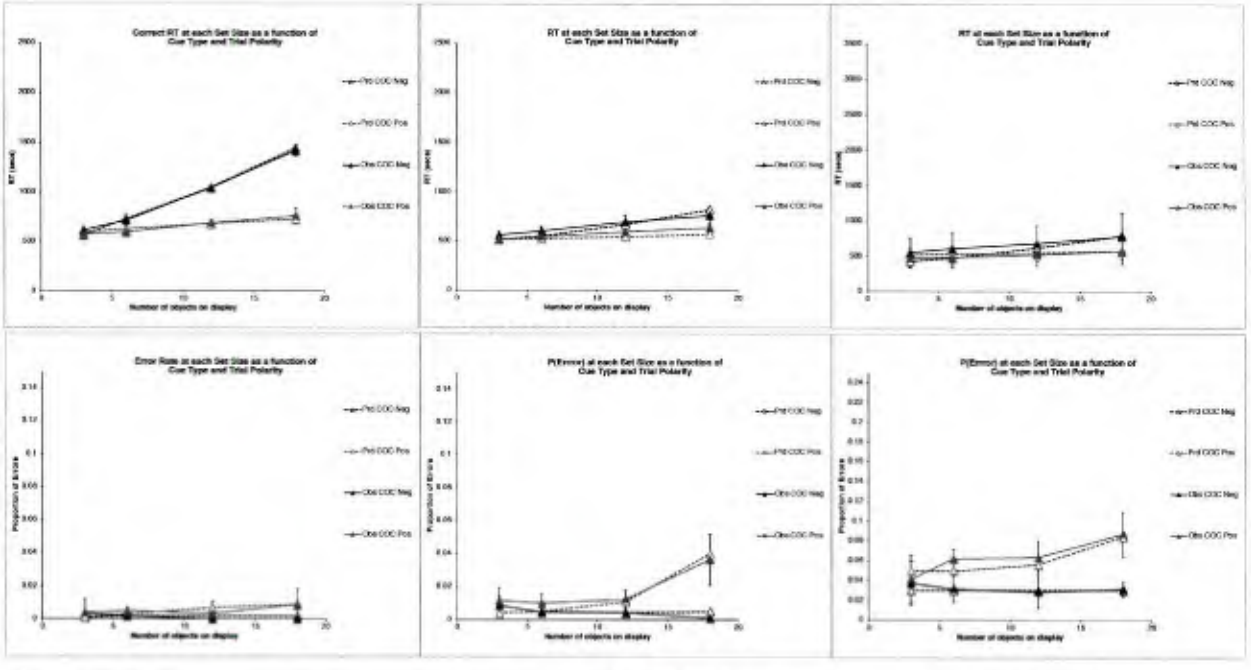

Figure 3.28. Observed (solid lines and points) and Predicted (dotted lines, open points) for the COC Cluster Means. RT: Top panels, ER: Bottom panels. Left panel: 310:Sloped/LowER. Middle panel: 1279:AlmostFlat/MedER. Right panel: 458:AlmostFlat/HighER.
Sources: COC_310_VM2eLS9c_15_025_25_05_0015_99_CP_200309, COC_1279_VM2eLS9c_1_075_15_075_0045_3_1_1_99_CP_200416, COC_458_VM2eLS9c_1_075_15_075_05_03_3_1_1_99_CP_200416.

the CSF clusters and most of the other graphs. All of the models use the Basic Search with Confirm-Positive strategy, but differ in whether fixations are limited.

The top panel in Figure 3.28 shows the predicted and observed RTs and ERs for the 310:Sloped/LowER, cluster which shows very sloped and very linear RTs and very low ER. The model for this cluster fits extremely well. As noted above, a very low observed ER can lead to a very large *aare* value, in this case, infinitely large since a couple of the observed ERs are actually zero; accordingly, this cell in Table 3.8 includes only the cases where the observed ER is greater than zero. The *aae* provides a good indication that the predicted values are very close to the observed values. Apparently these subjects had some difficulty seeing the Color and Orientation, as shown by the parameter values, and so adopted a methodical Basic Search with Confirmation strategy with unlimited fixations, and thus minimized errors, both with a very small Oops ER, and a strategy that produces very few Miss errors as well. Note that this strategy choice is different from that required to fit the overall COC data.

The other two clusters used the same strategy as the model for the overall data in which the fixations were limited to 3 to produce both fast and fairly flat RTs and Miss rates that increase with Set Size. The middle panel shows the fit to cluster 1279:AlmostFlat/MedER. These subjects had very good Color and Orientation availability, leading to fairly fast reaction times overall, and relatively large crowding probabilities, which produces more Miss errors than in the first cluster, but very few False Alarm errors, thanks to the Confirmation step in the strategy. In this portion of the parameter space, the negative acceleration of RTs observed in one of the SHP clusters from limited fixations does not appear. While in terms of the goodness-of-fit metrics, the model is a good fit, there are small but visible discrepancies in the RT slopes. A better fit to this cluster could not be obtained with the current model mechanisms.

