# EPIC MODELING OF A TWO-TALKER CRM LISTENING TASK

*Gregory H. Wakefield, David Kieras*

University of Michigan
Computer Science and Engineering
Ann Arbor, MI, USA
{ghw, kieras}@umich.edu

*Eric Thompson, Nandini Iyer, Brian D. Simpson*

Air Force Research Laboratory
Wright Patterson Air Force Base
Dayton, OH, USA
eric.thompson.22.ctr@us.af.mil
nandini.iyer.2@us.af.mil
brian.simpson.4@us.af.mil

## ABSTRACT

An extension of the auditory module in EPIC is introduced to model the two-talker coordinate response measure (CRM) listening task. The construct of an auditory stream is employed as an object in the working memory of EPIC's cognitive processor. Production rules are developed that execute the two-talker CRM task. Analysis of these rules reveal two sources of possible error in the output of the auditory processor to working memory. Each is explored in turn and the production rules modified to provide a corpus-driven model that accounts for human performance in the listening task.

## 1. INTRODUCTION

A common problem in the design of auditory displays is how to manage multiple sources so that the user can maximize the information gained from each acoustic source. The coordinate response measure (CRM) speech corpus was designed to assess the limits of human performance in listening to multi-talker environments [1]. From studies using the corpus, much has been learned about the factors that contribute to degraded performance. Nevertheless, strong predictive models have been a challenge, in part because of a paucity of tools for modeling the complex auditory task. Working from the stimulus forwards, the formulation of receivers (ideal or otherwise) using the standard tools of signal detection theory can fail due to the complexity of the auditory task, the complexity of the stimuli, or both. In contrast, cognitive architectures are adept at modeling complex human tasks at a more global level. The present paper explores the use of one such cognitive architecture, EPIC, to model human performance in a two-talker listening task. EPIC (Executive/Process-Interactive Control) is one among several architectures whose goal is to provide a comprehensive account of human abilities and limitations in perception, cognition, and action ([2, 3, 4]. Following a review of the two-talker CRM listening task, an overview of EPIC will be presented and key extensions of the auditory processing module will be introduced. Within the framework imposed by these extensions, a strategy for the two-talker CRM listening task will be proposed, modeled, and fit to the human data.



## 2. REPLICATION OF A TWO-TALKER CRM LISTENING TASK

For over a decade, the *coordinate response measure* (CRM) has been used to study the perception of two or more temporally-overlapping speech signals. The CRM corpus is a collection of commands

"Ready {Call sign} go to {Color} {Number} now"

spoken by one of four females or four males, where the *Call sign*, *Color*, and *Number* 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 *Call Signs*, *Colors* and *Numbers* [1].

In the two-talker listening task, participants respond to verbal commands from the CRM corpus by selecting the appropriate element from a matrix of colored blocks containing numbers which is presented on a visual display. On any given trial, a *target* is drawn from those utterances bearing the *Baron* call sign and presented simultaneously with a randomly selected *masker*, with the restriction that the *Call sign*, *Color* and *Number* of the *masker* differ from those of the selected *target*. Both target and masker utterances are presented diotically over headphones. A color response is scored *target*, and, therefore, correct, if the participant selects an element bearing the *target* color; similarly, a number response is scored *target* if the participant selects an element bearing the *target* number. When an error occurs, it is classified either as a *masker* if the color (or number) reported agrees with that of the *masker* or as *neither* if the color (or number) reported differs from those of the target and masker.

The six panels of Figure 1 show the initial results from a replication of a study that was originally published by Brungart and his colleagues in 2001 [5]. Each panel plots the probabilities of target (blue), masker (red), and neither (green) responses as a function of the target-to-masker ratio (TMR) in dB[1]. In addition, the joint probabilities of target color and target number are shown in black. The upper and lower rows display the results for *color* and *number* responses, respectively. The columns display the results based on the gender and identities of the target and masker talkers. From left to right, the stimulus conditions are different genders (TD), same gender but different talkers (TS), and same talker (TT). 95%

[1]The data represent average thresholds for each condition based on 10 subjects. For each subject, threshold probabilities were determined based on 40 trials/condition. The standard error of the mean is reported as the variability across subject means.

