Extending and Applying the EPIC Architecture for Human Cognition and Performance: Auditory and Spatial Components

Final Report
Project N00014-13-1-0358

David E. Kieras
University of Michigan

Report No. FR-13/ONR-EPIC-19

Period Covered: 1 JAN 2013 – 31 DEC 2015

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.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1 MAR 2016	Final Report 01-JAN-13 – 31-DEC-15

4. TITLE AND SUBTITLE
Extending and Applying the EPIC Architecture for Human Cognition and Performance: Auditory and Spatial Components

5. FUNDING NUMBERS
N00014-13-1-0358

13PR04643-00

6. AUTHOR(S)
David E. Kieras

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-13/ONR-EPIC-19

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research (Code 341)
875 N. Randolph St.
Arlington, VA 22203-1995

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

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

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

This is the final report for a project that was in a series of projects on the development and validation of the EPIC cognitive architecture for modeling human cognition and performance. This project focussed on extending the architecture 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. Additional work concerned application of the EPIC architecture to Navy research problems.

14. SUBJECT TERMS
Cognitive Architecture, Human Performance Modeling

15. NUMBER OF PAGES
55

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

20160308041

Extending and Applying the EPIC Architecture for Human Cognition and Performance: Auditory and Spatial Components

Final Technical Report

ONR Grant N00014-13-1-0358
Period Covered: 1 JAN 2013 – 31 DEC 2015

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: N00014-13-1-0358

D. Extending and Applying the EPIC Architecture for Human Cognition and Performance: Auditory and Spatial Components

Reporting Period: 1 JAN 2014 - 31 DEC 2015

II. Scientific and Technical Objectives

The basic goal of this project was to develop the EPIC architecture computational architecture for modeling human cognition and performance so that it supports predictive modeling of human-computer interaction involving spatialized audio and speech activity and does so at least as well as it currently supports modeling the visual, manual, and procedural aspects of such tasks. The primary focus of the proposed work was to continue the work on modeling human performance in multitalker speech perception tasks. This project contributed uniquely to the developing capability of human performance modeling to help design maximally effective human-machine systems. The specific project goals were:

Goal 1. The EPIC architecture components for audition and speech perception will be expanded to include a basic stream tracking capability that will provide a robust modeling and prediction capability for tasks involving multitalker speech perception (Goal 1.1), including spatial location of speech sources (Goal 1.2), and spatial perception of signals and cues (Goal 1.3).

Goal 2. The EPIC visual search models and associated architecture components will be revised as needed to take into account new results on visual search such as the effects of moving objects and new detailed results on multiattribute search over large displays

Goal 3. The models and architecture will be developed and tested in applications to militarily-relevant tasks in collaborative arrangements with DoD researchers.

Goal 4. The EPIC architecture software and models will be updated, improved, and made easily available as needed to support this work and that of other researchers.

III. Approach

The technical approach in the project is to develop the auditory architecture by starting with the previously developed visual architecture. Rather than determine how images and waveforms are analyzed by sensory mechanisms, the approach is to determine how to "black-box" low-level sensory and perceptual processing in a way that enables robust and useful models of how the visual, and now auditory, information is used in a complex and practically significant task.

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. However, it must be kept in mind that audition is very different than vision in multiple ways; the need to integrate information over long periods of time is more prominent in audition, and this may impose severe processing constraints that are not seen in vision.

IV. Concise Accomplishments

Goal 1.1. We developed a model for a two-channel listening task and tested it against empirical data; the model includes explicit mechanisms for the perception and tracking of auditory streams. It continues the previous novel line of work that combines the current computation cognitive architecture models with the primarily mathematical and psychophysical theoretical concepts that dominate the literature on audition and speech processing. EPIC's cognitive processor implements a strategy for performing the task, while the auditory processing represents masking phenomena in which interference between two speech streams obscures both word content and stream identification. A stream tracking component is driven by the acoustic properties of the input, which we derived after creating an accurate segmentation of the speech corpus used in the key experiments. The model accounts for the data extremely well, and promises to generalize to the 3- and 4-talker case. We refined our model for the two-talker task by exploring improvements to the stream tracking mechanism.

Goal 1.2. We extended the model to situations in which the speakers are spatially separated – long known to be a powerful cue to speech stream segregation, but surprisingly little understood.

Goal 1.3. To support the spatial separation modeling, we revised the auditory architecture to support auditory spatial perception of azimuth, which also supported modeling work with our NSMRL collaborators on using localized sound to assist visual search.

Goal 2. In another collaboration involving the visual search of complex displays (Kieras &

Hornof, 2014) additional analysis was performed of visual search eye movement data collected by Yunfeng Zhang and Anthony Hornof of the University of Oregon (Zhang & Hornof, 2013). This is a replication of the classic Williams(1967) study that used motion-picture-film methodology to study how people did multiattribute search of displays of many (100) objects. The Zhang-Hornof dataset provides a high-quality fixation-by-fixation collection of eye movement traces that can be used to test and refine the earlier EPIC models of visual search. The analysis showed that key effects predicted by the EPIC models were present in the data, and preliminary models were constructed for these effects. Such studies are especially relevant to display-intensive military tasks such as radar watchstanders.

Goal 3. Multiple discussions and visits have been held with Dr. Michael Qin's group at NSMRL, and a formal CRADA is in place. Work related to Goal 1 and this goal involved further development of auditory localization models, including new data collected at NSMRL.

Goal 4. Some useful extensions and additions were made to the EPIC software. The auditory architecture was stabilized, additional support for mixture models and confusion matrix models was added, and most importantly, the EPIC software was repackaged so that it can run on clusters such as mindmodeling.org under linux/gcc for parameter search and other purposes.

V. Expanded Accomplishments

Goal 1.1: Modeling Multi-Channel Speech Processing

Summary of Previous Work

As before, during this period, the project work was focused on Goal 1. The Annual Reports for 2013-2015 summarizes a set of key accomplishments which involved making 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.

Background

A classic problem in cognitive psychology is the "cocktail party effect" in which a person is surrounded by several people speaking simultaneously, and is nonetheless able to follow a single speaker well enough to maintain a conversation, although some information about what the other speakers are saying appears to be available under some conditions. The early study of these phenomena (e.g. Cherry, 1953; Moray, 1959) defined the current concept of selective attention; the human listener was said to be able to selectively attend to one of the signal sources and "filter out" the others. In the decades since, a body of additional studies and theoretical work has clarified what properties of the acoustic and perceptual situation contribute to the effect (for surveys see Yost, 1997; Bronkhorst, 2000; Haykin & Chen, 2005; Schneider, Li, & Daneman, 2007).

The most common experimental paradigm is that the subject listens to speech from two or more talkers who are speaking simultaneously, and responds to information provided by only one of them, called the *target*, and ignores the information provided by the other talkers, called the *masker(s)*. The research has focussed on characterizing what aspects of the messages contribute to interference between the target and the maskers, both in terms of the perception of sound in general, and of speech in particular. A general psychoacoustic effect is *masking*, in which a sound becomes less perceptible if another sound is simultaneously present. In the case of simple sounds or signals, masking effects are generally considered to be a result of interactions in the cochlea itself, for example, the excitation pattern on the basilar membrane produced by the target sound is disrupted by the masking sound, making it less detectible. In the context of simultaneous speech messages, a distinction is made between *energetic masking*, which refers to the interference produced at the acoustic or sensory level, as in the masking of simple sounds, and *informational masking*, which is interference produced at the higher perceptual and cognitive levels and makes it difficult for the human listener to follow the target message in the presence of the masker message, above and beyond the effects of energetic masking.

Related to informational masking is the concept of *auditory streams* (Bregman, 1990), the notion that the acoustic field sensed by the ears is decomposed by the auditory system into one or more temporally coherent sound sources (for reviews, see Darwin, 1997; Moore & Gockel, 2012). While, in general, the mapping between sources in the acoustic field and those perceived by the auditory system is not one-to-one, in a two-talker task, it is believed that each talker is perceived as a distinct stream, and the listener's task is to determine which sounds go with which

stream and choose the appropriate response. Performance in the two-talker task thus reflects a combination of the energetic masking effects and informational masking effects on stream formation and segregation (Schneider, Li, & Daneman, 2007).

Choice of paradigm and principal data set

Early studies (e.g. Cherry, 1953; Moray, 1959) on two-talker listening involved the *shadowing task*, in which the subject was required to immediately repeat out loud each word of one of the messages, and the accuracy of the shadowing and the memory (or lack of it) of the other message were the primary performance measures. The messages themselves were extended chunks of naturalistic text. However, Spieth, Curtis, and Webster (1954) used messages with a simple fixed structure that asked a question and the subject had to respond with the answer to the specified message using call signs in a radio communication protocol. More recent work has used tasks that were similarly face-valid for practical application, allowed more complete experimental control, and used manual rather than verbal responses. The *coordinate response measure* (CRM) task and speech corpus is a highly simplified form of the command and control messages used in military settings, and have been widely used to provide precise experimental control (Bolia, Nelson, Ericson, & Simpson, 2000) in multi-talker speech experiments.

The CRM corpus is a collection of recorded command utterances in the form of

Ready <Callsign> go to <Color> <Digit> now

spoken by one of four females or four males, where the Callsign, Color, and Digit are drawn from sets of 8, 4, and 8 items, respectively. The corpus was recorded and edited to maintain a high degree of temporal overlap among the spoken Callsigns, Colors and Digits (Bolia, et. al., 2000).

In the two-talker CRM listening task, participants respond to commands by pointing to the appropriate Color/Digit pair on a computer display. A particular Callsign is designated as the Target Callsign, which was always *Baron* in the studies used in this paper. On each trial, a Target message is drawn from those utterances bearing the Target Callsign and is presented simultaneously with a randomly selected Masker message, with the restriction that the Callsign, Color and Digit of the Masker differ from those of the Target. The participant thus hears two messages at the same time, and must choose the color-digit pair associated with the Target callsign, and is instructed to ignore the Masker message. The responses are scored as matching the Target message, the Masker message, or Neither.

An important study by Brungart (2001) stimulated our first modeling. He manipulated the acoustic similarity of the two talkers, varying from Different Sex (DS), to Same Sex (SS), to Same Talker (ST), and also manipulated the relative loudness of the two messages, with a Signal-to-Noise ratio (i.e. the Target-to-Masker ratio) ranging from -12 to +15 dB. The same combined signal was provided to both ears via headphones, making this a *diotic* listening task. This study is important because in addition to reporting the proportion of Both-Correct responses (both Color and Digit are Target), he also reported the proportions of responses that matched Target, Masker, or Neither separately for Color and Digit.

Rather than show his results, however, we present the results for a methodologically

improved replication which is very similar in design and results to Brungart (2001). The replication (Thompson, Iyer, Simpson, Wakefield, Kieras, & Brungart, 2015) followed the conditions and procedures of Brungart (2001) in all respects except two: (1) The SNR, which ranged from -12 to +15 dB in the original study, was shifted to a lower range (-18 to +9 dB) in the interest of studying performance at SNRs closer to masked detection thresholds; (2) the replication clarified the task instructions with a point reward system for correct performance, and provided performance feedback at the end of each trial and during the experiment. The 2013-2015 Annual Reports and Kieras & Wakefield (2014) provide more details about the effects of subject strategy in this paradigm.

Results

Because of the multiple factors and measures involved, these results are somewhat complex. The six panels of Figure 1 show the Thompson et al. results. Each panel plots the proportion of responses that matched the Target, the Masker, and Neither, as a function of the Target-to-Masker signal-to-noise (SNR) ratio in dB. The upper panels display the proportions for Color responses; the lower panels display the proportions for Digit responses. In addition, the panels show the proportion of Both-Correct responses in which both Color and Digit are from the Target message. These black curves are the same in the upper and lower panels. The left-to-right panels

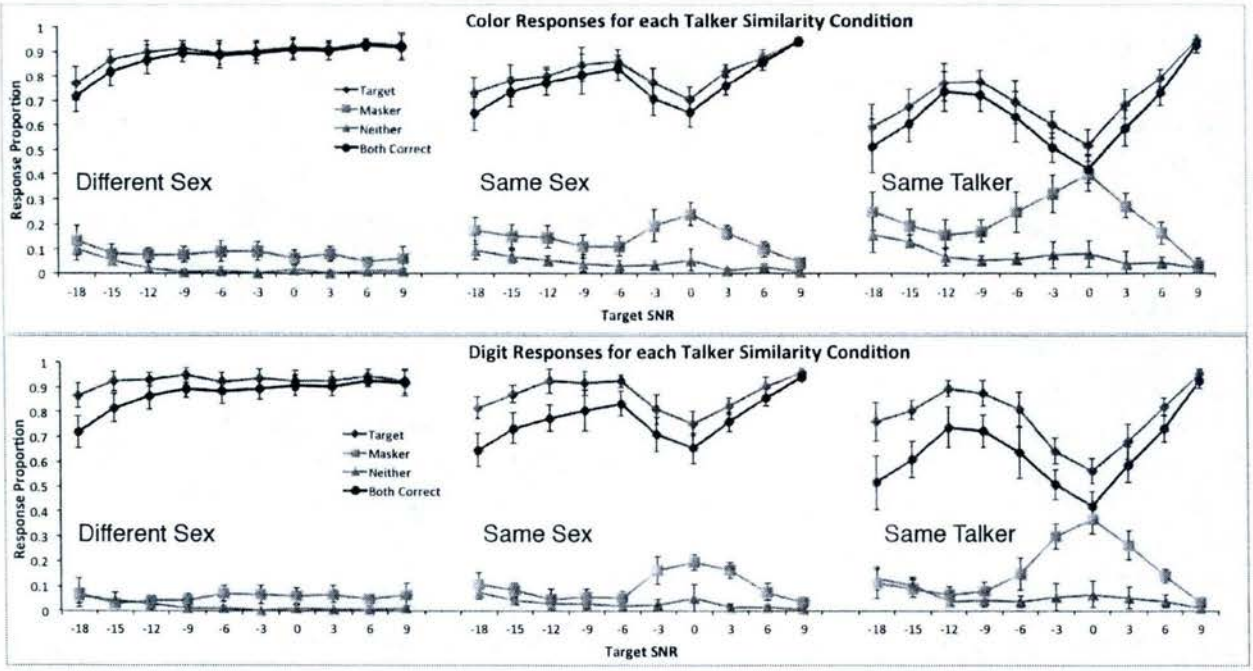

Figure 1. Observed probabilities for each response choice in the Thompson et al. (2015) data as a function of signal-to-noise ratio (SNR) of the target message relative to the masker message. The vertical axis is probability of response; the horizontal axis is Target-to-Masker SNR in dB. Color responses are in the upper panel; digit responses are in the lower panel. The three talker similarity conditions are shown left-to-right in the three subpanels: Different Sex (DS), Same Sex (SS), and Same Talker (ST). Blue curves are for the target color or digit, red for masker color or digit, green for responses that are neither target or masker, and black curves show the probability that both the color and digit are correct (same in both upper and lower panels).

display the results for the similarity of the Target and Masker talkers. From left to right, the stimulus conditions are Different Sex, Same Sex but different talkers, and Same Talker.

The basic effects are as follows: Overall, with increasing positive SNR, the Both-Correct and Target Color and Digit responses are chosen more often, and Masker and Neither responses are chosen less often. The overall performance when the messages are delivered by Different-Sex talkers is better than that for Same-Sex talkers, which is turn is better than that when the two messages are from the Same Talker. For the Same-Sex and Same-Talker conditions, accuracy is very poor at the lowest (most negative) SNRs, but then improves, and then declines again in the vicinity of 0 dB SNR, and then improves again.

A key empirical fact is that the incorrect responses were almost always from the Masker message, which places a basic constraint on the cognitive processes in any model, in that it implies that Masker message content was being perceived and remembered, and then chosen as a response, rather than being simply filtered out, as would be expected from a simple selective attention model.

Accounting for the Phenomena

To date, a theoretical account of the two-talker CRM results remains incomplete. Discussions have focused on the relative importance of informational masking over energetic masking, the roles of selected and divided attention, and the formation and maintenance of auditory streams. However, none of these concepts have been operationalized to the point of providing a quantitative theoretical account of experimental outcomes. The cognitive-architectural modeling work in this project attempts to bridge this gap.