The bottom panel shows cluster 458:AlmostFlat/HighER. As shown in Table 3.8, the visual parameters and number of fixations limit is the same as the previous cluster fit. While the RT slopes are slightly different, the same parameters provide a satisfactory fit and cannot be improved on. The distinct feature of this cluster is the extremely high error rate. Apparently these subjects were very much less careful than the previous one; if the False Alarm ER is used to estimate the Oops ER at 0.03, then the other parameters, which fit the RTs reasonably well, cause a consistent underestimate of the Miss ER. The fit shown here takes the unusual step of setting a separate Oops ER for Positive trials, namely 0.05, which produces a reasonable match. Perhaps this oddity is due to the unusually heterogenous character of this cluster, as shown in the above individual subject graphs; in fact the subject with the highest ER also has RTs that are very fast and almost identical for both Positive and Negative trials. Refining this cluster might produce a clearer result.

The average goodness-of-fit metrics in Table 3.8 for these three clusters are on the whole are fairly good, taking into account that a couple of the ER *aare* values are very high due to very small observed ERs. It seems that most of the COC subjects tried to be very fast, as shown by the small number of allowed fixations, and do not compare well to the first cluster who were very slow and careful, showing that it is possible to do the conjunction task with high accuracy. One possibility is that the faster subjects adopted a mixture strategy, sometimes doing an Eyes-Fixed

trial, and the rest of the time doing a limited fixations search with confirmation. Such a mixture would flatten the RT functions and increase the ER. In the interest of simplicity, such strategy mixtures were not tested in this work.

As discussed above with the overall COC modeling, the COC task is inherently more complex in its strategy requirements due to how crowding effects render the two-feature stimuli ambiguous to an extent not involved in the single-feature tasks. Subjects responded to this strategy complexity in different ways. Some subjects buckled down and did the task carefully and with high accuracy, and their RTs reflect this. Other subjects did a "quick and dirty" task strategy of just making a few eye movements before halting and accepting a high error rate, and in one cluster, apparently producing many more Oops Miss Errors. The greater heterogeneity of performance in COC was likely aggravated by the fact that the experimental procedure, with its lack of speed-accuracy feedback, did not encourage subjects to occupy a small portion of the speed-accuracy space, but rather allowed them choose very freely.

Table 3.8

	Color		Orientation		NFix	OopsER	GoF: RT			ER		
COC Cluster	Avail	CrPr	Avail	CrPr			r^2	*aare*	*aae*	r^2	*aare*	*aae*
310:Sloped/LowER	0.15	0.025	0.25	0.05	U	0.0015	0.99	3%	24	0.55	43%	0.002
1279:AlmostFlat/ MedER	0.10	0.075	0.15	0.075	3	0.0045	0.90	6%	40	0.90	76%	0.003
458:AlmostFlat/ HighER	0.10	0.075	0.15	0.075	3	0.03/.05	0.61	10%	55	0.91	11%	0.005
Average fit metrics							0.83	6%	40	0.79	43%	0.003

Summary of model fits

Table 3.9 shows a summary of the goodness of fit for the presented models, both the model for the average subject data in the three tasks, and averaging the fit metrics over the clusters in each task condition. Not included in these averages are the zero CSF r^2 values for the flat RT curves in CSF. The RT is fit extremely well, with very high r^2 values and *aare* under 10% and *aae* on the order of only 50 ms. ER, being intrinsically less stable in these data, was fit not as well, especially when the different clusters differed in their error patterns; however in terms of absolute error *aae*, the predicted values were off by rather less than half a percentage point.

Table 3.9

	GoF: RT			ER		
Modeled data	r^2	*aare*	*aae*	r^2	*aare*	*aae*
Subject averages in all three tasks	0.99	6%	49	0.95	19%	0.003
Average over tasks of clusters within tasks	0.90	8%	55	0.74	33%	0.003

According to the architecture, subjects could differ in architectural parameters and/or their task strategy. In the presented fits, there were two perceptual parameters for each property: the availability slope and the crowding property, and an Oops ER, estimated from the False Alarm ER, except for one cluster, which seemed to require a separate Miss Oops ER. There were two fundamentally different task strategies, the Fixed-eye strategy, required for the CSF condition, and the Basic Search strategy for SHP and COC, which in some cases required a couple of options, one being a limit on the number of fixations, and the other a Confirmation step before responding. In the subject clusters in these tasks, there was some variation on the strategy choices. The choice of strategies and parameters produced a good set of fits for the overall average data in the three task conditions, and the three clusters of subjects that appeared in each of those three conditions.

Conclusions and future work on modeling of simple visual search tasks

The main reason for undertaking simple visual search was to test the generality of the architecture in a domain that has seen considerable empirical and theoretical work in cognitive psychology and has been used to justify an attention-centric fundamental architecture. These models of simple visual search tasks show that the basic concepts of the EPIC architecture are applicable to a class of visual search tasks that has not been previously modeled in EPIC. Rather than ill-specified attentional mechanisms proposed in the literature, performance is determined by low-level mechanisms of vision, visual memory, oculomotor characteristics, and task strategy. Furthermore, the analysis of the subject clusters and the models for them are an excellent demonstration of the frequently-ignored problem with human performance experiments: If the methodology does not incentivize subjects to approach the task in any particular way, subjects will devise their own strategies for the task, and their performance will reflect some combination of their own architectural characteristics and this strategy. Too often, researchers have simply swept this issue under the rug and controlled neither for strategy differences nor examined whether individual subject performance was consistent with their theories, which often failed to even take the strategy options seriously. The present results thus not only contribute to understand how the visual architecture and strategy choice plays a role in visual search tasks, but also demonstrate a critical methodological requirement for future work in this area.