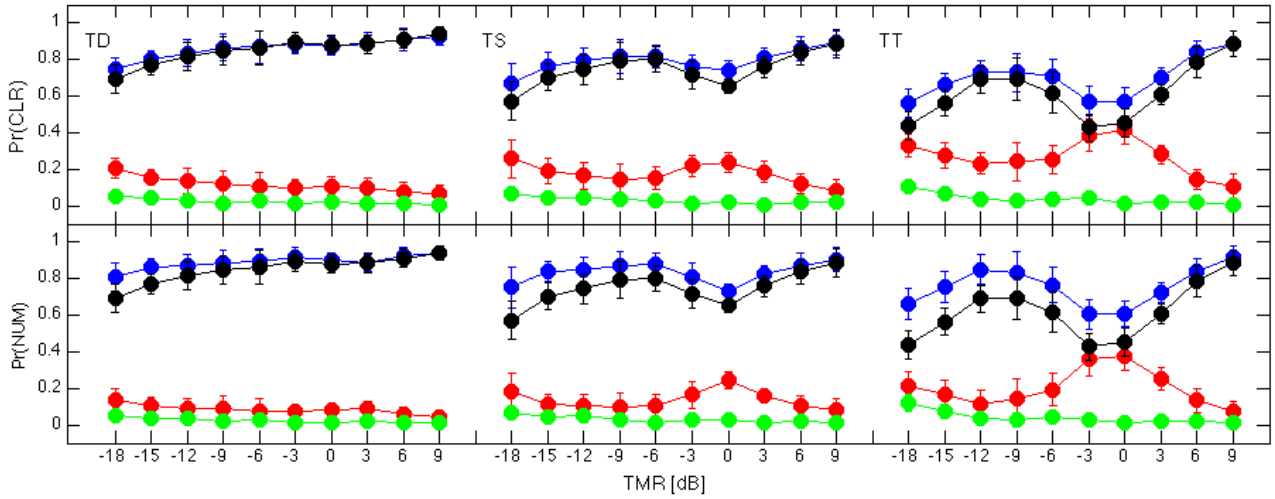


Figure 1: Results from a replication of the 2001 study by Brungart and his colleagues ([5]) are shown in the six panels. The data are broken down into types of response: target (blue), masker (red), neither (green) and target correct for color and number (black). Each panel plots these responses as a function of target-to-masker ratio (TMR) in dB. The top row shows the responses for color for three talker conditions: TD (a male talker and a female talker), TS (both talkers male or female), TT (both talkers drawn from the same talker utterances). The bottom row shows the responses for numbers. 95% confidence intervals are shown by the error bars.

confidence intervals are shown by the error bars, where the data has been pooled over the ten participants.

The replication followed the conditions and procedures of [5] in all respects except two. The TMR, which ranged from -12 to +15 dB in the original study, was shifted to a lower range of TMRs (-18 to +9 dB) in the interest of studying performance at TMRs closer to masked detection thresholds. In addition, the replication included feedback at the end of each trial following the listener's response in which the correct color-number option was highlighted on the visual display.

The trends of the replication follow those in the original study. When the target and masker genders differ, performance effectively depends only on the TMR and, even at the lowest TMRs, participants are able to respond correctly to the color and number well above chance. Comparing TS to TD, when the genders are the same, but the talkers differ, performance is suppressed slightly at the lowest TMRs, but the monotonic improvement in performance with TMR is interrupted in the neighborhood of 0 dB. When the target and masker are drawn from the same talker, performance is further suppressed at the lowest TMRs and substantial degradations in performance are observed over a broader neighborhood of 0 dB. Whereas differences exist between the color and number results, these are substantially smaller than those found in the original study.

To date, a theoretical account of the two-talker CRM results [5] 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 [6]. However, none of these concepts have been operationalized to the point of providing strong predictions of experimental outcomes. What follows is an attempt to help bridge this gap.