The focus of our work was to account for these results in terms of a basic concept of human cognitive architecture and a quantitative model based on that concept. The resulting model incorporates mechanisms that resemble both energetic and informational masking, but do so with considerably more theoretical precision; most importantly, the strategy that the subject follows to perform the task is directly represented, and this turns out to be critical in accounting for the specific effects in this data.

Basic modeling approach

At the most general level, the significance of our models is that they are apparently the first effort to marry the type of models typically used in audition and speech perception (essentially mathematical psychophysical models based on acoustic characteristics) with the type of cognitive architectural models like EPIC, ACT-R, and Soar, in which the cognitive processor implements a task strategy described with production rules. Applying an earlier lesson from EPIC (Meyer & Kieras, 1999; Kieras & Meyer 2000), even simple tasks can have sophisticated strategies, whose qualitative or logical nature is difficult to capture in conventional mathematical models. In summary, we built a psychophysical front end embedded in a set of information-processing stages, with a cognitive-strategic back end, to give a model that performed the entire task end-to-end. Details of the development of prior models can be found in the 2013-2015 Annual Reports and the Kieras & Wakefield (2014) Technical Report. For brevity, this summary will be limited to the current model.

The EPIC Cognitive Architecture

Since these models are constructed with the EPIC (Executive Process-Interactive Control) cognitive architecture for human cognition and performance, a summary is in order. 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 2 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 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

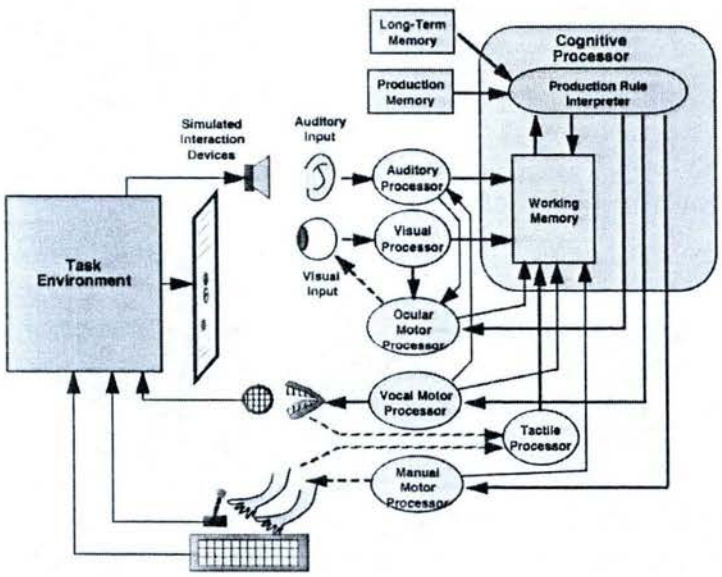

Figure 2. The EPIC architecture in simplified form. The simulated environment, or device, is on the left; the simulated human on the right.

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.

See Kieras (2016) for more discussion of the principles of the EPIC architecture along with a detailed example of its application to visual search tasks. A detailed technical description can be found in Kieras (2004).

Very early EPIC had crude mechanisms for representing auditory input of sound signals and single-channel speech input to support modeling tasks in which localized auditory signals or speech interaction was involved (Kieras, Ballas, & Meyers, 2001; Kieras, Wood, & Meyer, 1997). This project and its predecessors developed the auditory system in more detail to support modeling of multichannel speech processing. Figure 3 shows the basic components of the current auditory system in EPIC.

Constructing the models for multi-talker tasks required additions to the cognitive architecture in the form of more detailed auditory perceptual mechanisms, and then the models for the

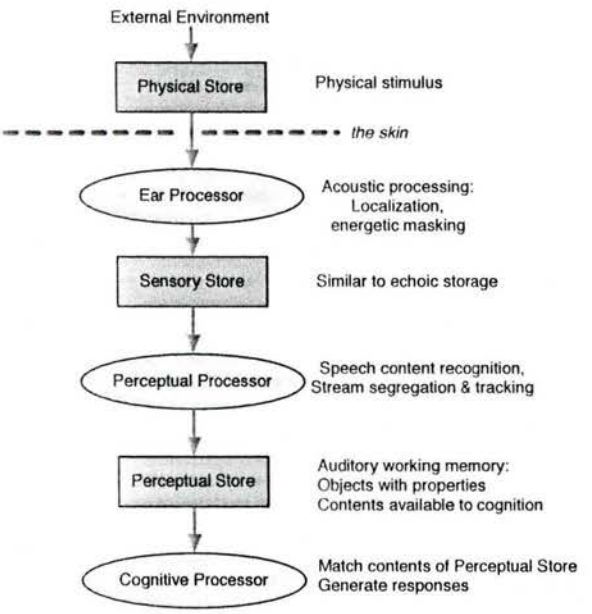

Figure 3. The expanded contents of the auditory processor shown in Figure 2. Low-level sensory acoustic processing is located in the Ear processor, while speech content recognition and stream functions are performed in the perceptual processor.

specific task required the development of a set of production rules for using the information provided by the perceptual system to determine and make the response on each trial. These rules thus implement the task strategy. Complex inference and response choice strategies are easy to represent in the production rules provided by a cognitive-architecture model like EPIC, but are typically clumsy to provide in a traditional mathematical model. We have implemented the architecture and models both in the traditional full simulation using simulated processors and stores, knowledge representations, and production rules, and also in stripped-down MATLAB models which allowed us to focus on specific mathematically-defined processes more easily, and evaluate models using efficient parameter-value searches. Often we have done preliminary exploration in the MATLAB models, and then implemented the full EPIC architecture version; this philosophy thus takes advance of what the two approaches do best, and because they should produce the same predictions, it also provides a cross-check on the implementations.

Model Summary

The application of a cognitive architecture to multichannel speech processing is novel, and so needs to be presented with some detail, but for brevity, low-level representational issues are not presented here. Rather, the emphasis is on the conceptual design of the architecture and model components, especially the auditory processor, taking into account that at this time many processes have to be "black boxed". The following is a compact description of the architecture and model components and processing involved in the two-talker CRM task, flowing from input to response. In some of what follows, the description is somewhat more complex because the mechanism is general enough to apply to more than two talkers.

Speech auditory input. Each utterance is pre-parsed into six segments corresponding to words (with go to being treated as a single word). The segments from the different sources are assumed to arrive at the auditory processor simultaneously and are each perceived as individual auditory events. Each segment pair is processed in order of arrival.

Auditory perception constructs auditory objects based on properties of the physical input. There are two kinds of auditory object: *word objects* represent individual perceived words that have a temporal duration; *stream objects* represent perceived sound sources for these word objects.

Word objects. Word objects have a variety of properties, but for the purposes of this model, they may or may not have *content*, which is the recognized semantic item (e.g. *red*); this allows for a word to be "heard" but not recognized. Words also have *stream attributes*, which in this model are average loudness level (for simplicity, perceived loudness is specified in dB) and average pitch (again for simplicity, perceived pitch is specified in *semitones*, where the number of semitones is defined as $12 \cdot \log2(\text{pitch in Hz})$), both averaged over the duration of the word. Semitones provide a logarithmic scale for pitch, analogous to decibels for loudness. This model assumes that the stream attributes are always perceived.

Whether the content of a word object is recognized in the presence of the other word objects is assumed to be a basic energetic masking phenomenon. The probability of content detection depends on the SNR, that is, the level (the physical intensity in dB) of the word relative to the

other word objects that are simultaneously present, and the pitch difference between the two word objects. With respect to the latter, studies show that discrimination of simultaneous vowel sounds improves with pitch difference, though increasing the difference beyond about 4 semitones produces no further improvement (Assmann & Summerfield, 1990). This effect was incorporated in the model by computing an *Effective SNR* that is the weighted sum of the level difference in dB (the SNR) and the pitch difference in semitones capped at 4.

Stream objects and stream tracking. The stream objects also have attributes of perceived loudness and pitch, but these represent the overall properties of the perceived sound source. In this model, a stream object carries the mean loudness and mean pitch of the words associated with the stream. For example, a typical female talker will be represented as stream percept with a higher mean pitch property than that for a typical male talker.

The auditory perceptual processor assumes that there are as many stream objects as input sources, each with a unique but arbitrary StreamID attribute, and attempts to assign each incoming word object to one of the streams, using the stream-related attributes of loudness and pitch to do so. Once the assignment is done, the stream percepts are updated to reflect the loudness and pitch properties of the words assigned to them, and the next pair of word objects will be assigned to the updated streams. Thus the auditory processor tracks the streams.

Cognitive strategy and response choice. The final output of perceptual processing, represented in the cognitive processor's working memory, is a set of word objects and a set of stream objects. Each word object will always be associated with a stream object, but it may or may not have recognized content.

Because the loudness and pitch of each word in the utterances varies within the same talker, it is possible for individual words from two different talkers to be mis-assigned to the streams, so that each stream is associated with a mixture of words from the two talkers. Figure 4 shows an example in which the Color words have been assigned to the wrong stream, while the Digit words were assigned to the correct stream. This will lead to a response with the Masker Color and the Target Digit.

The cognitive process for selecting a response makes use of the recognized content of the word objects together with the stream associated with each word object. For example, as in Figure 4, if the word object whose content is the target callsign, *Baron*, is associated with Stream2 and there are two word objects associated with the same stream whose content has been recognized as the color *Red* and the digit *8*, then *Red 8* will be used to specify the response to be made.

Some content might be unrecognized, but in many cases the model strategy can infer the missing information. For example, if only one of the Callsign contents was recognized, and it was a Masker Callsign, the model can infer that the unrecognized Callsign word object was the Target Callsign, and its assigned stream must be the Target stream, so the Color and Digit words associated with that same stream must be the Target Color and Digit. Thus the strategic component of the model tries to make use of partial information to perform the task.

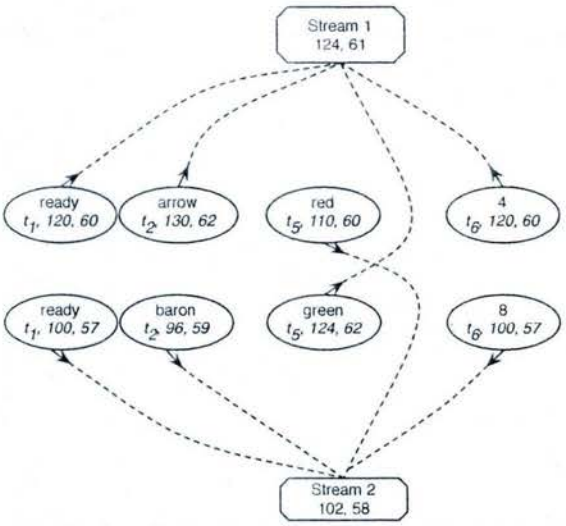

Figure 4. Example showing contents of working memory after erroneous stream tracking. The polygonal boxes top and bottom are the two stream objects, showing mean pitch (Hz) and loudness level (dB) values. The ovals are the word objects in each message in left-to-right time order (goto and now omitted for clarity), showing the content, time stamp, pitch, and loudness. During perception, each word was associated with its closest stream, but because the Color word pitches were discrepant, they were assigned to the wrong stream.

Theoretical summary. In terms of conventional attention theory, this is a "very late selection" model - all of the information produced by perception is available to cognition for choosing the response. The problems of trying to handle two simultaneous messages is not represented as a failure to select the correct stream prior to cognition, but rather that masking effects and errors in stream assignments will result in a collection of information about the perceived messages that may be incomplete or incorrect (e.g. as in Figure 4), and the task strategy must make use of this information to choose a response that meets the task requirements.

Model Details and Parameters

Corpus statistics drive the model. We computed the average level and pitch over each segment in each utterance in the CRM corpus, and supplied this information for each word (segment) that was "heard" by EPIC's auditory processor. An interesting result is that while female talkers had mean pitches about an octave higher than male talkers, individual talkers had somewhat different baseline pitches, which allows the stream tracking to often distinguish talkers within genders over the course of an utterance. Because this model was driven by the corpus properties, there are relatively few free parameters that affect its fit to data.

For each trial, the simulated experiment samples two utterances and then supplies EPIC's auditory system with the content, level, and pitch of each segment. The pitch was converted to semitones.

Content detection parameters. The content detection parameters are summarized in Table 1. The Effective SNR is the sum of the level SNR and the pitch difference in semitones weighted by a parameter w. The pitch difference was capped at 4 semitones, a constant value based on Assmann & Summerfield (1990) and not estimated to fit the data.).

The content detection process is modeled along the lines suggested by Wichman & Hill

(2001). With a low probability (the lapse rate α), subjects will fail to recognize content (even at very high SNRs); otherwise, the probability of content detection follows a gaussian detection function of Effective SNR, with parameters of mean μ and standard deviation σ. The parameters w, α and σ are assumed to be constant across the type of content word (Callsign, Color, Digit), while μ is assumed to have a different value for each type of content word (Callsign, Color, Digit). For completeness, the content detection functions for the filler words ready, goto, and now, were specified, but for simplicity were made the same as the Callsign detection function because the content of the filler words plays no role in stream tracking or response strategy.

Stream tracking details and parameters. The stream tracking parameters are also listed in Table 1. The stream perception model in the EPIC auditory processor uses an *averaging minimum-distance* stream tracking algorithm. Each stream object accumulates the mean pitch (in semitones) and mean loudness (in dB) of the word segments that have already been assigned to that stream. The stream predicts that the pitch and loudness of the next, or new, word segment will be the same as the current means. The stream perception model then calculates the prediction error between each stream and each new word segment as the weighted cartesian distance between the (pitch, loudness) values, where pitch differences are weighted by a parameter λ (0-1) and loudness differences are weighted by $(1 - \lambda)$. As noted above, the pitch difference was capped at 4 semitones. The new word segments are then assigned to streams so as to minimize the total distance between all words and their assigned streams. The streams are then updated to include their newly assigned word segments, and the resulting means used to predict the segment that follows.

The stream perception model included a noise component. After determining the minimum-distance assignment, the stream perception process compares the maximum and minimum total distance; if the difference is less than or equal to a threshold value θ, an assignment is chosen at random.

Cognitive processor strategy exploration. The auditory perception components in the EPIC architecture take the input utterance segments and perform content detection and stream tracking and provide the resulting content and StreamID attributes of the individual word segments, like that shown in Figure 4, to the cognitive processor, which is running a strategy implemented in the production rules.

Table 1 Best fit parameter values

Effective SNR pitch weight *w*	2.00
Callsign content detection μ	-20.00
Color content detection μ	-18.00
Digit content detection μ	-26.00
Content detection σ	10.00
Content detection lapse rate α	0.04
Stream tracking pitch weight λ	0.80
Stream tracking distance threshold θ	0.10

Over the course of constructing the model, a variety of task strategies were considered, and two key options were identified. The first is that in the 2-channel task, symmetrical inferences can be made; for example, if we know that one of the Color words is from the Masker stream, we can infer that the other Color word has to be from the Target stream. The present model strategy incorporates symmetrical inferences.

The second option concerns the "guessing" strategy. Note that in this forced-choice paradigm, the subject must respond even if they have not identified the Target Color or Digit. The optimum strategy would seem to be to always avoid responding with content known to be from the Masker, and choose some Neither Color or Digit instead. However, this Avoid-Masker strategy failed badly to fit the data - it could not account for how there are so many Masker responses in conditions where the Masker stream should be easily identified, such as at extreme negative SNRs. On the other hand, a strategy that always used available Masker content when Target content was missing seriously under-predicted the number of Neither responses. We realized that subjects might adopt a "use what you heard" heuristic: If the Target callsign content was not actually detected, then there is some uncertainty about whether the two streams were correctly identified, so responding using content that was actually detected might be better than a pure guess. Thus the Use-Maskers strategy will use content known to be from the Masker stream if Target content was not detected, but only if the identity of Target stream had been inferred from the detection of Masker callsign content. The model presented here achieved a good fit to the data with this Use-Maskers strategy.

Strategy summary. During the processing of the utterance, if Callsign content is present (detected), tag its StreamID as the Target or Masker stream accordingly. If not, infer the Target or Masker status from the other stream if its Callsign content is present. Then tag the Target or Masker status of each Color and Digit word, based on their assigned StreamIDs. Note that if neither Callsign is detected, it is still possible for Color and Digit words to be paired with their correct streams, but the model will not know which stream is the Target stream or the Masker stream.