Modeling simple visual search tasks is not as practically useful as good models of complex search, so the future work on this target will be focussed on developing the general architectural mechanisms that would play a role in complex search as well. The models contained in this report, the explanatory sequence justification of them, the final results on the individual difference clusters, is a substantial body of innovative work on simple visual search. This is being prepared as a substantial technical report to be made publicly available as well as supplied to ONR, and will also be submitted for publication.

Complex Visual Search

Background. Many military tasks are display intensive in that they involve using a display showing many objects with color and shape coding to perform a complex task; an example is the radar displays used in CIC stations. Previously, EPIC was used to construct models for how such displays could be searched, and basic concepts were added to the architecture, such as acuity functions for different visual attributes. For example, compared to shape, color can be detected in smaller objects further out in peripheral vision (at greater eccentricity). The first modeled dataset of this type was displays of 48 military-spec icons (Kieras & Marshall, 2006); the models were precursors of the one described here. Another dataset modeled in this effort was the classic Williams (1967) study that used motion-picture-film methodology to record eye movements while subjects searched a display of 100 objects for the one that matched a specification for some combination of color, size, and shape. Kieras (2010) and Kieras & Hornof (2014) presented models for the Williams task data.

The models for these tasks captured the basic effects of the different search cue conditions. If

color is a search cue, most fixations are to objects of that color; a weaker effect is observed for object size as a cue; object shape is remarkably weak cue, and the text label is even less effective. This effect is called *visual guidance*. The efficiency of the search in terms of number of fixations to find the target follows the same pattern. In the models, the task strategy chooses the next object to examine based on which visual properties are "available" or visible given the current eye position. Since color is visible at greater eccentricities than shape, fixations will be made to a matching color more often than to a matching shape. Providing more effective search cues in turn reduces the number of fixations, and the time, required to find the target.

However, the Williams study did not report some key aspects of the data such as the effect of object size and how it would interact with search attributes. In addition, modern eye movement methodology now provides considerably more reliability and detail on the properties of eye movements. Visual search eye movement data in a Williams-like task was collected by Anthony Hornof and Yunfeng Zhang of the University of Oregon (Zhang & Hornof, 2013) and has been analyzed and modeled in a collaborative effort. The Hornof-Zhang dataset provides a high-quality fixation-by-fixation collection of eye movement traces that can be used to further test and refine the earlier EPIC models of visual search.

The Complex Visual Search Experiment. The task was to locate a target object in a field of seventy-five distractor objects. Each object on the display had a unique two-digit number and a unique combination of color, size, and shape. Participants were precued with the number of the target, and some combination of the target's color, size, and shape. An example display is shown in Figure 3.29. See the 2017 Progress Report for more details on the experimental task.

Current Status of Complex Visual Search Work

During the prior reporting periods, a considerable amount of work was done on an extensive re-analysis of the data set to remove certain artifacts present in the original data set. For example, the number of fixations required to complete the search is an important performance metric, so it

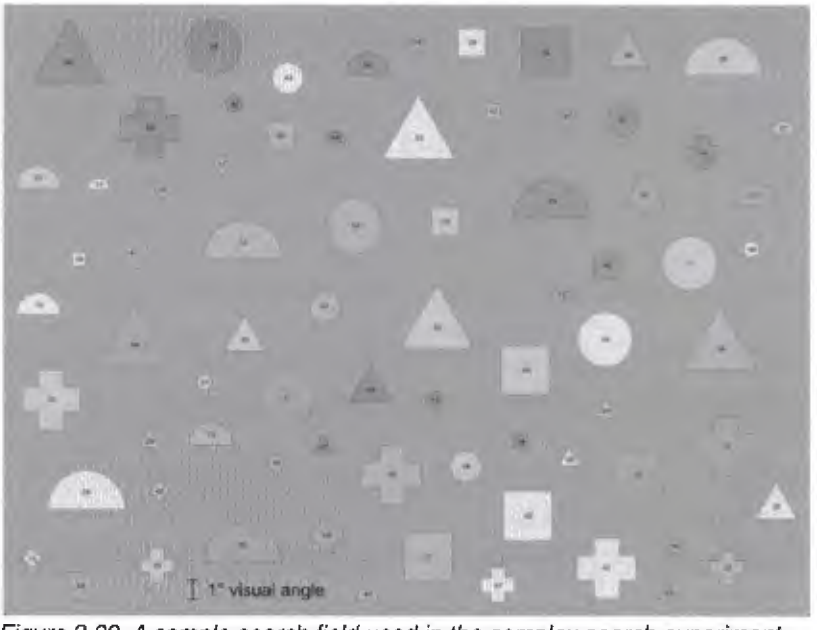

Figure 3.29. A sample search field used in the complex search experiment.

is essential to have a stable and well-behaved definition of what counts as a fixation. One issue was that sometimes subjects made anticipatory eye movements before the display appeared, so there was a subtle problem of identifying the true first fixation in a trial. Another issue was determining whether multiple successive fixations were made on the same object; this required both careful refinement of the algorithm for identifying saccades, and a new definition of a corrective saccade, in which a first fixation on or near an object was followed by a second fixation on the same object. A fairly general issue was that sometimes the limited temporal resolution of the eye tracker resulted in saccade durations which were implausible; such saccades were excluded from calculations of average saccade distance and duration.