## 3. OVERVIEW OF THE EPIC ARCHITECTURE

The EPIC architecture is summarized in Figure 2. A simulated human consisting of several processors is on the right, and a simulated task environment (often called the device) is on the left. Each box in the diagram corresponds to a component or set of components in the software that simulates the activity of a subsystem of the simulated human. By setting aside the usual preoccupation with learning, it has been possible to develop the architecture to provide useful approximations for a wide variety of mechanisms and processes that are important in realistic task settings.

In previous versions of EPIC, the components for audition and speech were adequate for modeling some tasks involving auditory signals and speech interaction (e.g. [7] ) but were not particularly well-structured. These components were further complicated by additions made to incorporate the phonological loop model of verbal working memory [8] in which the vocal motor processor deposits encodings of spoken words (either overt or covert) into the auditory working memory, as shown by the connection in Figure 2. Further modifications were made by Mueller ([9]) to support the complex retrieval and guessing strategies required to fit detailed patterns of recall in verbal working memory.

Another set of additions handled localized sound to model human performance in the "Ballas" dual-task paradigm [10, 11]. The major addition to the architecture was to combine visual and auditory spatial representations by postulating that an object with a perceived location in space could have both visual and auditory properties. A connection was added between the auditory processor and the oculomotor processor whereby a sound onset could trigger a reflexive eye movement to the location of the sound, and a production rule could specify an eye movement to a sound's location as well; this relationship between the auditory processor and the oculomotor processor is shown in Figure 2.

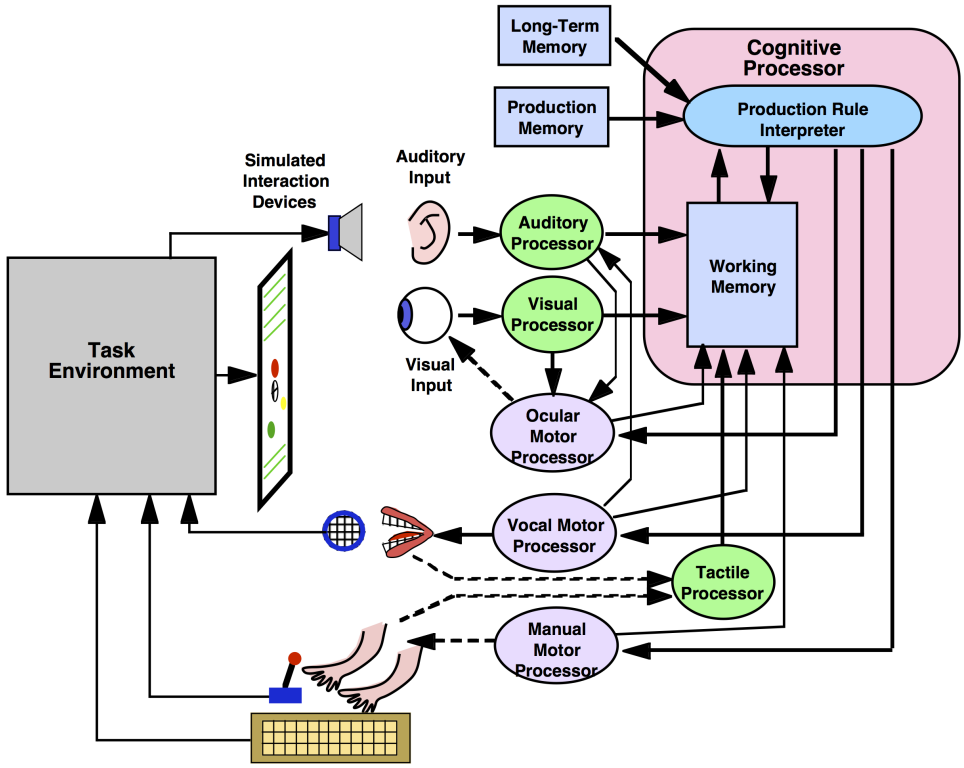


Figure 2: The overall structure of the EPIC architecture. A simulated task environment - usually a device is on the left; the simulated human, made up of a set of components, is on the right. All components of the simulated human and the device run in parallel. Note the connections between the Vocal Motor Processor, the Auditory Processor, and the Ocular Motor Processor.