When it is time to choose a response, the following rules are used for both choosing the color response and choosing the digit response, depending on what content was detected and which stream it is associated with: If the Target stream is known or inferred, then use the content from the Target stream if it is available. But if the Target stream was only inferred and the Target content is not available, then use the Masker content if it is available. Otherwise, use a color-digit content pair from the same stream if available, or use separate color and digit content if it is available; otherwise, make a pure guess.

Model Fitting and Results

The parameter values shown in Table 1 were determined by Monte-Carlo runs of the EPIC model using a grid search on high-performance computer clusters provided by AFRL through mindmodeling.org. The search goal was to maximize r^2 between predicted and observed values for the Target and Masker Color and Digit probabilities (blue and red curves in Figure 1). Each Monte-Carlo run used 3000 trials per talker/SNR condition. There are a total of 240 empirical

data points with at least 120 degrees of freedom; eight parameter values were varied in the search. The best-fit values are shown in Table 1.

Figure 5 shows the predictions from the EPIC model as open points and dotted lines. All three conditions are well handled with a small set of parameters that describe how the auditory perceptual process is affected by the acoustic properties of the input as provided by the corpus statistics based on the segmentation. It is especially noteworthy that unlike the model presented in Kieras et al. (2014) that was developed during last reporting period, there are no parameters that are specific to talker similarity conditions - the pitch difference used in content detection and stream tracking accounts for these effects.

As summary measures of goodness of fit, $r^2 = 0.99$ between predicted and observed values for the Target and Masker Color and Digit probabilities (blue and red curves), and $r^2 = 0.95$ for the Both-Correct probabilities (black). Only a few of the predicted values lie outside the confidence intervals in the data.

However, there is a clear tendency for the Both-Correct points to be generally under-predicted, probably because our simple model of the stream tracking is not efficient enough. To show this, in Figure 6 are the conditional probabilities of selecting the Target, Masker, or Neither Digit given that the listener has correctly identified the Target Color. On the whole, the likelihood of choosing the Target Digit is far higher than the others, meaning that if the listener has tracked the Target stream correctly as far as the Color, then he or she is very likely to get the

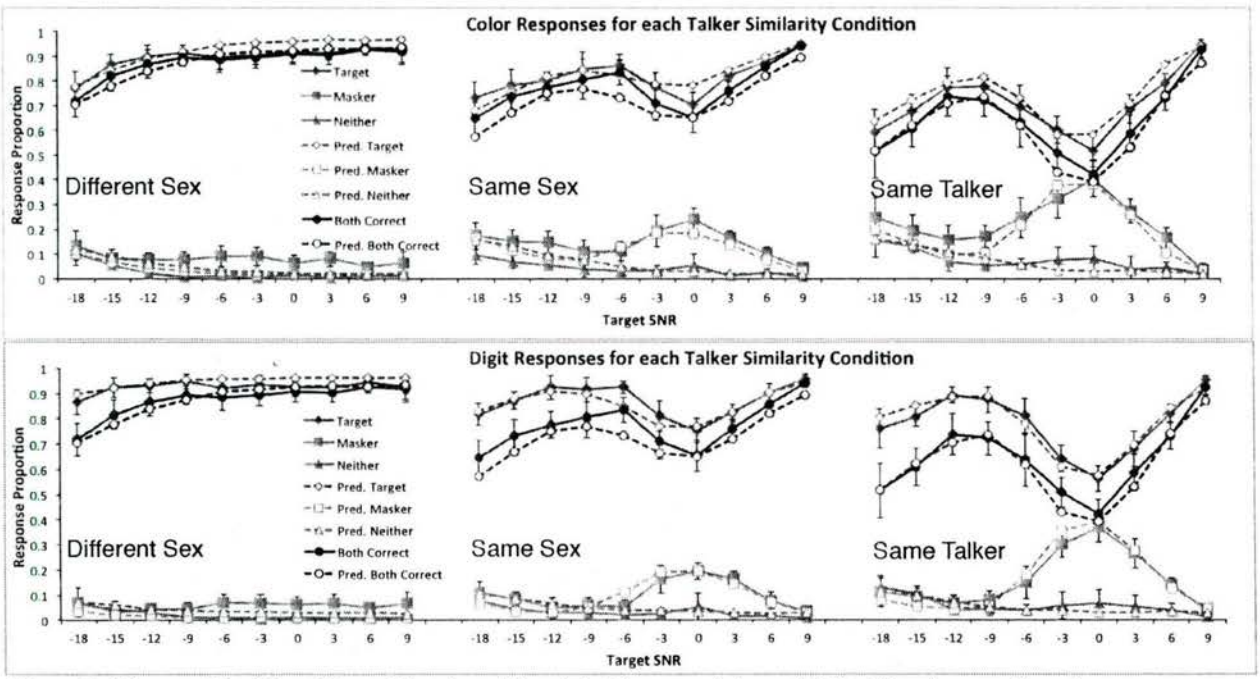

Figure 5. Observed (solid points and lines) and Predicted (open points and dotted lines) proportion of responses as a function of SNR and talker similarity. Top panel shows Color responses, bottom panel shows Digit responses. In order from the top down, the curves are as follows: Blue curves with diamond points are for Target responses, black curves with circle points are for Both-Correct responses (both color and digit from the Target), and are the same in the top and bottom panels; red curves with square points are for Masker responses, and green curves with triangles for neither Target nor Masker. Error bars show 95% confidence intervals for the means averaged over individual subject proportions.

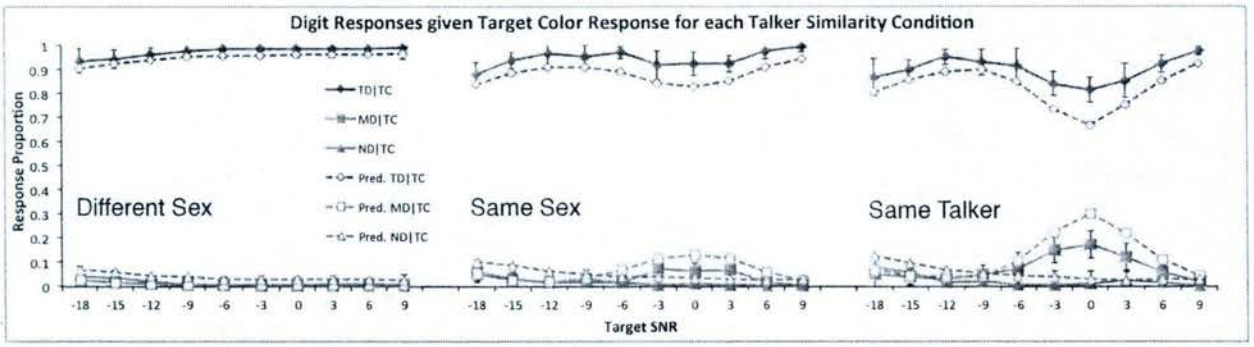

Figure 6. Observed (solid points and lines) and Predicted (open points and dotted lines) conditional probabilities of selecting a Target, Masker, or Neither Digit response given that the Color response was the correct target Color, as a function of SNR and talker similarity. In order from the top down, the curves are as follows: Blue curves with diamond points are for selecting the Target Digit (TD | TC), red curves with square points are for selecting the Masker Digit (MD | TC), and green curves with triangles are for selecting a Digit that is neither the Target nor a Masker Digit (ND | TC).

Digit correct as well because the stream tracking is very likely to be correct. In contrast, choosing a Masker Digit would be a case of the stream tracking "switching" to the Masker stream. Notice that the model is consistently less likely to choose the Target Digit than the subjects, and is more likely to switch to the Masker or Neither Digit, especially when the tracking is more difficult in the vicinity of 0 dB SNR for the Same-Talker condition. The result is a tendency to under-predict the Both-Correct responses, even though the individual Target and Masker responses are well predicted. Despite the overall good fit, this is a concern because the Both-Correct measure is by far the most common performance measure in the CRM literature.

The above-summarized model and its predecessors have been published in multiple conference papers (Kieras, Wakefield, Thompson, Iyer, & Simpson, 2014, 2015; Wakefield, Kieras, Thompson, Iyer, & Simpson, 2014), a published journal paper (Kieras, Wakefield, Thompson, Iyer, & Simpson, 2016), , and one in preparation (Wakefield, Kieras, Thompson, Iyer, and Simpson, in preparation). Additional details of the predecessor models are available in Kieras & Wakefield (2014). The segmented CRM corpus database is available in Wakefield (2014).

Exploration of improved stream tracking

Characterization of stream tracking with an abstract model. Because the Both-Correct performance measure is the most common measure in the CRM experimental literature, it is important to correct the model's systematic failure to account for it. While modifying some of the model parameters can improve the fit of Both-Correct, the result is that the Color- and Digit-Correct performance is over-predicted, so the problem is a true structural problem in the model. As shown in Figure 6, the problem is that if the model has determined the correct stream assignment at the time of the color word, it does not assign the digit words to the correct stream often enough. This should not happen if the color and digit words in each stream have similar loudnesses and pitch. But there is only a marginal degree of correlation in pitch and loudness values across consecutive words in the CRM corpus. Therefore, accounting for the joint behavior of listener responses requires a different formulation of the stream tracker.

A first line of work before implementing any specific new stream tracking approaches was to determine how an improved tracker should behave functionally. To do this, we returned to the approach we used to kick-start development of the current stream tracker, namely an *abstract model* (see the 2013-2015 Annual Reports and Kieras & Wakefield, 2014). Our abstract model, built and tested in MATLAB, provides a highly simplified functional description of the input-output relations in EPIC that allows exact probabilities to be computed and parameters easily estimated. Specifically, the input is represented by six states (the target/masker status of call sign, color, and digit contents of the target and masker utterances), the output is represented by two states (color and digit response), the input states are transformed by content-detection functions and stream-assignment functions, and output states are determined by applying production rules to these transformed input states. In this manner, probabilities can be calculated for a given set of production rules, which relate the input and output states, and a constrained optimization procedure can be used to determine the best-fitting parameters. The key simplification is that the model behavior can be summarized as "stream switching" - if the streams have been correctly assigned thus far, then what is the probability that the streams will "switch" to the incorrect assignment on the current word? So rather than implementing a stream tracker, the abstract model simply describes its behavior in terms of the probability of switching the stream assignment.

As is clear in Figure 6, the case in with the lack of "stickiness" is most pronounced is in the Same-Talker condition. Accordingly, we fit the abstract model to the Same-Talker condition data from Thompson et al. (submitted). The model used the same production rules as the current EPIC model. Content detection functions were fit by the same lapsed Gaussian form but with a fixed lapse parameter of 0.01. In our original abstract model, the probability of switching streams was specified with a function that in the Same-Talker case, had its highest value at 0 SNR, where the two streams are most confusable, and drops off at both higher and lower SNRs. The same switch probability function was applied independently to the color stream assignment and the digit stream assignment. In the new model, the switch probability for color is described similarly to the original abstract model, with a stream switch function that gives the probability that the color word was assigned to the same stream as the callsign or was switched to the other stream. It was fit under the constraint that it had to be symmetrical and monotonic above and below 0 dB - it reflects the *baseline* situation that the streams are hardest to distinguish when the levels are the same, meaning that the probability of switching to the other stream is highest at the point, and drops off on each side.

The key feature of the new abstract model is that the stream switching process incorporated a strong conditional relationship between the color and digit stream assignments. Determining whether the digit would be assigned to the same stream as the color or switched to the other stream is a two step process: First the model applies a "sticky assignment" function to see if the digit should be assigned to the same stream as the color. If not, it is then switched according to the same baseline switch function as color assignment. The "sticky" behavior would thus show up in the form of the "sticky assignment" function; therefore, it was allowed to freely vary to fit the data. The *fmincon* function in Matlab was used to optimize the model parameters. Keep in mind that the abstract model does not actually implement a stream tracker, but only describes its

behavior in terms of a probability of switching the stream assignment. By examining the probability functions that allow the model to fit the data, we can determine how the tracker should behave.

Figure 7 shows the best fit of the abstract model to the Same-Talker data and the best-fitting parameters. The left-hand panels show observed and predicted proportion correct using the same conventions as Figure 5 above, with color responses at the top, digit responses at the bottom. Target responses are the blue curves, masker responses are red, and Neither responses are green. The probability of Both-Correct are shown as the black curves in both the top and bottom plots. The key result is that the model fits not only the Target and Masker responses, but also the Both-Correct responses extremely well. The underlying model functions are shown in the right panels. The content-detection functions at the top are similar, though not identical, to those for the current EPIC model. In the bottom panel is shown the best-fit baseline switching function plotted with circles, with its symmetrical peak at 0 dB SNR. The best-fit "sticky assignment" function is plotted with squares. The fit shows that with a fairly high and constant probability, whatever

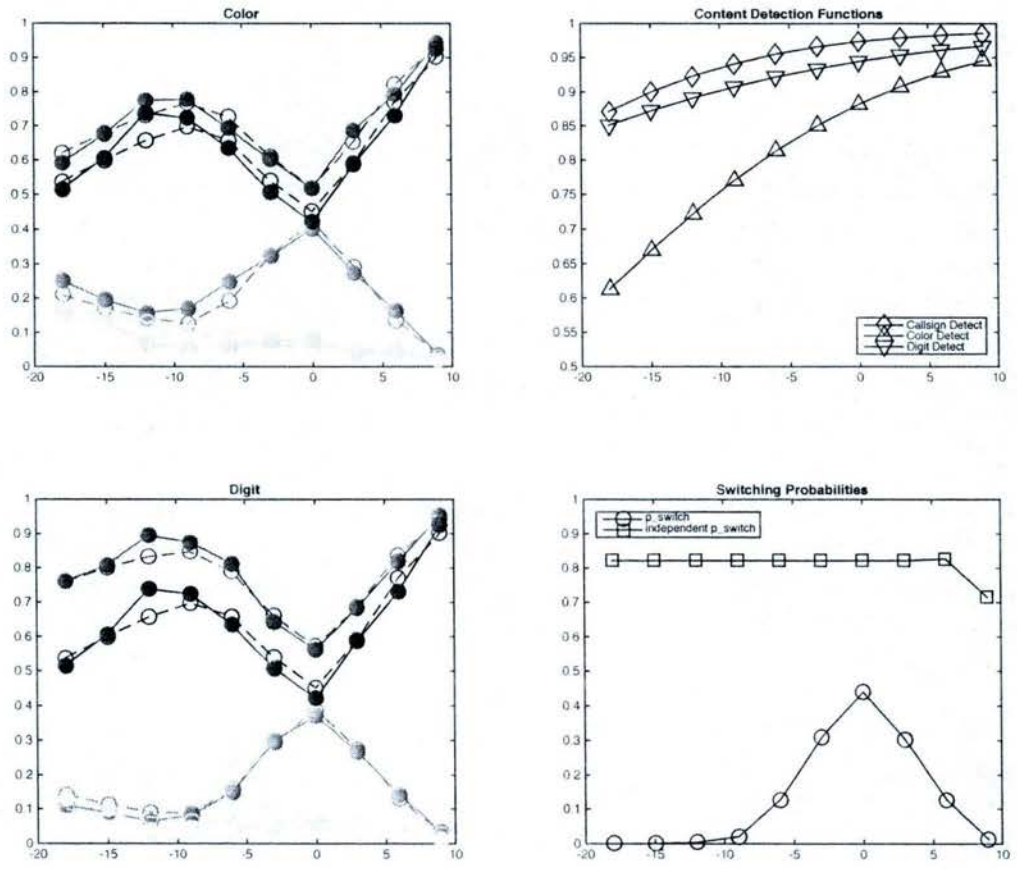

Figure 7. Results for a successful type of joint stream-assignment function are shown in the four panels for Same-Talker condition in the Thompson et al. (submitted). Human performance is shown by the solid symbols for color (top left) and digit (bottom left) responses, and model performance is shown by the open symbols. Symbol color denotes type of response: blue (target), red (masker), green (neither), and black (joint target correct). The curves shown in the panels on the right characterize the abstract model. The content-detection functions (upper right panel) are generally similar to the current EPIC model, but the stream-assignment functions (as specified in terms of stream-tracker behavior) differ considerably. See the text for further discussion.

stream assignment was made for color should be carried over to the digit. Thus the abstract model fits the data better than our current model by representing the notion that color-digit pairs usually stick together very strongly, and at a level that does not depend much if any on the SNR. A successful stream tracker will show these characteristics.