The dataset is now very stable and a report on the basic empirical effects is being prepared. We plan to make this very high-quality and refined dataset publicly available for download; the experience with the Wolfe, et al. dataset shows that this can be a valuable contribution to the field.

This dataset is a gold-mine of many different metrics of visual search performance; for example, saccade distance has not been explored before in the visual search literature, but it should be an indicator of search strategy. As described in the 2017 Progress Report, many of the effects in this dataset have been successfully modeled with high values of r^2, but accounting for the effects on multiple performance metrics simultaneously with a single strategy and set of parameters has not yet been achieved. The model strategies have become "crufty" in that they contain a convoluted mixture of mechanisms added on to the basic strategy in an effort to account for various aspects of the data.

Future Work on Models for Complex Visual Search

The proposed work was to determine whether an additional layer of low-level vision mechanisms, such as proto-objects, should be added to the architecture to account for important results. The strategy for pursuing this goal was to determine it certain important effects could be explained in terms of existing simple mechanisms in the architecture. The work with the simple visual search models indicates that crowding effects, an important visual mechanism, should be represented in the early-vision architecture; it is possible that crowding effects actually explain the phenomena that seem to call for concepts like proto-objects. It would seem like crowding might not play a visible role in the complex search stimuli because the amount of crowding is basically constant due to the high and relatively consistent density of the displays. But some of the more difficult-to-model effects in the data might be a result of crowding in addition to saccadic noise.

Note that all of the complex visual search modeling work has dealt with average data. Given the experience with the Wolfe, et al. dataset, it is possible that individual differences might result in the average data being too distorted to be fit by a single model. Perhaps cluster analysis would yield similarly fruitful results. The complex search dataset is a within-subject design and so individual parameters and strategies might be easier to identify. For example, heatmaps showing the locations of the first few fixations in a trial reveal that some subjects tend to start their search from the central fixation point and work outwards as expected, but others start at the upper-left corner of the display, corresponding to "reading order".

The collaboration with Anthony Hornof (University of Oregon) and analyzing and modeling

in detail the eye movement data from a complex visual search task will continue, having been paused while EPIC modeling work supported by a collaborative NSF grant was completed (now delayed by COVID-19). But we will complete a technical report on the complex visual search task data, and another on modeling it. Then a least one publication will be prepared from those reports.

Goal 4. A Preliminary Model of Auditory-Visual Integration

Background

In the introduction it was pointed out that the development of the auditory and visual architectures in EPIC were guided by the same principle, namely to attribute performance to perceptual issues as much as possible, rather than arbitrary cognitive mechanisms. Because the architecture allows task strategy to be explicitly represented in a simple way, it is is possible to work with alternative cognitive strategies, as well as different perceptual mechanisms, to help distinguish underlying perceptual abilities from task-specific strategy effects.

But beyond this methodological commonality, we proposed some preliminary work on how information in the visual and auditory modalities might be combined, or jointly processed, so as to model tasks in which both modalities are involved. The task domain for this work is strictly limited to tasks involving localized auditory cues in simple visual search tasks.

Current status

The proposed work depends on arriving at a stable model for auditory localization effects, which has not yet been achieved, so there is no progress to report under this goal at this time.

A small-scale effort on this topic might be possible. Previous work with NSMRL involved modeling the results of an Audio-Visual Integration task in which localized sound was used to cue the location of a target for visual search. Based on preliminary presentations of the eye-tracking data from that experiment, the auditory localization was not being used to guide eye movements as precisely as the model assumed; it would be interesting to see if a more accurate model could be devised that would explain whether the observed performance was limited by lack of resolution in the localization, or some other cause. This work would be feasible if the eye-movement data was available, which at this time, is still not the case.

References

Abrams, R.A., Meyer, D.E., & Kornblum, S. (1989). Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the oculomotor system. *Journal of Experimental Psychology: Human Perception and Performance*. 15 (3), 529-543.

Anstis, S.M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision research*, 14, 589-592.

Arbogast, T.L., Mason, C.R., & Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America*. 112(5), 2086–2098.

Bolia, R., Nelson, W., Ericson, M., and Simpson, B. (2000). A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America*. 107, 1065–1066.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, **226**, 177–78.

Brungart, D.S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*. 109 (3), 1101-1109.

Brungart, D.S., Chang, P.S., Simpson, B.D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America*, 2006, 120(6), 4007-18.