## 4. MODELING THE TWO-TALKER CRM TASK

### 4.1. Auditory streams

Although the extensions above have supported models of tasks that involve hearing, they are not defined in sufficient temporal grain to handle such common (but challenging) tasks as multi-talker listening, where performance is critically dependent on events at time scales on the order of 50-100 msec [5]. EPIC already supports visual objects, which are passed from the visual processor to working memory. As the cognitive processor is refreshed, such objects may be eliminated, either due to the action of a production rule or by a natural, probabilistic decay mechanism.

A simple extension of EPIC's representation of sound, which parallels the construct of a visual object, is an auditory stream. The role of the auditory processor is to parse temporal segments of the auditory input into *streams* and pass these into the cognitive processor's working memory. Like a visual object, a stream is defined by a number of features that are extracted by the auditory processor from the auditory input. To capture the intrinsic dependence of sound on time, the vector of features is time stamped so that the representation in working memory is a strictly ordered set of feature vectors

$$S = (\underline{f}(t_1), \underline{f}(t_2), \ldots, \underline{f}(t_k)$$

Such a representation assumes a process that updates, on a frame-by-frame basis, each of the *streams* in working memory, so that the number of elements in $S$ can grow over time.

### 4.2. Production rules for an oracle auditory processor

Any given task is modeled in EPIC at the cognitive-processor level as a collection of production rules, each of which fires when the contents of working memory match that rule's conditions. Firing may produce additional content in working memory, to be matched on subsequent cycles of the processor, remove content, or initiate motor output (e.g., finger point, shift in gaze, or speech). In general, task performance depends not only on the fidelity of the contents in working memory (the output from the relevant perceptual subsystems) but on the choice of production rules as well.

In order to write production rules for multi-talker listening tasks, such as the two-talker CRM task, it is sufficient for the *auditory stream* to represent "who is speaking" and "what is said" over the temporal course of the multiply-occurring synchronized speech utterances. This admits a much simpler form of auditory stream in which "who is speaking" is represented by each stream itself and "'what is said" is represented by an ordered sequence of *words*, e.g., the feature vectors of the stream. Thus, the production rules for the two-talker CRM task operate on target $T$ and masker $M$ streams in working memory, each of which grows incrementally a word at a time.

Given an *oracle* (ideal, perfect, all-knowing) auditory processor, the following (minimal set) of production rules form an optimum strategy for solving the two-talker CRM task:

- For each stream, if $S_i(w(2))$ is "Bravo", then $target = i$. [*Target call-sign test*]
- If $target$ exists and $S_{target}(w(5))$ exists, then $color = S_{target}(w(5))$. [*Color test*]
- If $target$ exists and $S_{target}(w(6))$ exists, then $number = S_{target}w(6))$. [*Number test*]
- If $color$ and $number$ exist, then press the appropriate (color,number) coordinate on the screen. [*Response*]

In EPIC, the cognitive processor executes each rule in the strategy as each word is added to the target and masker streams. During the first cycle, all tests fail. During the second cycle, the Call-sign test fires for one of the streams and a new proposition in working memory is created ($stream$). During the fifth and sixth cycles, the Color and Number tests fire, respectively, for the first time, and create two additional propositions ($color$, $number$). Finally, during the seventh cycle, the Response test fires.

In the actual programming and running of EPIC, the set of production rules above would require a halting rule based on knowledge that a response has been generated. For the sake of efficiency, auxiliary production rules can be included that remove rules that have already successfully fired from working memory. Finally, the Response rule requires a more detailed set of production rules that engage the working memory for visual objects (the matrix of colored blocks containing numbers) along with the manual motor and ocular-motor processing modules for executing the button press.

### 4.3. Production rules for a noisy auditory processor

Given all that is known about auditory processing, an *oracle* auditory processor is hardly an appropriate model of the human in this task. Nevertheless, modeling the task under the assumption that the sensory input is error-free leads to an optimum strategy that can be further refined as more realistic assumptions about auditory processing are introduced. As far as the task strategy is concerned, the exercise above identifies three potential types of error: errors due to guessing that occur when the target color (or number) words are missing, errors due to guessing which similarly occur when both call signs are missing (as could arise from energetic masking between the masker and target), or errors due to assigning masker words to the target stream and vice versa (as could arise from informational masking between the target and masker sources). In the case of missing words, additional production rules are required to completely specify the task strategy. In addition, the stream representation must be modified to allow for a feature vector $w(k)$ to be empty.