Now that we know how the tracker should behave, the question arises as to how to achieve the remarkably high level of color-to-digit "stickiness" required to account for the data. As noted above, the word-to-word correlations of pitch and level in the corpus are simply not strong enough to make the current tracker "sticky" enough. We have explored several hypotheses. The main alternatives we have considered thus far all involve a move towards finer temporal grain in the model. That is, rather than representing the two messages as series of six segments arriving in synchronized sequence, we consider a much finer segmentation of the utterances and/or the segments arriving at different times rather than simultaneously.

Using onset discrepancy information. A promising possibility is based on an observation that speech-on-speech masking can be reduced or eliminated if there is an asynchrony of about 100 ms or more in the onsets of the words in the two streams (e.g. Lee & Humes, 2012). Even though the CRM corpus was designed to make the two messages as synchronized as possible, our segmentation of the corpus (described in the 2013-2015 Annual Reports and Kieras & Wakefield, 2014) revealed that while almost all pairs of callsigns are effectively synchronous (onset times within 100 ms), around 13% of the color pairs, and 27% of the digit pairs are not - they have onset times more than 100 ms apart.

This suggests a stream tracker in which the termination of a word in a stream signals the beginning of the next word in that same stream. For example, if stream A has a color word just terminated while stream B has a color word still in progress, and a digit word is just beginning, the stream tracker could immediately assign that beginning digit word to stream A. If the two color words terminated at the same times (or within 100 ms), then the stream tracker would assign the digit words using the same pitch-loudness tracking logic as the current tracker. Such a tracker should be more sticky because the color-digit pairs are significantly more coupled along the lines suggested by the abstract model.

Accordingly, the stream tracker in the MATLAB implementation of the current EPIC model was modified to incorporate onset discrepancy. Specifically, if the delay in word onset exceeds a particular threshold, the new word is assigned to the same stream as the previous word. But if both words change within the threshold, the original stream tracker is used. Simulations of the EPIC model for various onset discrepancy thresholds were run and the parameters of the content-detection functions were optimized by hand to get a quick preliminary look at this approach. Results from one such simulation in the Same Talker condition are shown in the four panels for Figure 8. The model does a promising job of fitting the data. In fact, comparing the predicted and observed Both Correct results, it appears to be a little *too* "sticky," but it does exhibit switching behavior that is very similar to that of the abstract model in Figure 7. Therefore, using onset disparities as a source of stream assignment information is one way to get the abstract model behavior.

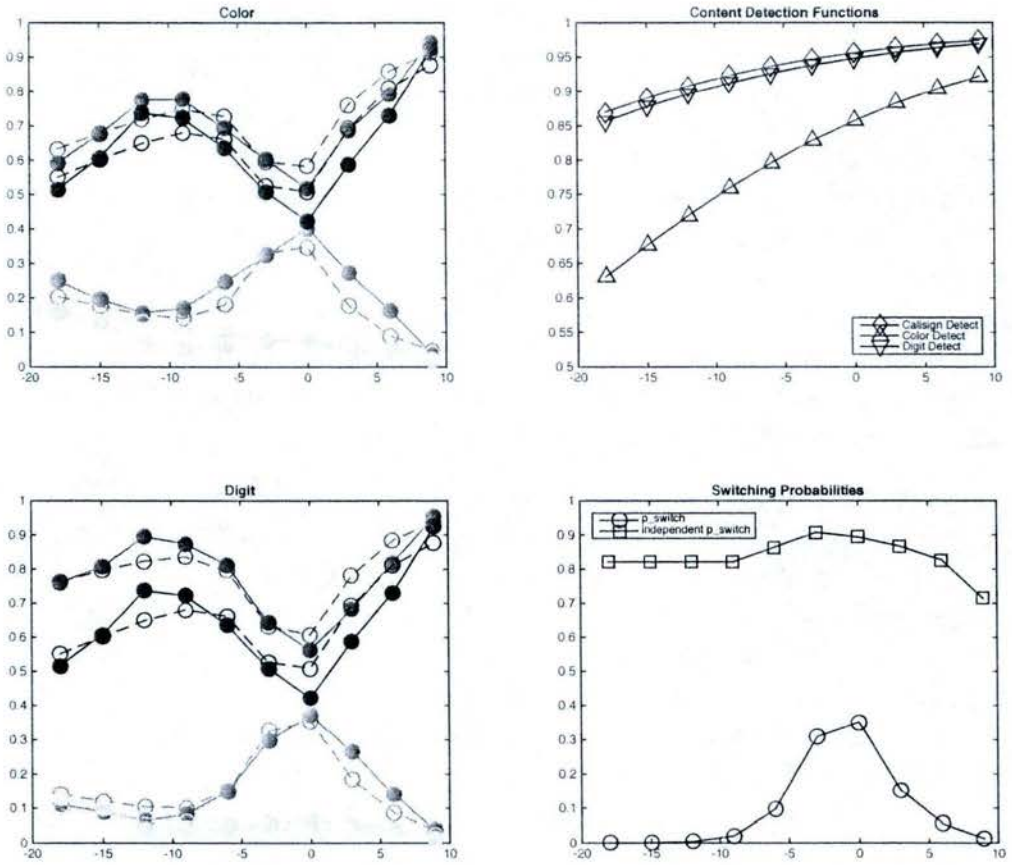

Figure 8. Performance of a modified stream tracker that utilizes onset disparity in stream assignment.

Further parameter fitting would be needed to fully test this model. Unfortunately, implementing the model in the full EPIC architecture would require some significant modifications to correct the current simplifying assumption that the utterances are synchronized, and then we would have to work out how to correctly represent masking effects in non-synchronized input. These are inevitable steps in the refining the EPIC architecture to be more realistic. But before making this move, we want to use the MATLAB implementation to more fully evaluate the onset discrepancy model. In addition, we want consider other stream tracking alternatives that depend on how we represent the temporal properties of speech input before modifying the full EPIC architecture.

Fine-grain tracking as clustering. As noted in our earlier reports, while it is gratifying that such gross measures as pitch and level at the scale of words does such a good job of accounting for the data, it is likely the auditory system is tracking pitch and level variations at much finer time-scales. With the discovery that onset disparities provide a means to counter one of the primary structural deficiencies in the current EPIC model, we decided to examine tracking

algorithms for pitch and level at finer temporal scales. This considerable effort is still ongoing, but key results are summarized below.

The current EPIC stream tracker is based on the concept that segments of the speech input have properties such as pitch and level, and stream percepts carry the mean pitch and loudness of the input segments that have been assigned to them. Each segment is assigned to a stream in a way that minimizes the total distance between segments and streams. In order to compute these distances, the loudness and pitch dimensions have to be combined in some way even though they have different scales and ranges. This was done with a rather simple scaling parameter whose value was determined by fitting the data. In the current tracker, only the means for the two dimensions are retained in the stream percepts. Another way to improve the tracker would be to combine the pitch and loudness values during stream assignment in a more statistically efficient way. The problem is that if the utterances are simply represented as a series of six word-length segments with pitch and loudness, there are no powerful statistical methods that will work with such small samples. But if we move from a word-level representation to a finer time scale, the amount of data available to drive the tracker increases substantially (from 6 words to 1600 1-ms samples, for example), so that incorporating a statistical model directly into the tracking algorithm becomes a sensible alternative. We have experimented with processing the input in terms of segments of from 1 to 100 ms duration.

During this period, we explored one such statistical method drawn from the clustering literature. Rather than trying to compute a Euclidian distance with assumed scaling parameter between each segment and each stream, we tried using the *Mahalanobis distance* instead. This measure optimally weights and rotates the data according to the mean and covariance structure of the underlying distribution, which is assumed to be multidimensional Gaussian. If the different dimensions have the same variance and are uncorrelated, then the Mahalanobis distance reduces to simply the Euclidean distance. However, in clustering data using a nearest-neighbor method, the Mahalanobis distance is often found to be superior to a simple Euclidean method. In our application, the stream tracker would work similarly to the current tracker, but rather than simply accumulating the mean pitch and loudness values, the stream accumulates *all of the values* under the assumption that they are being sampled from a two-dimensional Gaussian distribution of pitch and loudness, and so the accumulated data produces an increasingly accurate approximation to the actual distribution represented by the stream. When a new segment arrives, the tracker computes the Mahalanobis distance between the (pitch, loudness) value for the segment and the distribution maintained by each tracker, determines which assignments yield the smallest distances, and updates the tracker data accordingly.

This modified tracker has exhibited several interesting properties, which we are currently assessing. The Mahalanobis distance eliminates the need to weight the dimensions with a parameter (as expected), but also appears to capture some of the more interesting features of pitch variation that have, hitherto, been unexplained. Specifically, assignment errors do not require an initial truncation step as in the current EPIC model where pitch differences greater than 4 semitone are eliminated. Like our current tracker working at the word-segment level, the Mahalanobis tracker working with fine-grain segments appears to capture the differences

between female and male speakers.

However, getting good tracking behavior requires smoothing the input over larger durations. With 1-30 ms segments of the input, the sample-to-sample variations in pitch and level are so large that discriminating between a target and masker becomes very poor. Smoothing the input over 60 msec or more produces tracking much closer to the data. This agrees with a large number of studies concerning critical durations for pitch (and possibly loudness), where 60 msec is often reported as a minimal duration to achieve a stable pitch.

Glimpsing as an explanation for multiple-masker effects. The concept of glimpsing, first proposed by Miller and Licklider (1950) has become more prominent over that past twenty years to explain the perception of search recognition in the presence of noise, but has an potentially greater value for understanding the effects of speech maskers, especially multiple speech maskers.

The concept is that the listener does not have to get all of the speech waveform in order to understand the message. There is considerable redundancy in speech, if the listener is familiar with the language, and especially if the context is familiar. This means that it would suffice just to get segments of the speech signal, *glimpses* of it. That is, the interfering noise will not always obscure the signal; there will be brief intervals of time in which the noise will be momentarily low, and the speech signal will be relatively loud and clear. The auditory and the cognitive system will be able to infer the obscured parts of the message from the glimpses that were heard. The greater the SNR, the more often the message can be glimpsed in the noise, and the inference process will be more accurate.

More directly applicable to our problem, if instead of noise, the interfering signal is another speech message, then glimpses will generally be much more available. That is, it is characteristic of speech that its spectrum and intensity varies widely - compared to relatively constant noise, there are many opportunities for the other speech message to get revealed for brief times. So in general, detecting the content of a speech message at a given SNR is much better if the masker is another speech signal than if it is broadband noise. In fact, if there are two or three masking speech messages, the performance is so poor that it resembles the effect of a noise masker. The explanation is that the multiple speech maskers overlap so much that there are fewer opportunities for glimpses of the target message, just like for noise.

Our current two-talker model can account for the effects of multiple maskers if the content detection functions are changed to be less sensitive and much steeper with additional maskers (see our previous Annual Reports and Kieras & Wakefield, 2014). However, we currently lack a principled basis for this effect. A glimpsing model might provide this explanation, as well as a more detailed explanation for speech understanding in the presence of masking signals. At present we have only been surveying the relevant literature. What follows is a brief survey and some key conclusions.

Cooke (2006) demonstrated that the information present in a collection of glimpses was sufficient to recognize speech when presented in a noise masker. He constructed an automatic speech recognition system that was trained on a set of speech tokens to recognize consonants in

noise. Parametric study of glimpse support (e.g., the region in time and frequency over which a glimpse is defined) and threshold SNR (e.g., the value at which a glimpse is declared significant) revealed several sets of parameters for which the psychometric identification functions generated by the system matched those of human listeners.

Using the CRM procedure, Brungart, Chang, Simpson, and Wang (2006) found similar support for the notion that speech recognition is driven by glimpses. Listeners were tested using either the standard CRM stimuli or *ideal time-frequency segregated* (ITFS) forms of the stimuli in which only significant target glimpses (T-F units in the words of the authors) are retained. In general, when the masker was wideband noise, eliminating glimpses dominated by the noise improved performance by 2-5 dB. However, eliminating glimpses dominated by competing speech utterances (from one to three talkers) improved performance substantially beyond what would be predicted simply on the basis of SNR.

In his discussion, Cooke distinguishes between the process of detecting a glimpse and that of integrating across glimpses to reach a decision. Pre-processing a noisy speech signal into glimpses with highly reliable information (that is, improving detection) yields better performance simply because the process that integrates such information across glimpses is drawing on better data. In principle, such thresholding should be subject to the same criticisms leveled against high-threshold models by estimation and detection theorists over sixty years ago: that there is no sensory threshold is a more or less accepted fact when modeling, for example, the detection of sinusoids in noise. It should be emphasized that Cooke's approach is more akin to applying a priori binary weights to approximate the scoring that would normally occur when combining data from a variety of sensors with varying degrees of sensitivity, signal intensity, and external noise. Presumably, any loss of partial (but noisy) information at the point of statistical integration is offset by the highly redundant nature of the speech signal itself so that even a binary weighting of the data yields superior performance.

Cooke's treatment of glimpsing and detection aligns well with the notion of energetic masking effects in the Brungart et al study. The ITFS processing of the stimulus into T-F units (glimpses) that are dominated by the speech signal improves detection, in this case, by a relatively small amount (15%, on average, assuming that the slope of the psychometric function is 5%/dB). According to Brungart et al., the additional improvement in performance for speech maskers reflects a release from informational as well as energetic masking. By pre-processing the two-talker observation into glimpses using ITFS, the integration process is improved because highly-reliable glimpses for the target speech utterance are retained and highly-reliable glimpses for the masker speech utterances are eliminated. Thus, informational masking reflects a breakdown in the integration process: spurious glimpses are being (inappropriately) included at the front-end to the integration. That is, the detection process has not properly discriminated among glimpses.

Our current work has considered whether the concept of a glimpse and associated psychometric functions based on aggregating glimpses over time can account for the change in detection function slope observed as the number of speech maskers increases and then goes to a wideband noise maskers. At this preliminary stage, the following observations are appropriate:

- All attempts to model glimpses appear to require substantial prior knowledge about the signal and the masker. Even in the case of ITFS, an unrealistic degree of certainty is inserted up front in the model: the two utterances are known, a priori, so it is clear how to interpret high-energy regions appropriately as belong to the target or to the masker. This strategy generally works well for the wideband noise case, since the masker provides a baseline against which any variations substantially above baseline can be deemed as reflecting target energy. When both signals are speech, an additional uncertainty (which glimpse goes with which hypothesis) must be incorporated if the system is processing signals not known in advance.
- Attempts to model changes in slopes of psychometric functions may contain some conceptual errors. We have identified several published accounts which run counter with the older literature concerning temporal integration and multiple looks (see Viemeister and Wakefield (1991) for a more detailed discussion).
- Current attempts to model glimpses equate a masking speech utterance with noise and draw upon the slopes of *detection* functions. Curiously, a smaller but still well-founded literature exists from many decades ago on the slopes of *discrimination* functions. These are always much shallower.
- From our literature review and current mathematical investigations, we propose that much of what is called informational masking in the CRM task reflects the statistics of glimpses, assignment errors, and the statistical integration of *discrimination* judgments, rather than detection judgments.

Goal 1.2: Incorporating spatial location in multichannel speech tasks

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 sound presented over headphones and modified with 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 (submitted) 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.

Spatial location effects with two talkers. Most of the available studies have manipulated spatial location only in the azimuth, the horizontal plane at the level of the ears. By convention, "straight ahead" is an azimuth of 0°, 90° is to the right, aligned with the right ear, and -90 ° (or 270°) is to the left. Almost all studies using speech signals have focussed on the front half-circle, between -90° and +90°.

In our work on incorporating spatial location into our speech processing model, we have focused on experiments that used the CRM stimuli and task, because these would support a direct generalization for our diotic model. Unfortunately, most studies have not used a wide range of azimuth locations, resulting in data sets with few points. For example, Kidd et al. (2005), focused on whether knowing the target location in advance resulted in improved performance, used only three locations, -45°, 0°, and +45°. Datasets with such a small range of values for important parameters are ill-suited to developing and testing complex quantitative models - more data points constrain the model substantially, and complex models needs lots of constraint!