Buetti, S., Cronin, D.A., Madison, A.M., Wang, Z., & Lleras, A. (2016). Towards a better understanding of parallel Visual Processing human vision: Evidence for exhaustive analysis of visual information. *Journal of Experimental Psychology: General*. 145 (6), 672-707. http://dx.doi.org/10.1037/xge0000163

Buus, S. (1999). Temporal integration and multiple looks, revisited: Weights as a function of time. *Journal of the Acoustical Society of America*, 105, 2466.

Carney, L. (2019). https://www.urmc.rochester.edu/labs/carney.aspx

Carpenter, R.H.S. (1988). *Movements of the eyes* (2nd ed). London: Pion.

Carrasco, M., & Frieder, K.S. (1996). Cortical magnification neutralizes the eccentricity effect in visual search. *Vision Research*, 37, 63-82.

Cooke M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 2006, 119(3), 1562-1573.

Douven, I. (2017). Abduction. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), URL = <https://plato.stanford.edu/archives/sum2017/entries/abduction/> Retrieved March 19, 2020.

Durlach, N.I., Mason, C.R., Kidd, Jr., G., Arbogast, T.L., Colburn, H.S., & Shinn-Cunningham, B.G. (2003). Note on informational masking (L). *Journal of the Acoustical Society of America*. 113 (6), 2984-2987.

Engel, F. L. (1977). Visual conspicuity, visual search and fixation tendencies of the eye. *Vision Research*, 17, 95–108. https://doi.org/10.1016/0042-6989(77)90207-3.

Findlay, J.M., & Gilchrist, I.D. (2003). *Active Vision*. Oxford: Oxford University Press

Green, D. (1960). Psychoacoustics and detection theory. *Journal of the Acoustical Society of America,* 32, 1189-1203.

Green, D., Birdsall, T. & Tanner, S. (1957). Signal detection as a function of signal intensity and duration. Journal of the Acoustical Society of America, 29, 523.

Green, D. & Swets, J. (1966). Signal Detection Theory and Psychophysics. New York: Wiley.

Gruner, C. & Johnson, D. (2001). Calculation of the Kullback-Leibler distance between point process models. Proceedings of ICASSP, May 7-11, 2001, Salt Lake City.

Harris, C.M. (1995). Does saccadic undershoot minimize saccadic flight-time? A Monte-Carlo study. *Vision Research*, 35, 691-701. Levi, D.M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, 48. 635-654.

Henderson, J.M. & Castelhano, M.S. (2005). Eye movements and visual memory for scenes. In G. Underwood (Ed.), *Cognitive processes in eye guidance*. New York: Oxford University Press. 213-235.

Hillenbrand, J. (2019). Vowel data. accessed April, 2019. http://homepages.wmich.edu/%7Ehillenbr/voweldata.html

Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099-3111.

Hornof, A. J., & Zhang, Y. (2010). Task-constrained interleaving of perceptual and motor processes in a time-critical dual task as revealed through eye tracking. Proceedings of ICCM 2010: The 10th International Conference on Cognitive Modeling, Philadelphia, Pennsylvania, August 5-8, 97-102.

Hornof, A. J., Zhang, Y., Halverson, T. (2010). Knowing where and when to look in a time-critical multimodal dual task. Proceedings of ACM CHI 2010: Conference on Human Factors in Computing Systems, New York: ACM, 2103-2112.

Hulleman, J., & Olivers, C.N.L. (2017). The impending demise the item in visual search. *Behavior & Brain Sciences*, (40). Cambridge University Press. doi:10.1017/S0140525X15002794, e132

Kieras, D. (2009). Why EPIC was Wrong about Motor Feature Programming. In A. Howes, D. Peebles, R. Cooper (Eds.), *9th International Conference on Cognitive Modeling – ICCM 2009*, Manchester, UK.

Kieras, D. (2010). Modeling Visual Search of Displays of Many Objects: The Role of Differential Acuity and Fixation Memory. *The 10th International Conference on Cognitive Modeling – ICCM2010*, August 6-8, 2010, Philadelphia, PA.

Kieras, D.E. (2016). A summary of the EPIC Cognitive Architecture. In S. Chipman (Ed.), *The Oxford Handbook of Cognitive Science*, Volume 1. Oxford University Press. 24 pages. DOI: 10.1093/oxfordhb/ 9780199842193.013.003

Kieras, D.E. (2018). Visual search without selective attention: A cognitive architecture account. *Topics in Cognitive Science*, ISSN: 1756-8765 online, 1-18. https://rdcu.be/bfiJ7 DOI: 10.1111/tops.12406

Kieras, D.E & Hornof, A.J. (2014). Towards accurate and practical predictive models for active-vision-based visual search. In *Proceedings of CHI 2014: Human Factors in Computing Systems. New York: ACM, Inc.*

Kieras, D.E., & Hornof, A. (2017). Cognitive architecture enables comprehensive predictive models of visual search: Commentary on Hulleman & Olivers. *Behavioral & Brain Sciences*, 40, 29-30. doi:10.1017/ S0140525X16000121, e142

Kieras, D.E, Hornof, A., & Zhang, Y. (2015). Visual search of displays of many objects: Modeling detailed eye movement effects with improved EPIC. Poster in *Proceedings of the 13th International Conference on Cognitive Modeling (ICCM 2015)*, Groningen, The Netherlands, April 9-11, 2015.