*Missing target color (or number) words (Type 1 Error).* If either the color or number word of the target stream is missing, additional production rules are required to avoid "no response" at the end of the trial. In the CRM task, listeners are told to avoid using masker stream color or number, as they are known to be wrong. Therefore, additional production rules implement an optimum guessing strategy that first identifies the masker-stream color (or number) and removes it from the set of possible colors (or numbers) before guessing.

*Missing call signs (Type 2 Error).* The *Target call-sign test* fires if one of the streams bears the call-sign word "Bravo". That stream is subsequently labeled the *target* stream. In the absence of the target call sign, however, it is still possible to correctly label the target stream. An additional set of production rules utilizes the symmetry of the two-talker task to infer the identity of the target stream: the presence of a masker call sign in one stream can be used to correctly label the other stream as that of the target. If both call signs are missing, the strategy labels the stream with the more information (color and target words) as the target, or labels one stream at random, should the information be the same.

*Assignment error (Type 3 Error).* The third type of error reflects imperfect assignment of words to streams by the auditory processor. The cognitive processor is effectively blind to such errors. It can only intervene using additional production rules if there is insufficient information in working memory to solve the task. Switching what word goes with which stream as the streams are updated in working memory is unobservable.

### 4.4. Model of a noisy auditory processor

Based on the analysis above, it is sufficient to model the auditory processor at the level of word-content detection and word-assignment error.

*Content detection.* For each of the talker conditions (TD, TS and TT), content detection is assumed to depend on the SNR through a cumulative distribution function of the standard normal distribution

$$p_{det}(\mathrm{SNR}) = \Phi(\frac{\mathrm{SNR} - \mu}{\sigma})$$

where the mean ($\mu$) and variance ($\sigma^2$), in general, depend on the word (call sign, color, number) as well as talker condition, and

$$\mathrm{SNR} = \mathrm{TMR}$$

when computing the probability of detecting the target word in the presence of the masker word, and

$$\mathrm{SNR} = -\mathrm{TMR}$$

when computing the probability of detecting the masker word in the presence of the target word.

*Assignment error: Switch model vs. Stream tracking model.* In the following, two models of assignment error are considered. The **Switch model** treats the assignment error as a conditional Bernoulli random variable where $p_A$ is the probability of assigning the incoming pair of words (from the target and masker utterances) to the same streams as the previous incoming pair of words. $p_A$ is assumed to increase monotonically with $|TMR|$ and, for fixed TMR, $p_A$ is expected to be smallest for the TT condition and largest for the TD condition. In fitting $p_A$ to the data, the conditional nature of the model allows the product of the word-by-word probabilities ($w(k)$ to $w(k+1)$) to be represented as a single conditional probability from $w(k)$ to $w(k+n)$ . Accordingly, without loss of generality, the model assumes that the incoming pair of call signs are assigned to the correct auditory streams in working memory and fits $p_A(color|call\ sign)$ and $p_A(number|color)$.

The **Stream tracking model** treats assignment error as a consequence of errors in a simple tracking algorithm. Accordingly, two stimulus variables (pitch in semitones and intensity in dB) are used as surrogates for talker identification. The tracker is designed to estimate the mean pitch and intensity of each stream by updating its current estimates with the observed pitches and intensities of the incoming data. Which data is associated with which estimate determines how the auditory streams in working memory are updated. That is, for the first word ("Ready"), the stream tracker creates two auditory streams in working memory, loads each with "Ready", and associates each stream with the appropriate (pitch,intensity) values for that source's "Ready". For