The most promising dataset we have found thus far is one presented, but incompletely reported, by Brungart and Simpson (2004). This is a subset of a more complex study on demonstrating an optimal layout for a seven-source multichannel task (see Brungart & Simpson, 2005). In the subset of these results that we are using, only two talkers were used, each at one of two locations, both locations at 1 m distance, and with a level normalization which produced the

same levels at the ear closest to the sources, which were always in the same quadrant (i.e. azimuths between 0° and 90°). One source, Talker A, was located at one of {0, 5, 15, 30, 45, 60, 75, 90} degrees, and the second source, Talker B was located at one of {5, 15, 30, 45, 60, 90} degrees. The Target and Masker messages were sampled from the same talkers in the CRM corpus, and the assignment of each to Talker A or B was done at random on each trial. In terms of the Brungart (2001) and Thompson et al. (submitted) study previously described, the stimuli were always drawn from the 0-dB Same-Talker condition, which is the lowest-performing condition.

Figure 9 shows the basic effects of spatial position. The y-axis is the percent of Correct responses in which the color and digit are both from the Target message. The x-axis is the azimuth of Talker A. Each curve is the performance at a specific Talker B location. Overall, the lowest performance, about 40% correct, appears when the location of the two talkers is the same - for example, if both are at 45°. Performance improves if the two talkers are at different locations, but the improvement is asymmetric. If Talker B is at 5°, the performance for Talker A at 0° or 15° azimuth is high and gets even higher at larger azimuths. But if Talker B is at 60° or 90°, then performance does not get high until Talker A is much more in front, at 30° or less.

Brungart & Simpson explain this effect by referring to the classic work on the *minimum*

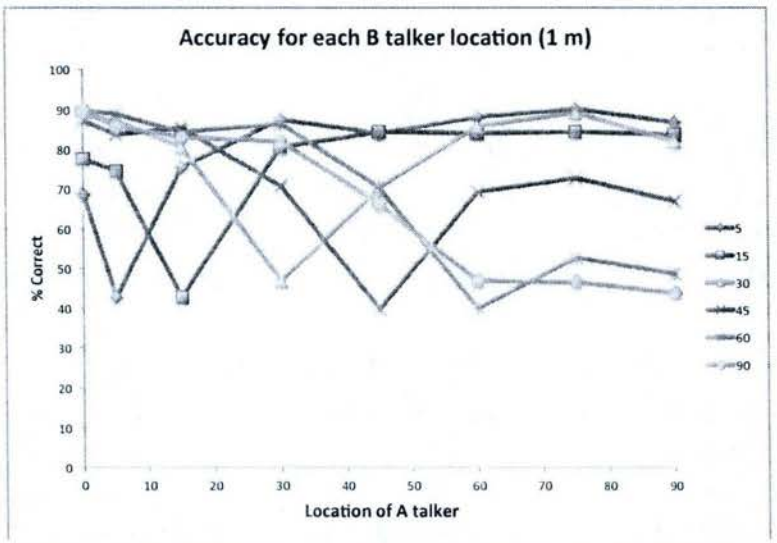

Figure 9. Performance as a function of the azimuth location for each Talker. The location of Talker A is on the x-axis. Each curve is for a particular Talker B location. For example, in the lowest leftmost corner, if Talker B is at 5°, performance is about 70% if Talker A is at 0°, almost 80% if Talker A is at 15°, and about 90% at all larger azimuths.

audible angle, which is smallest detectable difference in the location of a sound (Mills, 1958). Mills presented blind-folded listeners with a brief reference tone at a certain azimuth, and then moved the sound source and presented the tone a second time. The listener judged whether the source had moved to the left or right from the reference position. Repeated trials produced a

psychophysical function which for a reference azimuth, shows the accuracy of this judgement as a function of difference in azimuth; half the distance between the 25- and 75-percentile points on the function was defined as the minimum audible angle (MAA) at that reference azimuth. The MMA has a complex dependence on both azimuth and frequency of the tone, as shown in Figure 10 below.

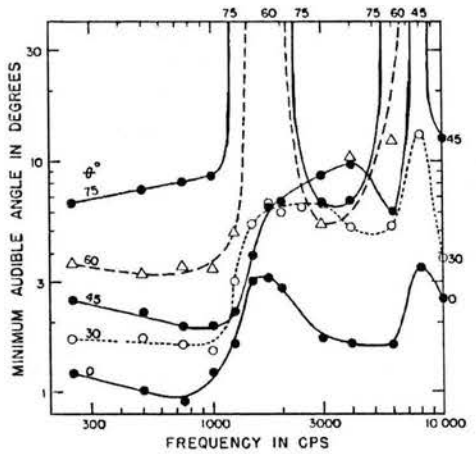

Figure 10. From Mills (1958), Figure 5. The average MAA as a function of stimulus frequency; each curve corresponds to the azimuth of the reference tone.

The key result from Mills (1958) can be seen in this figure. Notice that for low frequencies at the large 75 degree reference azimuth (the top curve on the left), the MAA starts relatively large, but then at higher frequencies becomes so large that it is "off the the chart," with the perceived location becoming essentially indeterminate. In contrast, at small reference azimuths, the lower curves, the MAA remains relatively small, at 3° or less, though it does increase with generally higher frequencies. Figure 11 shows a subset of the data; the MAAs for 500 and 1000 Hz stay on

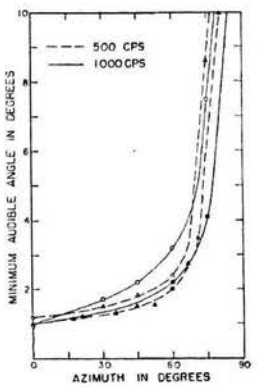

Figure 11. From Mills (1958), Figure 6. The average MAA as a function of azimuth for two different frequencies whose MAAs are in bounds in Figure 10.

the chart, but increase very rapidly as the azimuth increases past 45°.

Thus location discrimination is best at locations in front of the listener, and is much worse at for locations near either ear. Brungart and Simpson (2004, 2005) use this principle to propose that an optimal arrangement of source locations would have close spacings for the front-most sources, with wider spacing towards the sides. The detailed effects in Figures 9 are consistent with this. If one of the talkers is at the high-resolution 5-degree azimuth, then the other talker can be pretty much *anywhere else*, and performance will be high. But if one of the talkers is at 90°, the other will have to be far away to produce high performance.

Perrot (1984) pointed out that Mills's procedure involved successive rather than simultaneous judgements, which would relevant if we are concerned with discriminating the location of simultaneous speech messages. He presented two concurrent tones differing slightly in frequency, and ask the listener to judge whether the higher-pitched one was to the right or left of the other. Rather than plotting the results in terms of the azimuth of the reference tone, he plots the MAA in terms of the azimuth midway between the two sources. As shown in Figure 12, he obtained effects qualitatively similar to Mills's results, but the absolute values of the MAA were much higher, ranging from 5° at the in-front azimuth to as much as 45° at the side azimuths, with MAA increasing more rapidly when the frequency difference was smaller.

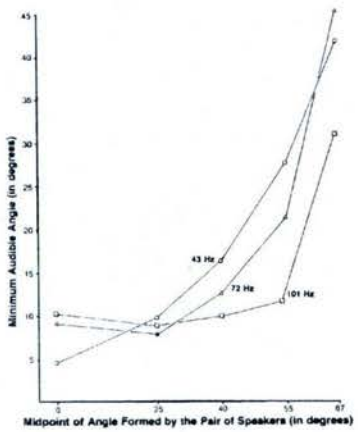

Figure 12. From Perrott(1984), Figure 4. MAA as function of azimuth for each stimulus frequency difference.

Hartman and Rakerd (1989) recast the MAA methodology into decision-theoretic terms and showed that different experimental procedures determine how to estimate the variability of the perceived locations from the measured MAA. The Brungart & Simpson (2004) results make it clear that the MAA results are important for understanding the effects of presenting speech messages at different locations. The application to our models is that we can assume that the perceived location of a sound source is sampled from a distribution whose mean is the true location, and whose variability can be estimated from the value of the MAA for that true location.

Thus location discrimination is best at locations in front of the listener, and is much worse at for locations near either ear. Brungart and Simpson (2004, 2005) use this principle to propose that an optimal arrangement of source locations would have close spacings for the front-most sources, with wider spacing towards the sides. The detailed effects in Figure 9 are consistent with this. If one of the talkers is at the high-resolution 5-degree azimuth, then the other talker can be pretty much anywhere else, and performance will be high. But if one of the talkers is at 90°, the other will have to be far away to produce high performance.

The Brungart and Simpson (2004) results make it clear that the MAA results are important for understanding the effects of presenting speech messages at different locations. The application to our models is that we can assume that the perceived location of a sound source is sampled from a distribution whose mean is the true location, and whose variability can be estimated from the value of the MAA for that true location. Unfortunately, MAA has not been measured for speech stimuli generally, nor for CRM stimuli in particular. However, the general form of the MAA results in the literature can be approximated with an exponential function whose parameters can be estimated to fit the data.

A preliminary model for spatial location effects. Incorporating different spatial locations into our multichannel speech model was always a goal. But by starting with the simpler diotic presentation paradigm, we were able to develop our basic model of stream tracking. Our approach to adding perceived spatial location information is to treat it as an additional attribute of the speech segments and stream objects, in direct analogy to how loudness and pitch are used.

As a first pass at extending the architecture and model, we simply added spatial location (in the azimuth) as an additional attribute of words (segments) along with their mean pitch and loudness. The effective SNR for content detection was expanded to include a weighted log absolute location difference as a contributor to the effective SNR, analogous to how pitch differences contribute to content detection. The stream tracking mechanism was expanded to include the location as a stream attribute. During the phase in which words are assigned to streams, the distance in a 3-space of weighted loudness, pitch, and location, is computed, and the words are assigned in a way the minimizes the total distance in this 3-space.

Just on this basis, with any reasonable weighting, the location information greatly improves performance. However, matching the effects in the Brungart & Simpson (2004) data requires that the location difference be noisy, corresponding to the variability in the pitch and loudness of the words. As mentioned above, we represented the variability with a simple exponential function. Thus the perceived azimuth of a segment is sampled from a gaussian distribution whose mean is the actual azimuth, and whose standard deviation is determined from the exponential function. This model produces a perceived location whose variability sharply increases with azimuth. Thus if the two sources are both at large azimuths, their perceived locations will often be very different from the actual, leading to stream assignment errors analogous to those for loudness or pitch. If at least one of the sources is at a small azimuth where the variability in perceived location will be much less, there is a better chance of the stream assignment being correct.

In addition, ILD effects were taken into account. In the Brungart and Simpson (2004) experiment subset, the levels of the two sources were adjusted to be equal at the ear on the same side as the sources, corresponding to an SNR of 0 dB. However, due to head-shadowing effects, the levels at the other ear would be different, leading to a non-zero SNR. A variety of studies suggest that performance in a binaural masking task will be determined primarily by the SNR at the ear that has the better SNR - the better-ear effect. To approximate this head-shadowing better-ear effect, we used the results from Shaw (1974) who combined many studies of free-field acoustics to produce functions showing the dependence of intensity at the ipsilateral (closer to source) and contralateral ears on the azimuthal location of a sound source. These functions are strikingly non-linear, and depend on frequency. For the preliminary model, we used Shaw's reported function for 2500 Hz. The model computes the level of each source at each ear, and then applies the "better ear" rule to choose which ear has the largest SNR. The resulting loudness SNR combined with adjustments from the pitch differences, yielded the effective SNR used to determine content detection.

In summary, the model starts with our previous non-spatial model, adds spatial location as a stream attribute to be used in stream tracking, and as a contributor to content detection both in terms of improving discriminability, and in modifying the loudness SNR. The reliability of perceived spatial location is based on the MAA effects, so that the reliability of location differences depends not only on the differences in azimuths, but also on the absolute azimuths of the two sources. The task strategy represented in the production rules is unchanged from the previous non-spatial model, as are the content detection and stream tracking parameters; the pitch and loudness weights are unchanged. If there are no azimuth differences in the stimuli, the model predicts the same performance as in the model for the diotic stimuli.

Figure 13 shows the accuracy data as in Figure 9 along with the predicted accuracy, shown as open points and dotted lines. Keep in mind that this is a preliminary fit; only a few iterations

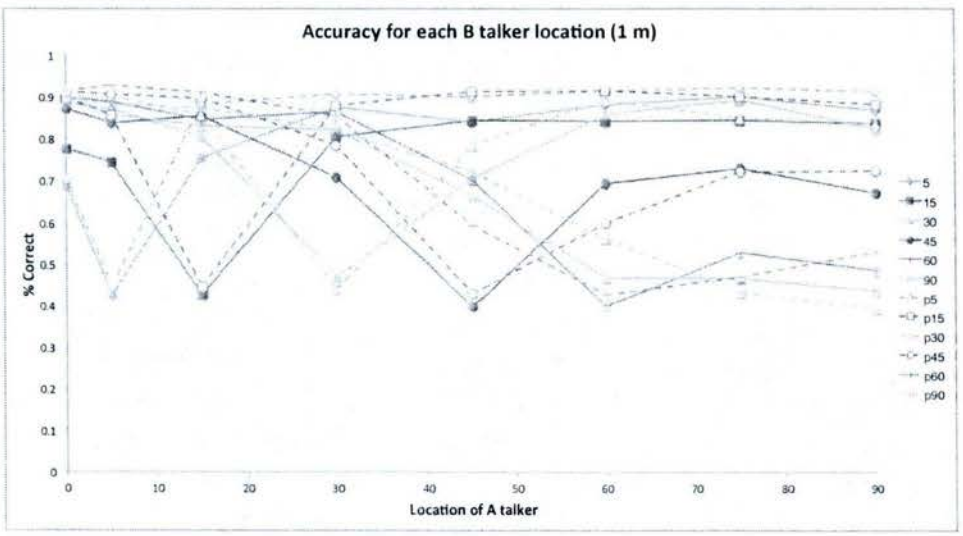

Figure 13. Predicted (open points, dotted lines) and observed (solid points and lines) performance as a function of the azimuth location for each Talker. The location of Talker A is on the x-axis. Each curve is for a particular Talker B location. The r^2 is 0.93.

were done on the relevant parameters. In spite of this, the r^2 is 0.93, an excellent result for a preliminary model. Note that the basic effects are there - performance is best if one source is at a small azimuth, or there is a large separation between the two sources. A conference paper has been submitted on these results (Kieras, Wakefield, Brungart, & Simpson, submitted).

The complex and asymmetric results are to be expected given the complexity of the effects of sound localization phenomena in general. Thus is it very encouraging that this extremely preliminary model is already producing many of the effects, and suggests that we are on a good path for incorporating spatial location in our model of multichannel speech processing. The details on how we incorporate spatial location may change if we move to a different stream tracking concept, such as those described in the earlier section.

Goal 1.3: Models for Auditory Localization Accuracy

Our original plan was to model the accuracy of perceived spatial location as preparation for including spatial location in our multichannel speech models. The Annual Reports summarize a body of work we performed to contribute to the use of localized sound in display systems, as in the NSMRL project with Dr. Michael Qin, discussed under Goal 3 below. We attempted to make use of the literature in which auditory localization accuracy was directly measured by having subjects indicate with a pointing response the perceived location of a sound source in the two dimensions of elevation and azimuth. Example studies are Oldfield & Parker (1984), Makous & Middlebrooks (1990), Carlile, Leong, & Hyams(1997), and Carlile, Delaney, & Corderoy (1999). By combining these results, we isolated three contributions to localization error, focussing first on perceived azimuth, the most immediately useful aspect of localization.

The first contributor is *bias* - there is a systematic tendency to perceive sounds located behind the head as forward of their true position, and vice-versa, to an extent that depends strongly and in a complex way, on the actual azimuth.The second contributor to localization error is *noise* - assuming that the biased location is the mean of a gaussian distribution, and the perceived location is sampled from this distribution, then a listener's response will vary according to the standard deviation of the distribution. Empirical results on response variability are much sparser than those for bias, so developing this model is not straightforward. The third contributor is *reversals*, in which a sound behind the head can reported as in front of the head, or vice-versa. We concluded that the most promising model for reversals was that there could simply be very high noise variability in the vicinity of 90° and 270°, so a significant number of stimuli are reported in the wrong front/back quadrant.