Kieras, D., & Knudsen, K. (2006). Comprehensive Computational GOMS Modeling with GLEAN. In Proceedings of BRIMS 2006, Baltimore, May 16-18, 2006.

Kieras, D.E, & Marshall, S.P. (2006). Visual Availability and Fixation Memory in Modeling Visual Search using the EPIC Architecture. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 423-428.

Kieras, D.E., & Meyer, D.E. (1995). Predicting performance in dual-task tracking and decision making with EPIC computational models. *Proceedings of the First International Symposium on Command and Control Research and Technology*, National Defense University, Washington, D.C., June 19-22. 314-325.

Kieras, D. & Meyer, D.E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction.*, **12**, 391-438.

Kieras, D. E., & Meyer, D. E. (2000). The role of cognitive task analysis in the application of predictive models of human performance. In J. M. C. Schraagen, S. E. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis*. Mahwah, NJ: Lawrence Erlbaum, 2000. 237-260.

Kieras, D., Meyer, D., & Ballas, J. (2001). Towards demystification of direct manipulation: Cognitive modeling charts the gulf of execution. In M. Beaudouin-Lafon & R.J.K. Jacob (Eds.), *Proceedings of the CHI 2001 Conference on Human Factors in Computing Systems*. New York, ACM. Pp. 128 – 135.

Kieras, D. E., Meyer, D. E., Ballas, J. A., & Lauber, E. J. (2000). Modern computational perspectives on executive mental control: Where to from here? In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 681-712). Cambridge, MA: M.I.T. Press.

Kullback, S & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*. 22 (1), 79–86.

Levi, D.M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, 48, 635-654. doi:10.1016/j.visres.2007.12.009

Leshowitz, B. (1969). Receiver operating characteristics and psychometric functions determined under simple- and pedestal-detection conditions. The Journal of the Acoustical Society of America 45, 1474-1484.

MacPherson, A. & Ackeroyd, M. (2014). Variations in the Slope of the Psychometric Functions for Speech Intelligibility: A Systematic Survey. *Trends in Hearing*, 1-26.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*, 104, 3-65.

Meyer, D. E., & Kieras, D. E. (1999). Precis to a practical unified theory of cognition and action: Some lessons from computational modeling of human multiple-task performance. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII. Cognitive regulation of performance: Integration of theory and application* (pp. 17 -88). Cambridge, MA: M.I.T. Press.

Motter, B.C, & Simoni, D.A. (2008). Changes in the functional visual field during search with and without eye movements. *Vision Research*, 48(22), 2382-2393.

Miller G.A., & Licklider J.C.R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, 1950, 22, 167-173.

van Opstal, A.J, & van Gisbergen, J.A.M. (1989). Scatter in the metrics of saccades and properties of the collicular motor map. *Vision Research*, 29(9), 1183-1196.

Põder, E., & Wagemans, J. (2007). Crowding with conjunctions of simple features. *Journal of Vision*, 7(2):23, 1–12, doi:10.1167/7.2.23.

Pelli, D.G., & Tillman, K.A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, 11(10), 1129-1135. doi:10.1038/nn.2187.

Penner, L. (1978). A power law transformation resulting in a class of short-term integrators that produce time-intensity trades for noise bursts. *Journal of the Acoustical Society of America*, 63, 195-201.

Praat, P. & van Heuven, V. (2001). Speak and UnSpeak with Praat. *Glot International*, 5 (9/10), 341-347.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, 2, 437–57. doi: 10.1146/annurev-vision-082114-035733

Sachs, M., Winslow, R. & Sokolowski, B. (1989). A computational model for rate-level functions from cat auditory nerve fibers. *Hearing Research*, 41, 61-70.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001).Virtually perfect time-sharing in dual-task performance: Uncorking the central cognitive bottleneck. *Psychological Science*, 2001, 12, 101-108.

Shaw, E.A.G. (1974). Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *Journal of the Acoustic Society of America*, 56(6), 1848-1861.

Thompson, E.R., Iyer, N., Simpson, B.D., Wakefield, G.H., Kieras, D.E., & Brungart, D.S. (2015). Enhancing listener strategies using a payoff matrix in speech-on-speech masking experiments. *Journal of the Acoustical Society of America*. 138(3), 1297-1304.

Townsend, J.T., & Wenger, M.J. (2004). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review*, 11(3), 391-418.

Treisman, A. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97-136.

Viemeister, N.F. & Wakefield, G.H. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America*, 1991, 90, 858-865.

Viemeister, N.F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *Journal of the Acoustical Society of America*, 66, 1364-1380.

Virsu, V. & Rovamo, J. (1979) Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, 37:475–94.

Wakefield, G. H. (1994). Temporal integration, multiple looks, and signal uncertainty. *Spring meeting of the Acoustical Society of America*, Cambridge, MA.

Warren, R.M. and Bashford, J.A. Perception of acoustic iterance: Pitch and Infrapitch. *Perception & Psychophysics* (1981) 29: 395.

Wertheim, A. H., Hooge, I. T. C., Krikke, K., Johnson, A. (2006). How important is lateral masking in visual search? *Experimental Brain Research*, 170, 387-402. DOI 10.1007/s00221-005-0221-9

Williams, L.G. (1967). The effects of target specification on objects fixated during visual search. In A.F. Sanders (Ed.) *Attention and Performance*, North-Holland. 355-360.