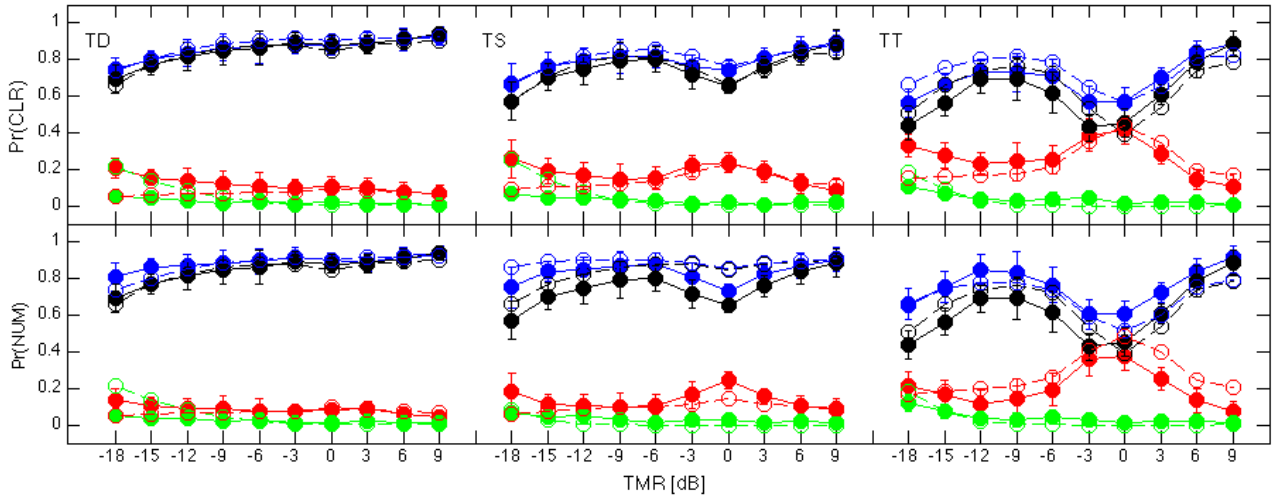


Figure 3: Results of fitting the switch model to the replication data of Figure 1. The open-symbol/dashed lines are the predicted responses; the solid-symbol/solid lines are the observed data.

each subsequent word, the incoming (pitch,intensity) values of the target and masker talkers are compared with the current mean (pitch,intensity) values for each stream. The pairing that minimizes the weighted difference between the incoming and current mean values determines which data are associated with which stream. Following this association, the mean estimates for both streams are updated and the observed word contents are appended to the appropriate auditory streams in working memory. To fit the data, the relative weighting of pitch differences to intensity differences is adjusted as a function of stimulus condition.

## 5. PERFORMANCE ANALYSIS

### 5.1. Switch model

The optimal production rules of 4.1 and 4.2 were programmed in EPIC and the auditory processor was black boxed to create auditory streams in working memory according to the Gaussian models of content detection and the conditional probabilistic switch model of assignment error. Whereas EPIC provides considerable utility in modeling the cycle-by-cycle behavior of a set of production rules, it is not well suited for optimizing the parameters of its black boxed components to best fit a given set of data. Accordingly, the production rules were reduced to a decision tree and Matlab was used to optimize the parameters of the content detection functions and the conditional probabilities of the switch model.

The results are shown by the open-symbol/dashed lines in Figure 3 and are based on best fits obtained from an interior-point constrained optimization procedure. In general, the model accurately captures the general behavior of the data. Correct target color and number detection increases monotonically for $\mathrm{TMR} \geq 0$ dB. The probability of reporting "Neither" decreases monotonically with TMR. For intermediate TMRs (in the neighborhood of 0 dB), non-monotonicities in Target and Masker responses are observed. These are greatest for the TT condition and smallest for the TD condition.

Figure 4 shows the psychometric functions for content detection (upper panels) and the conditional probabilities of staying in the currently assigned stream (lower panels) for color (left panels) and number (right panels). In general, it appears that performance in the two-talker CRM task is dominated by informational masking, as modeled by stream assignment error. For any given TMR, listeners are most likely to confuse masker and target words in the TT condition, i.e., when the utterances are drawn from the same talker, and least likely to do so in the TD condition, i.e., when the utterances are drawn from a male and a female talker. There is an orderly progression about the point of symmetry at 0 dB TMR.