Our attempt to construct a mathematical model for the localization error was frustrated by the fact that measurements of variability in localization responses are very sparse in the empirical literature, and the few key papers that do report variability, do so with idiosyncratic graphic techniques rather than numerically. The 2014 Annual Report summarizes some confusion matrix data collected at NSMRL in their actual laboratory test-bed for using localized sound to aid visual search. During this period, a module was added to the EPIC architecture to make use of this confusion matrix data directly to support modeling their visual search task results.

Our discovery that the minimum audible angle (MAA) paradigm can provide estimates of variability of perceived location is a new opportunity for revisiting our models for location accuracy.

Goal 2: Modeling 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). One 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). Kieras and Hornof (2014) presented a model for the Williams task and then showed how the results could be incorporated into the easier-to-use GLEAN tool for GOMS modeling to produce usefully accurate results with much less modeling effort. The models 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. 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. In another collaboration, new 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). 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.

This data was analyzed and some preliminary modeling work conducted. A poster (Kieras, Hornof, & Zhang, 2015) presenting these results and preliminary modeling results was accepted for presentation at the ICCM 2015 conference. Since the analysis of this experimental data provided new levels of detail on complex visual search, the results are described in some detail here, followed by a summary of the modeling work thus far. In addition, various details in the experiment data led us to prepare an entirely new reduction of the data using a velocity-based algorithm to identify saccades and allow the eye tracking samples to be more reliably grouped into fixations. This reanalysis did not change the "big picture" of the data, but because the basic data analysis has removed some ambiguities. For example, the original data reduction produced a

significant number of "double fixations" - apparent artifacts in which there were two fixations on the same object with very short saccades and dwell times. Cleaning up these artifacts increases the clarify of the basic phenomena.

The 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.

Figure 14 shows one of the search fields. Each search field was preceded by the presentation of a cue that described the target in text and included the target's two-digit number and, depending on the condition, some combination of the the target's color, size, and shape, for example *red square 23*. Because each cue could include any combination of the three features, including none, there were a total of eight possible cue types. Search fields contained seventy-five objects on a gray background that subtended 39° by 30° of visual angle. Each object had a unique combination of color, size, and shape. Colors were blue, green, yellow, red, and purple. Sizes were small (0.8°), medium (1.6°), and large (2.8°), measured as the diameter of the circular object of that size, with other shapes normalized to the same area. Shapes were circles, semi-circles, squares, equilateral triangles, and crosses. Each object had a one-pixel black border.

The seventy-five unique objects were randomly distributed across the search field with at least one degree of visual angle between adjacent objects. A unique two-digit number from 01 to 75 appeared in the center of each object with a height of 0.26°. The cue appeared in the center of the display in the same typeface, with each feature listed on a separate line. Participants started each trial by clicking on an XX above the cue.

Each successful trial proceeded as follows: (1) The cue appeared in the center of the display. (2) The participant moved the mouse and clicked on the XX. (3) The cue disappeared and the search field appeared. (4) The participant found the target. (5) The participant moved the mouse and clicked on the number in the target. Participants were constrained to not move the mouse

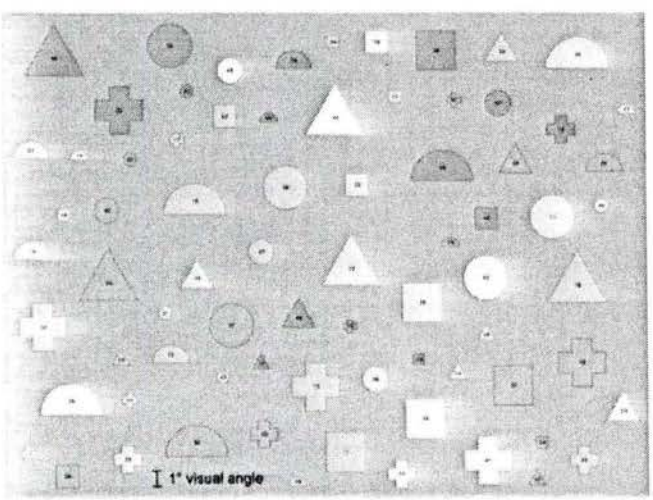

Figure 14. A sample search field used in the experiment.

until they found the target by using a point-completion deadline (Hornof, 2001). Participants practiced until they were comfortable with the deadline. Participants were financially rewarded for fast and accurate performance according to a payoff scheme.

Results

The error in the eye tracking data was reduced using the method of required fixations, as described in Zhang & Hornof (2014), yielding a series of fixations for each trial by each subject, for a total of about 64 thousand fixations.

Accuracy of fixations and fixation data. It is well-known that eye movement tracking results contain a substantial amount of both systematic and random error. Hornof and his students have developed elegant ways to eliminate much of the systematic error by using fixations on locations that are required by the task (e.g. the target object) as calibration points and adjusting the baseline and scale of the recorded locations so that they are correct for these required locations. These corrections were made in this data. However, there remains some variation in fixation locations. One symptom of this is that many fixations to do not land directly on one of the visual objects. What is the source of these in-between fixations?

To begin to characterize the in-between fixations, three different criteria for designating the fixated object were compared: (1) In the *Closest Object* criterion a fixation is counted as simply being on the object whose center is closest to the fixation point. This was used in an analysis previously summarized in an Annual Report and the ICCM poster (Kieras, Hornof, & Zhang, 2015). (2) The *Contained in Object* criterion: a fixation is counted as being on an object if it is within the circle corresponding to the object's nominal size. (3) Zhang and Hornof (2013) did their analyses in terms of an *Area of Interest* (AOI) criterion: a fixation was counted as being on a object if it was within a certain radius beyond the circle for the nominal size of the object; this radius was made as large as possible until overlapping AOIs began to appear, which was at a radius of 0.6°. They note that the 0.6° AOI "buffer" around the objects encompasses the mean residual eye movement error after applying their correction technique.

The Closest criterion, which is often used in eye movement studies, by definition produces zero fixations being classified as between objects, while the Contained criterion produces about 40% in-between fixations in these data, which means that a large number of fixations are not directly on part of the object. The AOI criterion produces about 20% *in-between* fixations, meaning that a substantial fraction of fixations are in-between objects, but because the AOI diameter depends on the object size, it is difficult to interpret some of the effects of object size. On the other hand, it provides a simple metric concerning saccade accuracy, and can yield less noisy results because fixations are only counted as being on a relatively nearby object. It is useful to note that of the fixations with an identified AOI object, almost all (about 99.5%) of these objects were also the Closest object, meaning that these two criteria are compatible for this subset of the fixations. In the results summarized here, the AOI criterion was used except where noted.

Fixations assigned to an object with the AOI criterion tended to be close to the center of the closest object, averaging 0.73° away from the center. This value is comparable to the 1° minimum distance between object bounds in the displays; it is also similar to the residual

measurement error, which we were able to estimate from the fixations on the required objects. This error is approximately gaussian, with means in the x- and y-dimensions of -0.02 and -0.07 degrees, and standard deviations of 0.30 and 0.35 degrees. Current work incorporates this measurement noise into the simulated experiment software. Alternatively, fixation error could be a result of noise in the oculomotor system, which will be considered below.

Analysis of effects. The proportion of fixations in which the properties of the fixated object matched the cue properties were calculated. Similar calculations were made for other statistics, such as the saccade distance - the difference between the current and previous fixation locations. These statistics were accumulated for each subject in each condition, and means computed for each condition by averaging over the 22 individual subject values. In the figures that follow, the observed mean values are plotted with 95% confidence intervals for the mean based on the 22 values from the individual subjects.

Replication of Williams (1967) effects. The results were consistent with those reported by Williams(1967). Figure 15 shows the proportion of fixations on objects that matched the cued properties. E.g., if the color was the only specified cue, about 70% of the fixations were on objects with the specified color. If the cue is all three properties, 56% of the fixations are on objects with the specified color, 38% on objects with the specified size, and 26% on objects with the specified shape. These fixated objects could match either a single cue, or two cues; only one object would match all three cues. The color cue produces the highest proportion of matches, followed by object size, whereas object shape produces the lowest proportion of matches. In the Number Only condition, whether the fixated object matches a property of the target object is equal to the distribution in the display (five colors and shapes, three sizes), which means that

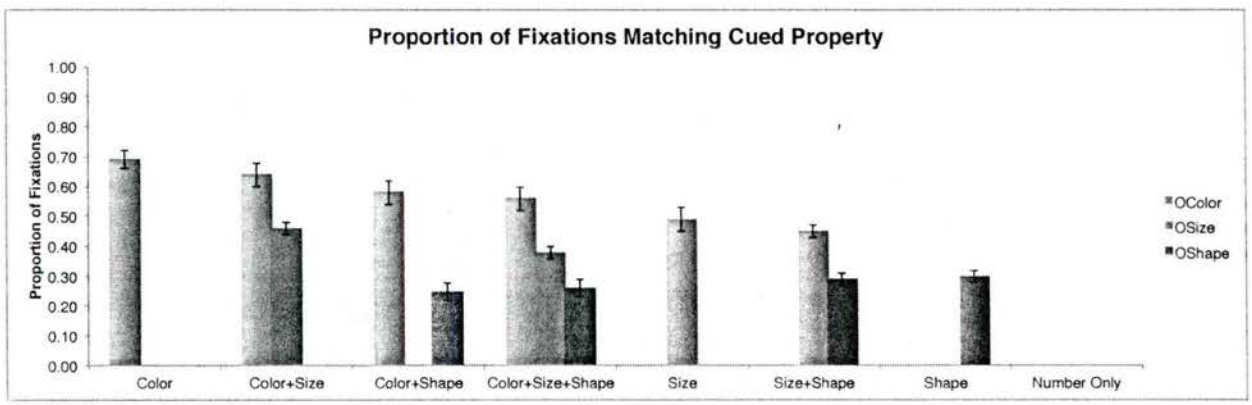

Figure 15. Proportion of fixations that match each of the cued properties in each cue condition. Fixations that match the cued color are shown in red; the cued size as green; the cued shape as blue. See text about the proportions in the Number Only condition.

fixations in this condition were random with regard to the color, size, or shape of the object.

These results replicate the Williams results quite well, showing that color is the most effective cue in guiding visual search, and shape is the least. But size appears to be more effective in these data compared to Williams, being similar to color, perhaps because there were only three different sizes, rather than four as in Williams that may have been difficult to

discriminate.

To further compare with Williams (1967), Figure 16 shows the number of fixations required to complete the task for each cue type. The color cue requires the fewest fixations, followed by size, then shape, with the Number Only cue requiring the most. These effects also basically replicate the Williams results, but are more precise due to better eye tracking methodology.

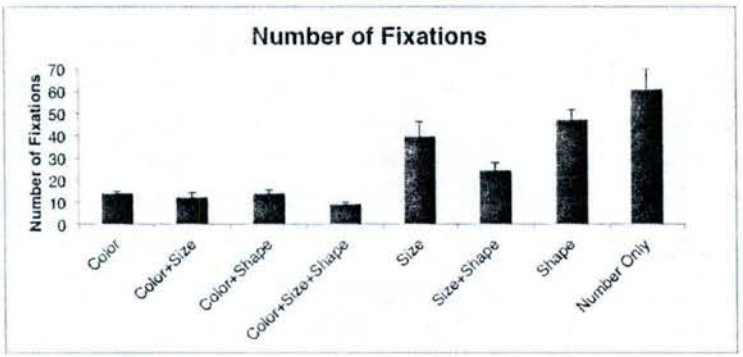

Figure 16. Number of fixations in each cue condition.

Additional visual search effects. These data have several new effects compared to the Williams data. The first is some precise information on revisit fixations; that is, fixations on an object that was previously fixated after some other object was fixated. Figure 17 shows the proportion of revisit fixations in each cue condition. Generally, it appears that the proportion of revisit fixations is proportional to the number of fixations, which would govern the total time required to locate the target. This is consistent with an earlier hypothesis that revisit fixations are a result of a loss from memory of information about previously inspected objects, and so the longer the task requires, the higher the proportion of revisit fixations.

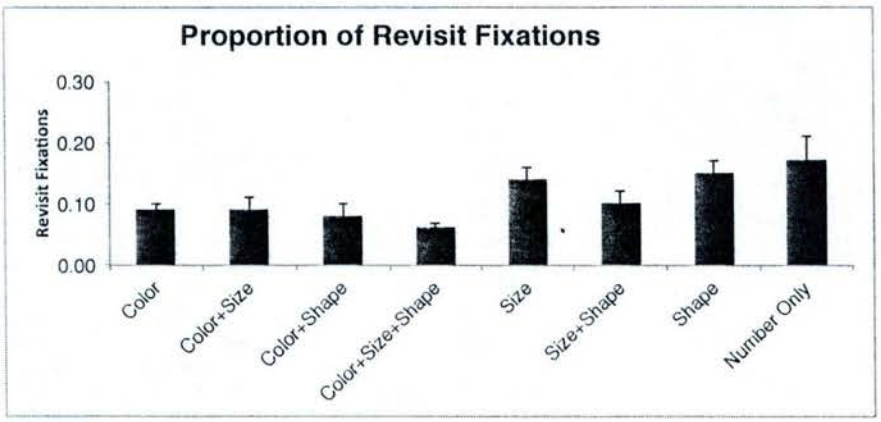

Figure 17. Proportion of repeat fixations in in each cue condition.

An important new effect in these data is an object size effect. If a visual property such as color guides the visual search, and the availability of this property outside the fovea depends on the size of the object, then the likelihood of fixating an object should be greater for larger objects

than for smaller. Figure 18 shows this effect. In left-to-right order are shown the proportion of fixations on small, medium, and large objects in each cue condition regardless of whether they match or mismatch the cue. There is a clear effect in that large objects are fixated more often than small, with medium size objects lying in between. The only exception to this pattern is the Size-only cue where there is no difference between small and medium sizes. The effect is quite large when color is one of the cues, and smallest with the Shape- and Number-only cues, but it is still present to some extent.

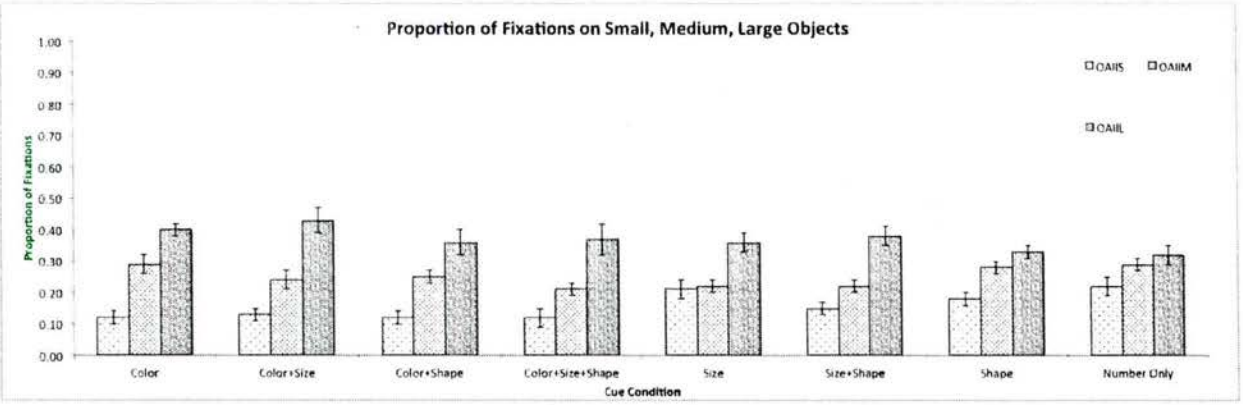

Figure 18. Proportion of fixations on objects of small, medium, and large size (left to right) in each cue condition, regardless of whether the object matches the cue.

Another new effect in these data concerns the saccade distance. If a cue is more effective than another in guiding visual search, the corresponding property of an object should be visible at a greater eccentricity, meaning that saccades should be longer on the average for more effective cues. Figure 19 shows this effect; color cues produce the longest saccades, ranging from 6.84° to 7.31°, followed by size and shape ranging from 5.50° to 6.35°, then Number Only at 5.01°. The effect is fairly small, less than 2°, but reliable, as shown by the non-overlapping confidence intervals. The small size of the effect could be due to averaging over saccades to both matching and mismatching objects.