Wolfe, J. M. (2014). Approaches to Visual Search: Feature Integration Theory and Guided Search. In A.C. Nobre & S. Kastner (Eds), T*he Oxford Handbook of Attention*. Retrieved from DOI: 10.1093/oxfordhb/ 9780199675111.013.002.

Wolfe, J.M., Cave, K.R., & Franzel, S.L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3), 419-433.

Wolfe, J.M., Palmer, E.M., Horowitz, T.S. (2010). Reaction time distributions constrain models of visual search. *Vision Research*, 50, 1304-1311.

Yashar, A., Xiuyun, W., Jiageng, C., & Carrasco, M. (2019). Crowding and binding: Not all feature-dimensions behave the same way. *Psychological Science*, September 2019. DOI: 10.1177/0956797619870779

Yee, V. and Wakefield, G. H. (1997). "Modeling the auditory re-set mechanism in a decision-based approach to temporal integration," *Abst Midwinter Res Mtng, Assoc Res Otolaryngol.*

Yost, W. A. (1997). The cocktail party problem: Forty years later. In R. Gilkey & T. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments*. Mahwah, NJ: Erlbaum.329–348.

Zhang, Y., & Hornof, A. J. (2013). The effect of target specification and visual acuity on objects fixated during visual search (Tech. Rep. No. CIS-TR-2013-03). University of Oregon: Department of CIS Technical Report.

Zwislocki, J., Hellman, R. & Verillo, R. (1962). Threshold of audibility for short pulses. *Journal of the Acoustical Society of America*, 34, 1648-1652.

V. Training

Nothing to report.

VI. Dissemination

In addition to the usual publications and presentations, a major channel for dissemination of our work is through our collaborators. These are:

- Dr. Jeffrey Bolkovsky at NSMRL.
- Anthony Hornof at the University of Oregon, and Yunfeng Zhang, now at IBM Research.
- AFRL also granted access to the AF HPC Clusters through MindModeling.org which we have used for model-fitting (POC: Kevin Gluck, WPAFB, 711 HPW).
- The source code for the current version of the EPIC architecture and some representative models has been made open source with the assistance of the University of Michigan Intellectual Property Office, and may be found on github.

More detail is provided for two collaborations in the auditory work:

Collaboration with AFRL. The collaborative project with Air Force Research Laboratory (AFRL) is being conducted with the 711th Human Performance Wing, Human Effectiveness Directorate (711 HPW/RH), Warfighter Interface Division, Battlespace Acoustics Branch (RHCB) which operates out of the Wright-Patterson Air Force Base (WPAFB). The three AFRL scientists involved are Dr. Brian Simpson, Dr. Nandini Iyer and Dr. Eric Thompson, who, along with Dr. Doug Brungart, have been involved in over 15 years of studies following up on the original Brungart (2001) study. They previously provided technical help in understanding the Brungart data, particularly with respect to unpublished background information pertaining to subject training, instructions and data analysis.

The collaboration over the past year has produced one new line of research into better understanding the dynamics of a Kullback-Leibler classifier as well as a new body of CRM data in which spatial position was manipulated. With respect to the former, Dr. Wakefield spent four weeks at AFRL from May to August, 2016, in which he explored both infrapitch signals, which serve as working models for machinery noise and had been studied by Dr. Wakefield in previous research, and granular sources, which serve as sparse models of complex naturalistic acoustic sources as rain, wind, and brooks. In follow-up visits, Dr. Wakefield and AFRL have been refining several hypotheses for experimental evaluation; this work, in turn, also forms a new source of experimental data for better modeling the processes involved in auditory source formation and tracking. Work during the current funding period continued informally at a much lower effort while Dr. Wakefield developed his granular synthesis/experimental tools at Michigan.

Collaboration with NRL/NSMRL. In the past, the auditory model, particularly the stream-tracking component, has been of interest to Dr. Paul Bello and Dr. Will Bridewell at NRL and their ARCADIA project. They had been kept informed of our progress, including the Kullback-Leibler classifier and our computationally-efficient approach to modeling the auditory front-end.

The research direction at NRL shifted away from sensory matters and the primary line of questions related to ARCADIA and audio became a new line of research at NSMRL. We have had a series of fruitful conversations with Dr. Bolkhovsky at NSMRL over the past year and look forward to further collaboration.

VII. Plans

Goal 1. Models of Multitalker Speech Tasks. Over the summer, the research on the self-regulating pooling detector will be prepared for publication and the acoustically-driven models for stream tracking and content detection will be used to re-fit the Thompson et al. (2015) data. Pending the outcome of this work, the research will be prepared for publication. Finally, as it relates to issues concerning late vs. early attention in cognitive architectures, the original "word-based" black-box EPIC account of the CRM data will be prepared for publication.

Goal 2. Modeling the Effects of Spatial Location. In collaboration with Thompson and his colleagues at AFRL, results from a relatively new CRM experiment in which spatial position was varied are available for two- and three-talker conditions. Modeling these data still awaits a more accurate account of the effects of transitioning from two- to three- and four-talker scenarios. As time permits, we will explore this monaural extension, as part of Goal 1, and work on the recent Thompson data, if successful.