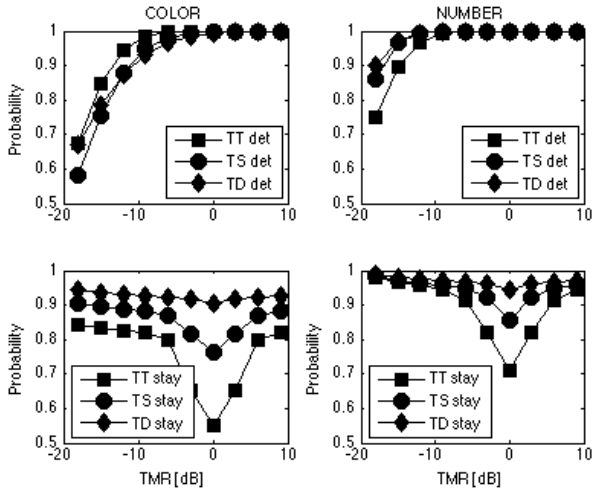


Figure 4: Best-fitting psychometric functions (top panels) and conditional probabilities of staying in the current stream (bottom panels) are shown for the Switch model. Separate functions are fit for each talker condition (TD, TS, and TT) and for each word (color, number).

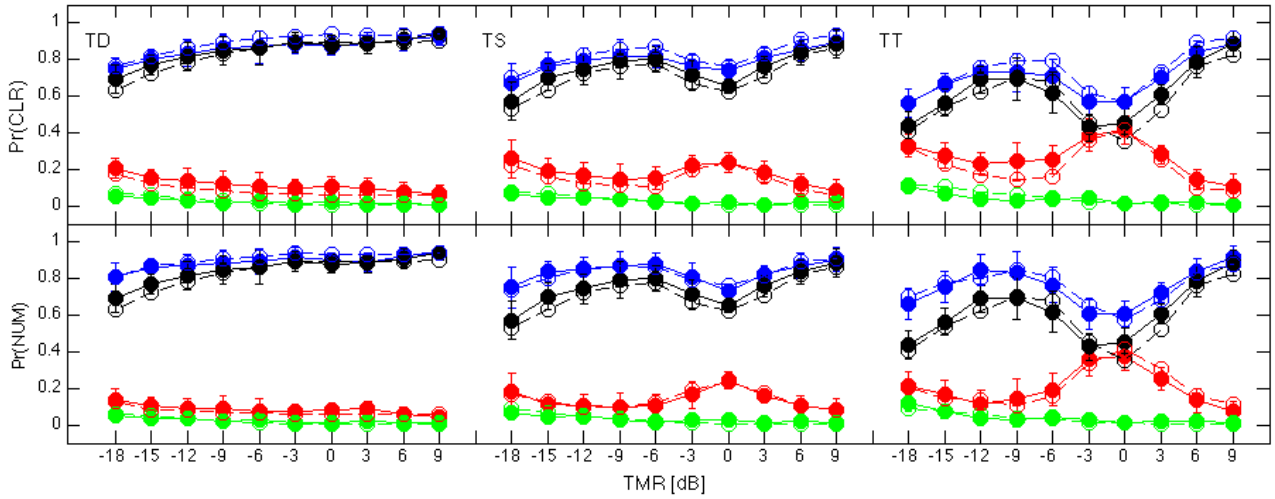


Figure 5: Results of fitting the stream tracking model to the replication data of Figure 1. The open-symbol/dashed lines are the predicted responses; the solid-symbol/solid lines are the observed data.

When utterance level can't serve as a cue, listeners are the most likely to confuse masker and target words for any given talker condition. As utterance level becomes a more reliable cue, the stream assignment process also becomes more reliable and performance improves. Finally, even though the number word immediately follows the color word in each utterance, there remains a non-trivial probability that the words will be switched during their assignment.

If assignment error were the only source of error, target response should be high at both low and high TMRs. The content detection psychometric functions indicate why there is a drop off in performance at the negative TMRs. There doesn't appear to be a significant difference among the three psychometric functions for color (upper right panel). For all three talker conditions, detection drops rapidly below -12 dB TMR. Comparing the psychometric functions for numbers with those for color, it appears that numbers are, in general, more easily detected than colors, but this improvement is much less in the case of numbers drawn from the same talker.