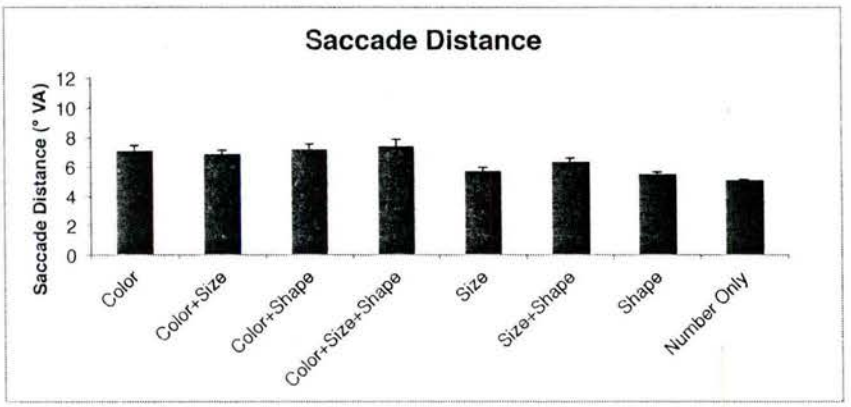

Figure 19. Average saccade distance (length) in each cue condition.

Accordingly, Figure 20 compares the saccade distance when the fixated object matches the cued color with when it mismatches. The mismatch saccades are considerably longer than the shorter ones (8.85° vs 6.39°) , and apparently more variable as well. In contrast, a difference between matching and mismatching target saccade distances is much smaller for the cued size, as shown in Figure 21 and for the cued shape, shown in Figure 22, being less that 0.5° in both cases. The size and shape distances average 6.56°, which is similar to the color matching distance.

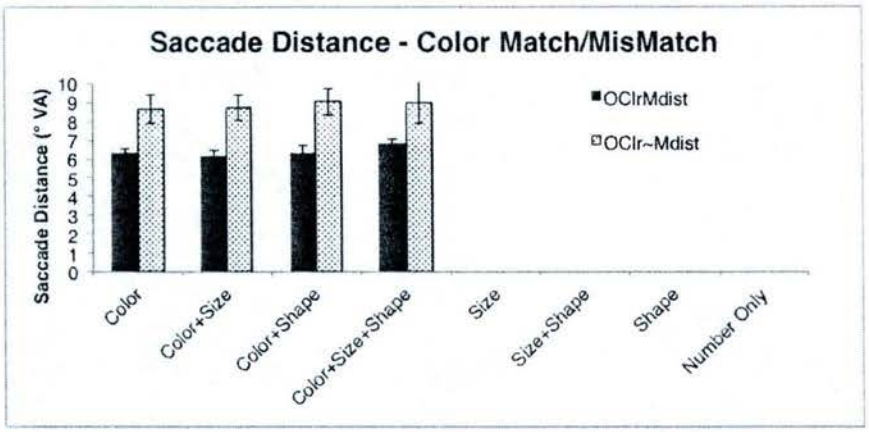

Figure 20, Saccade distances to objects that match the cued color (black bars) and that mismatch the cued color (gray bars). Closest-object criterion was used.

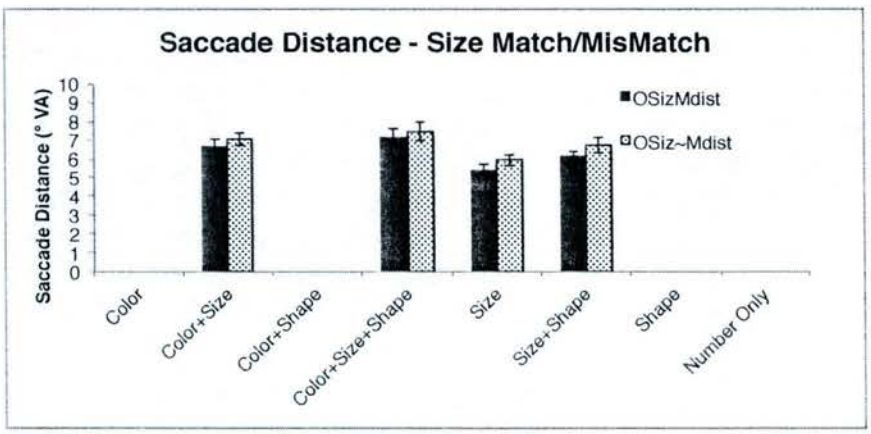

Figure 21, Saccade distances to objects that match the cued size (black bars) and that mismatch the cued size (gray bars). Closest-object criterion was used.

Given that most color-cue saccades are to a matching object, the result is that the overall average saccade distances in Figure 19 are very similar.

Modeling work

The detailed data analysis was followed by an effort to determine whether the EPIC architecture could account for these detailed eye movement effects; it was immediately apparent that two modifications to the architecture were required.

Revised eye movement model. In the past, EPIC's saccade to an object was always accurate -

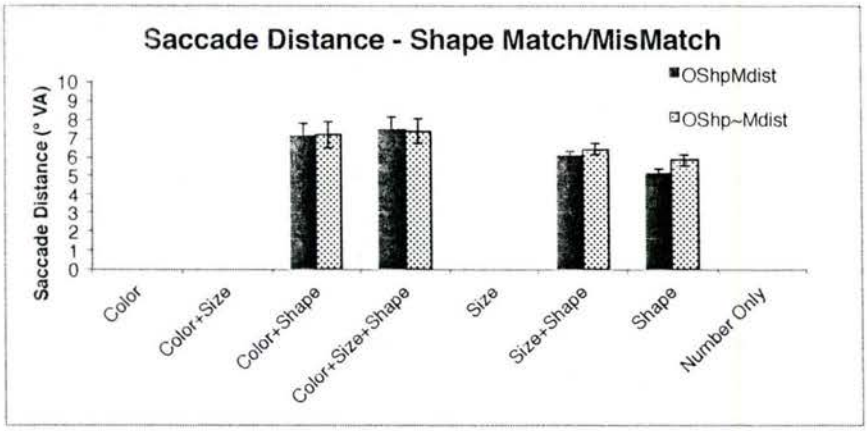

Figure 22, Saccade distances to objects that match the cued shape (black bars) and that mismatch the cued shape (gray bars). Closest-object criterion was used.

the eye always landed exactly at a specified location, such as the center of the target object. The data clearly show that this is incorrect. The properties of saccade accuracy need to be characterized. A variety of studies (see Abrams, Meyer, & Kornblum (1989) and the review in Harris, 1995) 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. Thus, the revised 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%. Unlike previous EPIC models for this visual search task, a model using this revised oculomotor processor often misses the object to be fixated, which decreases the probability that (e.g.) its text label will be available. The task strategy must either attempt to fixate the object again, or choose a entirely different object to fixate. On the other hand, if the acuity functions are such that most fixations are close enough that the relevant properties are available, there may be little effect of inaccurate saccades.

*Revised acuity functions.*The form of the acuity functions used in the eye processor were modified to align better with the available empirical data. Previously, EPIC used simple forms of acuity functions that were adequate to fit the limited data such as Williams(1967). The new work here was to anchor the acuity functions closer to the available psychophysical data. Of special interest are studies of "cortical magnification" which is based on the reasoning that a constant amount of visual cortex (presumably supporting a certain number of receptive fields) are 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 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, and then quite sharply in the further periphery. A cubic function with a moderate linear coefficient, a zero quadratic coefficient and a very small cubic coefficient provides a good fit.Visual search studies such as Carrasco & Frieder (1996) 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 measured functions, search time becomes flat with eccentricity. However, it appears that magnification functions measured for individual objects greatly overestimate the acuity for objects in dense visual fields (e.g. see discussion in Anstis, 1974). To measure acuity in dense displays would be very difficult, and the literature does not contain useful parametric studies.

To deal with this non-definitive picture, a simple family of acuity functions were proposed, and their parameters determined by a combination of general constraints set by the literature and iterative maximization of fit in the models. A separate function was specified for each property: color, encoded size (small, medium, large), shape, and text label. The acuity function is a Gaussian detection function that gives the probability that the property will be available (or detected) for an object with size s at eccentricity e:

$$P(detection) = P(s > N(\mu, \sigma)),\ \mu = a + be + ce^2 + de^3$$

The form for μ (which can be interpreted as the 50% threshold for object size) reflects the commonly fitted cubic form of cortical magnification functions. The value of σ governs the steepness of the ogival detection function; smaller values of σ make it look more like an all-or-none threshold-like process. To reduce the number of parameters to be fit in the models described here, the a term was held at 0.05, and c was held at 0. The b term describes the availability near the fovea, and the d term describes how quickly availability declines in the periphery. To give a sense of the values of these parameters, some typical fits estimated b in the range of 0.1, $d = 0.0005$ for color, 0.025 for shape, 0.05 for text, and σ was 0.5 for color, size, and shape, and 1.0 for text.

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 the same object can fluctuate, and will not be reliably available from one fixation to the next. However, the information, once acquired, will remain for some time in the perceptual store.

Revised task strategy. Two important changes from the previous Williams task model were made to the production system strategy. First, unlike before, the strategy *continuously* nominates candidate objects to inspect, so that the next object can be chosen using any new information resulting from the just-completed eye movement. Second, the rules are optimized. Early models seemed to fit the data satisfactorily with suboptimal rule sets, but it seems that a better research tactic is to assume an optimum strategy so that performance is unambiguously limited by the available visual information. Thus the next object to fixate is chosen to be one that matches the most or most diagnostic cues. If no object matches a cue, then a "guessed" object is selected in which the cued properties are unknown, so that fixations are not wasted on objects known to be

incorrect. Note that in the Number-Only case, there are no cued properties, so any object whose label is currently unknown is a potential candidate. Current work has been explored various strategies for making "guesses" given these optimizations, taking into account that the guessing strategy may depend on the cue condition.

Modeling results summary. The current models with the modified architecture are at least as good as the previous models in capturing the basic effects of the different search cue conditions. For example, the preliminary model in Kieras, Hornof, and Zhang (2015) accounted for data in Figure 15 above, the different proportions of fixation on objects depending on the search cue (r^2=0.89) , and the corresponding effects (Figure 16) on number of fixations (r^2=0.92). However, the real goal with these results to move to a finer grain of modeling detail that confirms the architectural additions and the more sophisticated visual search strategy. In addition, all of the effects should be accounted for by a model with a single task strategy and set of architectural parameters. Many models and parameter fits have been tested, but unfortunately, at the time of this writing, a fully satisfactory model and parameter value set has been found, even though there has been some success at accounting for individual effects that are new in this data.

One purely methodological result from this work is that even the large-scale cluster parameter search facility provided by mindmodeling.com is not up to the task of grid searches over the entire parameter set outlined above; apparently, compared to the models typically used with mindmodeling.com, because EPIC models include many perceptual and motor parameters, their parameter space is too large to be handled by simple grid searches. In addition, an ideal model fitting process would fit multiple dependent measures (like those shown in the above graphs) simultaneously; the techniques for this are poorly explored in the cognitive modeling literature. This experience provides a counterexample to the common perception that complex cognitive architecture models can always be made to fit any data set – getting a model that fits complex data is difficult! In fact, the purpose of a cognitive architecture is to constrain the possible models for a set of data, so the failure to find even one good model thus far means that the architecture constraints and the data are working together to very strongly constrain the possible models, so much so that the difficulty of constructing an adequate model conveys useful information.

The object-size and saccade-length effects. Figure 18 above shows that larger objects are fixated more often than smaller objects. At first glance, this effect should be a simple consequence of how the acuity functions work - the properties of larger objects are more available than for smaller objects. Some parameter settings produce this effect for some of the cue conditions, but not for others. Analysis of the model processes give some clues about the problem. First, the range of object sizes is not very large, and the lengths of the saccades tends to be only moderate (see Figure 19 above). Under these conditions, the acuity functions would have to be very "steep" to produce differences due to property availability.

A more subtle problem is similar to that underlying the Figure 19, and shown in the subsequent match-mismatch saccade distances. Any given saccade to an object is either a *guided* saccade, meaning that the object matches one or more of the cue properties and these properties were available, which would depend on the size of the object as well as its eccentricity, or it is a

"guessed" saccade, meaning that some potentially arbitrary or random choice was made to fixate that object. Since a guessed object might match the cued properties by chance, only fixations to non-matching objects provide information about the guessing strategy. It is possible that the effects of property availability will be diluted by the guessing strategy, explaining why the object size effect is elusive.

As mentioned above, an optimal guessing strategy requires that the cued properties of a qualified object be unknown, but that potentially covers a large number of objects in the display, and only one can be chosen to fixate at a time. Two simple guessing alternatives are to choose the closest qualifying object, or to choose a qualifying object at random from the whole display, which on the average would be further away. At present, it appears that the guessing strategy should depend on the cued properties. For example, since the text label in the Number-Only cue condition is generally unavailable from any distance away, almost all saccades in this condition will be guessed saccades, and they might as well be the closest object with an unknown label. This results in relatively short saccades in this condition. In contrast, in a Color cue condition, since the color is widely available, a guessed saccade will be to an object whose color is unknown, meaning that it is relatively far away. Clearly, the effects of object size are dependent on the guessing strategy as well as the guided saccade strategy, meaning that even straightforward properties of the data depend on the model strategy in a subtle way.

Currently, models that make a guessed saccade to the closest object tend to under-predict the mismatching saccade length, and those that make a guessed saccade to an object chosen at random from the display tend to badly over-predict the mismatching saccade length. A similar problem was noted in the models reported in Kieras & Marshall (2006).

Possible architectural solution. As also suggested by the Kieras & Marshall (2006) results, it might be time to modify an important simplifying assumption in EPIC's visual architecture. This is the assumption that all of the objects in the visual field are known to be present or "visible" - it is only their visual properties, such as color or shape, that might not be available. Since all the objects are present, it is simple for a strategy to select one of the them to be fixated, which would then make all its properties available. Likewise, it is simple to ensure that all objects get fixated eventually – simply fixate any objects whose properties are unknown. However, perhaps objects are known to present or visible in a way similar to their other properties - whether an object is available to serve as a fixation target might depend on its eccentricity or size, just like its color or shape. Making this change in the architecture would be simple enough; the problem is how to make it possible for a task strategy to choose to look at an area of the display that currently contains no visible or targetable objects - this would require a "philosophical" change in how the ocular motor processor can be commanded and how the visual space is represented.

One approach would be to incorporate the common concept of *salience maps* or the related concept of *proto-objects*. Both of these mechanisms provide an explanation for the so-called *center of gravity* effect, in which saccades are made to a location that is in-between a group of objects with the sought-for property. The salience map concept holds that the visual field is represented as regions in which a feature such as color has a particular value, such as a red area. If a red object is sought, then a fixation could be made into the red region, and then the

individual objects could be resolved. Alternatively, perhaps these regions are themselves represented like objects in low resolution; for example, a cluster of small red objects at a distance might be represented a single large red object that could be the target of a saccade; following a fixation on this proto-object, the individual objects could then be resolved.

In terms of the visual search task, if a guessed saccade is made, it would be to a salient region or proto-object rather than to an individual object. This would make it possible for guessed saccades to be guided by visual features in a way that this different from the current architecture, and thus might be made consistent with the data on guessed saccade sizes.

Adding a salience map or proto-object layer to the visual representation would be a major architectural change; while it might allow some additional phenomena to be accounted for, it also increases the complexity of the architecture and brings in additional parameters that would have to be estimated from the fragmentary available data.

Goal 3. Application to Military Tasks

This work would be considered a examples of technology transfer, and will be listed under that heading below. However, it was made an explicit goal of the project to ensure that the time and resources would be available for such work, reflecting a commitment that the scientific development of cognitive architectures such as EPIC fundamentally benefits from attempts to apply them to actual practical problems in system design or human performance analysis; in addition to clarifying what parts of the architecture are successful and which need to be built out or improved, such work acts as a counterbalance to the often insular and parochial concerns of purely academic research.