Goal 3. Extending the Visual Architecture. The current model for simple visual search, and its theoretical implications will be prepared for publication.

The now-stable dataset for complex visual search will made public along with a technical report and empirical publication on the main findings. Based on the simple visual search work, we will characterize individual differences in the complex task data with cluster analysis before returning to refreshing the models for important effects in the data. These models will first be limited to the four simple search tasks (cues of color, size, shape, and number-only) and then elaborated to cover the full set of conditions, which will also provide a strong validation. The current and additional modeling results will be prepared for publication.

Goal 4. A Preliminary Model of Auditory-Visual Integration. Work on this goal will await developments in Goals 1-3, together with additional results on relevant tasks developed by our collaborators. A possible example of this was noted above; an improved model of the NSMRL Audio-Visual Integration task could be developed if the eye-tracking data was made available.

VIII. Major Problems / Issues

As noted last year, the major issue to be reported is that the new modeling of auditory stream tracking, and accounting for all of the details in the complex visual search modeling, are both proving to be very difficult scientifically. Roughly speaking, we are now working on the "hard part" of the modeling.

A new issue to report is that our planned collaborative work with the NSMRL group was "on hold" through most of this report period due to changes in our point of contact at NSMRL.

IX. Honors

Nothing to report in this period.

X. Technology Transfer

The work with Dr. Qin's group at NSMRL on audio-visual integration had a formal CRADA (Agreement Number NCRADA-NSMRL-13-9182); a renewal of this CRADA was completed with Dr. Jeffrey Bolkovsky at NSMRL (Agreement Number NCRADA-NSMRL-18-10406).

XI. Participants

David E. Kieras, Professor, Electrical Engineering and Computer Science & Psychology, University of Michigan. Approximately 3.3 person-months/year during the reporting period.

Gregory H. Wakefield, Associate Professor, Electrical Engineering and Computer Science, University of Michigan. Approximately 3.3 person-months/year during the reporting period.

XII. Products (Publications)

A. Refereed Journal Articles

Kieras, D.E. (2018). Visual search without selective attention: A cognitive architecture account. *Topics in Cognitive Science*, ISSN: 1756-8765 online, 1-18. https://rdcu.be/bfiJ7 DOI: 10.1111/tops.12406

Kieras, D.E., & Hornof, A. (2017). Cognitive architecture enables comprehensive predictive models of visual search: Commentary on Hulleman & Olivers. *Behavioral & Brain Sciences,* 40, 29-30. doi:10.1017/S0140525X16000121, e142

B. Non-Refereed Significant Publications

None in this report period.

C. Books or Chapters

None in this report period.

D. Technical Reports

None in this report period.

E. Workshops and Conferences

Kieras, D. (2018). Visual Search without Selective Attention: A Cognitive Architecture Account. In *Proceedings of the International Conference on Cognitive Modeling (ICCM 2018)*, Madison, Wisconsin, July 21-24, 2018.

Wakefield, G. H. (2018) EPIC modeling of CRM data, Poisson modeling of temporal integration, and Infrapitch stream formation. Hartmann Lab presentation, Michigan State University, June, 2018.

Wakefield, G. H. (2018). Temporal integration and multiple looks, revisited. Invited paper. 175th meeting of the Acoustical Society of America, Minneapolis, May, 2018. [Note - abstract published but paper could not be presented due to a death in the author's immediate family.]

Wakefield, G. H. and Kieras, D. (2017). Modeling a Two-Talker Listening Task Using the EPIC Cognitive Architecture. Invited talk. Neuroscience Seminar, Psychology Department, Loyola University Chicago, November 14, 2017.

Kieras, D.E. (2017). EPIC lessons for the proposed standard model of the mind. Invited presentation at the AAAI 2017 Fall Symposium on A Standard Model of the Mind, Arlington, VA November 9-11.

Kieras, D. (2017). EPIC Lessons for the Proposed Standard Model of the Mind. Position paper in the Proceedings of the AAAI 2017 Fall Symposium on the Standard Model of the Mind, Arlington, VA, November 9-11, 2017.

Kieras, D.E., Wakefield, G.H., Brungart, D.S., & Simpson, B.D. (2016). A preliminary cognitive-architectural account of spatial separation effects in two-channel listening accounts. *Proceedings of the Human Factors and Ergonomics Society 2016 International Annual Meeting*, Washington D.C., September 19-23, 2016.

F. Patents

None in this report period.

G. Awards/Honors

None in this report period.

Final Report Distribution List

Thomas McKenna
ONR HUMAN & BIOENGINEERED SYSTEMS
875 N. Randolph Street
Arlington VA 22203-1995

ONR REG Office Chicago
Telephone: (312) 886-5423
230 South Dearborn
CHICAGO IL 60604-1595

Defense Technical Information Center
8725 John J Kingman Road Ste 0944
Fort Belvoir, VA 22060-6218

Naval Research Laboratory
ATTN: CODE 5596
4555 Overlook Avenue SW
Washington, DC 20375-5320