Auditory-aided Visual Search

As described in the Annual Reports, collaborative work was done with Dr. Michael Qin's group at NSMRL and a CRADA is formally in place to allow data sharing. The basic problem: There is developing a new generation of submarine periscopes based on 360° digital cameras; this provides a high-resolution panoramic view, but presenting this view on display compatible with existing space in the submarine control room is a problem. Even a 360° visual display is not necessarily a good idea; although it preserves the spatial compatibility of the original optical periscope, it would in effect preserve the problem that has to be "walked around" to view the whole horizon. In addition, the physical space requirements of such a display is likely to be impractical for a long time. So one set of design issues is how to map the output of the 360° camera to one or more standard-size flat displays; reasonable comprises would involve some combination of panoramic view and one or more "zoomed in" views. In addition, perhaps localized sound information, such as that from sonar, might improve the effectiveness of the display. Qin has defined a critical periscope task, that of checking for nearby vessels prior to surfacing; fast and accurate performance is important for safety. Sonar information can help identify the bearing for likely close vessels, so looking there first would improve performance. However, this involves an additional problem of how the user can map a sound appearing at a relative egocentric bearing of say 135° to a particular area of the panoramic display, and whether both the perceived sound location and this translation will be reliable enough to actually support

the task.

Qin believes that modeling could be used to explore some of the consequences of different design decisions to help guide and interpret empirical studies. To this end, the EPIC software and some "starter" models have been provided to his group, and the PI (Kieras) has made multiple visits there to advise on how to do the modeling, and the EPIC architecture and a basic version of the task model have been installed at NSMRL. The work involved adding a component to the EPIC architecture that captures the empirical effects on accuracy of sound localization, which acts as a limit on the value of a localized sound cue, as well as a component that represents the time and accuracy aspects of recoding an auditory azimuth to a location on the visual display. To support using empirical localization accuracy data in the form of a confusion matrix, the EPIC software was augmented with a delightful generator for random sampling from an arbitrary discrete distribution, and this was provided to the NSMRL group.

Goal 4. Software Maintenance and Distribution

A rewrite of the EPIC GUI code into the modern "Cocoa" API was completed during this project, and the current version has been stable for some time. Some contributions to this effort have been made by Yunfeng Zhang at University of Oregon (see Zhang and Hornof, 2014). Another long-time EPIC user, Travis Seymour at U. has been working on a new Windows port of EPIC.

The Auditory processor components involved in the multichannel speech processing and localization models were developed to a relatively stable state to support additional modeling work. This software was distributed to collaborators as needed.

A small but significant extension was made to the production-rule language - this is a new rule action, `Add_with_probability`, action that makes it much simpler to implement mixture models

A version of the EPIC software was developed that is tailored for running on the MindModeling.org clusters. We have been granted access to some Air Force computational clusters (part of the Mindmodeling project) by the Cognitive Modeling and Agents Branch at WPAFB, Kevin Gluck being our POC. This has been used for the two-talker task models and for the visual search models. The cluster software include build facilities that will build the architecture and model in any Standard C++ environment (such as gcc under Linux), and a customizable top-level driver module that maps the parameter values provided by the cluster system into the architecture, and computes goodness-of-fit values that returned to the system to guide the parameter space search. Because of its relative complexity in modeling perception and action as well as cognitive processing, this work is pushing out the envelope on cluster-based model fitting, and has led to some useful methodological discussions with Kevin Gluck.

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.

Assmann, P.F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *c*

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.

Bregman, A. S. (1990). *Auditory scene analysis: the perceptual organization of sound*. Cambridge, MA: MIT Press.

Bronkhorst, A.W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86, 117-128.

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.

Brungart, D.S. & Simpson, B.D. (2004). Optimizing the performance of multitalker speech displays. Presented in the *Symposium on Speech Separation and Comprehension in Complex Acoustic Environments*, Nov 4 - 7, 2004, Montreal, Quebec.

Brungart, D.S. & Simpson, B.D. (2005). Optimizing the spatial configuration of a seven-talker speech display. *ACM Transactions on Applied Perception*. 2, 430-436.

Brungart, D.S., Simpson, B.D., Ericson, M.A., & Scott, K.R. (2001) Informational and energetic masking effects in the perception of multiple simultaneous speakers. *Journal of the Acoustical Society of America*, 110 (3), 1101-1109.

Carlile, S., Leong, P., Hyams, S. (1997). The nature and distribution of errors in sound localization by human listeners. *Hearing Research*, 114, 179-196

Carlile, S., Delaney, S., Corderoy, A. (1999). The localisation of spectrally restricted sounds by human listeners. *Hearing Research*, 128, 175-189.

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

Cherry, E.C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *Journal of the Acoustical Society of America*, 25 (5): 975–79

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

Darwin, C.J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, 1(9), 327-333.

Fant, G. and Kruckenberg, A. (1996). On the Quantal Nature of Speech Timing. *Fourth International Conference on Spoken Language Processing, ICSLP 96*. SuA1L3.3.

Findlay, J. (1997). Saccade target selection during visual search. *Vision Research*, 37, 617–631.

Harris, C.M. (1995). Does saccadic undershoot minimize saccadic flight-time? A Monte-Carlo study. *Vision Research*, 35, 691-701.

Hartman, W.M., & Rakerd, B. (1989). On the minimum audible angle – A decision theory approach. *Journal of the Acoustical Society of America*, 85 (5), 2031-2041.

Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation*, 17, 1875-1902.

Hopkins, K., & Moore, B.C.J. (2009). The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *J. Acoust. Soc. Am.* 125(1), 442-446.

Kidd, G., Arbogast, T.L., Mason, C.R., & Gallun, F.J. (2005). The advantage of knowing where to listen. *Journal of the Acoustical Society of America*, 118(6), 3804-3815.

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 & 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., & 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. & Marshall, S.P. (2006). Visual availability and fixation memory in modeling visual search using the EPIC architecture. In *Proceedings of the Annual Cognitive Science Society Meeting*, July 26-29.

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.

Kieras. D.E. & Santoro, T.P. (2004). Computational GOMS Modeling of a Complex Team Task: Lessons Learned. In *Proceedings of CHI 2004: Human Factors in Computing Systems*. New York: ACM, Inc. 97-104.

Kieras, D. & Wakefield, G. (2014). Developing Models for Multi-Talker Listening Tasks using the EPIC Architecture: Wrong Turns and Lessons Learned. Technical Report EPIC-17, University of Michigan Electrical Engineering and Computer Science Department. Publicly available at University of Michigan Deep Blue repository at http://hdl.handle.net/2027.42/108165

Kieras, D.E., Wakefield, G.H., Brungart, D.S., & Simpson, B.D. (submitted). A preliminary cognitive-architectural account of spatial separation effects in two-channel listening accounts. Submitted to *The Human Factors and Ergonomics Society 2016 International Annual Meeting.*

Kieras, D.E., Wakefield, G.H., Thompson, E., Iyer, N., and Simpson, B.D. (2014). A cognitive-architectural account of two-channel speech processing. In *Proceedings of the 2014 International Annual Meeting of the Human Factors and Ergonomics Society,* Chicago, October 27-31, 2014.

Kieras, D.E, Wakefield, G.H., Thompson, E., Iyer, N., Simpson, B.D. (2015). Modeling two-channel speech processing with the EPIC cognitive architecture. *Proceedings of the 13th International Conference on Cognitive Modeling (ICCM 2015),* Groningen, The Netherlands, April 9-11, 2015.

Kieras, D.E, Wakefield, G.H., Thompson, E.R., Iyer, N., Simpson, B.D. (2016). Modeling two-channel speech processing with the EPIC cognitive architecture. *Topics in Cognitive Science*, 8, 291–304. DOI: 10.1111/tops. 12180.

Wakefield, G.H., Kieras, D., Thompson, E., Iyer, N., Simpson, B.D. (in preparation). A cognitive-architectural approach to modeling listener strategy: An EPIC account of performance in speech-on-speech masking experiments. To be submitted to the *Journal of the Acoustical Society of America.*

Lee, J.H., & Humes, L.E. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *Journal of the Acoustical Society of America*, 2012, 132(3), 1700-1717.

Leachtenauer, J.C. (2003). Resolution requirements and the Johnson criteria revisited. In G.C. Holst (Ed), I*nfrared Imaging Systems: Design, Analysis, Modeling, and Testing XIV, Proceedings of SPIE* Vol. 5076.

Makous, J.C., & Middlebrooks, J.C. (1990). Two-dimensional sound localization by human listeners. *J. Acoust. Soc. Am.* 87 (5), 2188-2200.

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.

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

Mills, A.W. (1958). On the minimum audible angle. *Journal of the Acoustic Society of America*, 30(4), 237-246.

Moore, B.C.J, & Gockel, H.E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society,* 367, 919-931.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 27, 56-60.

Oldfield, S.R, & Parker, S.P.A. (1984). Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions. *Perception*, 13, 581-600.

Perrott, D.R. (1984). Concurrent minimum audible angle: A re-examination of the concept of auditory spatial acuity. *Journal of the Acoustic Society of America*, 75(4), 1201-1206.

Santoro, T.P., Kieras, D.E., & Pharmer, J.A. (2004). Verification and validation of latency and workload predictions for a team of humans by a team of computational models. *U.S. Navy Journal of Underwater Acoustics, Special Issue on Modeling and Simulation*, 54, 281-304.

Schneider, B.A., Li, L., & Daneman, M. (2007). How competing speech interferes with speech comprehension in everyday listening situations. *Journal of the American Academy of Audiology*, 18, 478-591.

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.

Spieth, W., Curtis, J.F., & Webster, J.C. (1954). Responding to one of two simultaneous messages. *Journal of the Acoustical Society of America*, 26(3), 391-396.

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. DOI: 10.1121/1.4928395

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

Virsu, V. & Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, 37, 475-494.

Wakefield, G. (2014). Keyword Co-articulatory Boundary Segmentation of the CRM Speech Corpus. Database, Publicly available at University of Michigan Deep Blue repository at http://hdl.handle.net/2027.42/108223

Wakefield, G.H., Kieras, D., Thompson, E., Iyer, N., Simpson, B.D. (2014). EPIC modeling of a two-talker CRM listening task. In *Proceedings of the 20th International Conference on Auditory Display (ICAD-2014)*, New York, June 22-25, 2014.

Wakefield, G.H., Kieras, D., Thompson, E., Iyer, N., Simpson, B.D. (in preparation). A cognitive-architectural approach to modeling listener strategy: An EPIC account of performance in speech-on-speech masking experiments. To be submitted to the *Journal of the Acoustical Society of America.*

Wichman, F.A., & Hill, N.J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8), 1293-1313.

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.

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.

Zelinsky, G.J., Rao, R.P.N, Hayhoe, M.M, and Ballard, D.H. (1997). Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science*. 8(6) 448-453.

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.

VI. Technology Transfer

Auditory Aiding of Visual Search

This work with Dr. Qin's group at NSMRL has been summarized above; this is a clear transition project, reinforced with a formal CRADA (Agreement Number NCRADA-NSMRL-13-9182).

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). This group was previously led by Dr. Doug Brungart, who is now at Walter Reed Army Medical Center. The two AFRL scientists involved are Dr. Nandini Iyer and Dr. Brian Simpson, who have been involved in over10 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.

This interaction resulted in a Seedling proposal funded by AFRL to collect, analyze, and model new data for the speaking-while-listening (Broadbent) task that was a topic in the previous ONR project. We still have under review a proposal in parallel with one by the AFRL WPAFB group for a joint project to do further empirical and modeling work on multichannel speech processing, but with a main focus on the speaking-while-listening problem.

Human-in-Control Modeling

This period of this six-month STTR project was 5/17/13 - 8/15/13. It was sponsored by MDA, with the prime contractor being Intelligent Automation, Inc. This project is ITAR restricted and so will be only lightly summarized here. The context is large-scale simulations for modeling the C2 aspects of BMD system, which need simulated human operators for evaluation and training purposes. Thus the problem is to have realistically-performing simulated humans integrated into these larger simulation systems. Another system, GLEAN, was used in this work.

Under earlier DOD and ONR support, GLEAN was developed as a simplified cognitive architecture modeling tool, using the well-established GOMS model concept instead of production rules for representing human procedural knowledge. GOMS models are much easier to develop for routine computer interaction tasks than production rule models. GLEAN was usefully applied to the analysis of teams using a new CIC watchstation concept for Anti-Air Warfare (Santoro, Kieras, & Pharmer, 2004; Kieras & Santoro, 2004). The GLEAN and EPIC work have informed each other, and share many software components, and so this project, though separated supported, contributes to the development of EPIC. For example, the details of implementing models for teams of simulated humans were developed for GLEAN and are being used in this project, and can then be ported to EPIC easily.

VII. Productivity

A. Refereed Journal Articles

Kieras, D.E., & Hornof, A. (in press). Cognitive architecture enables comprehensive predictive models of visual search: Commentary on Hulleman & Olivers. *Brain & Behavioral Sciences.*

Kieras, D.E, Wakefield, G.H., Thompson, E.R., Iyer, N., Simpson, B.D. (2016). Modeling two-channel speech processing with the EPIC cognitive architecture. *Topics in Cognitive Science*, **8**, 291–304. DOI: 10.1111/tops.12180.

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. DOI: 10.1121/1.4928395

Wakefield, G.H., Kieras, D., Thompson, E., Iyer, N., Simpson, B.D. (in preparation). A cognitive-architectural approach to modeling listener strategy: An EPIC account of performance in speech-on-speech masking experiments. To be submitted to the *Journal of the Acoustical Society of America.*

B. Non-Refereed Significant Publications

None during the period covered by this project.

C. Books or Chapters

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., & Butler, K.A. (2014). Task analysis and the design of functionality. In H. Topi & A. Tucker (Eds.) Computing Handbook: Information Systems and Information Technology, Chapman and Hall/CRC Press. Print ISBN: 978-1-4398-9854-3, eBook ISBN: 978-1-4398-9856-7. (pp. 33-1 - 33-26).

D. Technical Reports

Kieras, D. & Wakefield, G. (2014). Developing Models for Multi-Talker Listening Tasks using the EPIC Architecture: Wrong Turns and Lessons Learned. Technical Report EPIC-17, University of Michigan Electrical Engineering and Computer Science Department. Publicly available at University of Michigan Deep Blue repository at http://hdl.handle.net/2027.42/108165

Wakefield, G. (2014). Keyword Co-articulatory Boundary Segmentation of the CRM Speech Corpus. Database, Publicly available at University of Michigan Deep Blue repository at http://hdl.handle.net/2027.42/108223

E. Workshops and Conferences

Kieras, D. (2014). Searching and listening: EPIC models for perceptually intensive tasks. Invited talk presented at the AFRL Cognitive Modeling Brownbag, sponsored by AFRL Wright-Patterson AFB, Dayton OH, June 10, 2014.

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.* (Best Paper award).

Kieras, D.E., Wakefield, G.H., Brungart, D.S., & Simpson, B.D. (submitted). A preliminary cognitive-architectural account of spatial separation effects in two-channel listening accounts. Submitted to T*he Human Factors and Ergonomics Society 2016 International Annual Meeting.*

Kieras, D.E, Wakefield, G.H., Thompson, E., Iyer, N., Simpson, B.D. (2014). A cognitive architectural account of two-channel speech processing. In *Proceedings of the Human Factors and Ergonomics Society 2014 International Annual Meeting*, Chicago, October 27-31, 2014.

Kieras, D.E, Wakefield, G.H., Thompson, E., Iyer, N., Simpson, B.D. (2015). Modeling Two-Channel Speech Processing with the EPIC Cognitive Architecture. *Proceedings of the 13th International Conference on Cognitive Modeling (ICCM 2015),* Groningen, The Netherlands, April 9-11, 2015.

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.

Wakefield, G.H., Kieras, D., Thompson, E., Iyer, N., Simpson, B.D. (2014). EPIC modeling of a two-talker CRM listening task. In *Proceedings of the 20th International Conference on Auditory Display* (ICAD-2014), New York, June 22-25, 2014. (Received Best Paper Award.)

F. Patents

None during the period covered by this project.

G. Awards/Honors

ACM CHI 2014, Best Paper Award for Kieras & Hornof (2014).
ICAD 2014 Best Paper Award for Wakefield, Kieras, Thompson, Iyer, & Simpson (2014).

Final Report Distribution List

Paul Bello (Code 341)

Office of Naval Research

875 N. Randolph St.

Arlington, VA 22203-1995

Office of Naval Research

Regional Office - Chicago-N62880

230 South Dearborn, Room 380

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