Despite fitting the general trends of the data, the model does break down in several important ways. First, it under-predicts target response and over-predicts masker response at positive TMRs and trends in the opposite direction at the most negative TMRs. Second, the model predicts far more "neither" responses at the lowest TMRs than are observed in the data for the TT and TS conditions. Both of these phenomena can be traced to single attribute of the model: the symmetry of assignment error around 0 dB TMR. Adjusting the conditional probabilities to better fit the target response at positive TMRs leads to greater target response rates at negative TMRs, which is clearly not in the right direction. At the lowest TMRs where confusion is the least likely, loss of content does play a role in suppressing the target responses. However, as the production rules have clearly identified the masker stream, the masker color (or number) is eliminated from the guessing options and the loss of target content is offset by an increase in responding "neither". This is a fatal flaw of the Switch model and the optimal strategy. There are just too few masker intrusions at the lowest TMRs, and this trend is the likely cause of the poorer fits at $\pm 9$ and $-12$ dB TMR.

### 5.2. Stream tracking model with suboptimal production rules

In light of the failure, the production rules can be modified to include a "use-what-you-heard" clause. Under this case, content from the masker stream is used should the integrity of the target stream be sufficiently degraded, as measured by missing call sign, color, or content words. A version of this suboptimal strategy was implemented with the stream tracking model of the auditory processor. Paraphrasing the production rules, the strategy is as follows:

- If callsign content is present, label its stream as Target or Masker. If not, infer Target/Masker status from the other stream.
- If Target stream is known, and Target color (or Number) content is available, use it.
- If Target stream has only been inferred, and Target color (number) content is not available, but Masker content is, use it instead.
- Otherwise, use a content pair from the same stream if available.
- Otherwise, use separate color, number content if available.
- Otherwise, do a pure guess of color, number content.

Unlike the switch model, the stream tracking model generates assignment errors by virtue of imperfect tracking of the means of the two incoming talkers. Since pitch and loudness are proposed as surrogates for talker identification, it is necessary to weight the relative importance of the differences in pitch and loudness. This was accomplished by adjusting a mixture parameter, $\lambda$, across all talker conditions. In addition, the literature suggests that the effect of differences in pitch between two talkers in such listening tasks saturates beyond a critical range [12]. Accordingly, pitch differ-

ences greater than 4 semitones were capped at 4 semitones. Initial experiments with these modifications suggested that the model was substantially better at the task than a human observer. Therefore, to degrade the model's performance, a weighted uniformly distributed random variable was added to the (pitch,intensity) statistic and a "Bernoulli flip" was introduced at the final stage of stream assignment.

The fits for the stream tracking model are shown by the open-symbol/dashed curves in Figure 5. Comparing the fits with the data, there appears to be excellent agreement over the entire range of TMR. By invoking a suboptimal strategy ("use-what-you-heard"), the model is able to achieve very good fits at positive TMRs and also accurately follow the proportionate rise in masker responses at the lowest TMRs. In general, the model accounts for 99% of the variance in the target, masker, and neither responses. It is slightly weaker at handling the joint occurrence of target responses (black lines) where it accounts for 96% of the variance.

The three panels of Figure 6 show the best fitting content detection psychometric functions for the call sign, color and number words. These psychometric functions were fit by constraining the standard deviation of all functions and conditions to be the same and allowing the means to vary freely. Similar to the observations made for the Switch model, content detection improves with each succeeding word in the stream. In addition, there is a strict ordering of performance within each word from TT (poorest detection) to TD (best detection). The optimal mixture parameter was determined to be 0.75, favoring pitch over loudness. Although allowed to freely vary, a single value for the additive random term (0.35) was required. Similarly, although allowed to freely vary as a function of talker condition, the perturbation probability to emulate a noisy stream assignment was found to be 0.04 for TD and 0.05 for TS and TT. Thus, the model requires 14 parameters[2] to fit the complete two-talker CRM replication data.

Although it accounts for the data, why listeners apparently adopt the suboptimal strategy of "use what you heard" in the face of instructions that declare the contrary remains an interesting, open question. From the standpoint of task analysis, there is a possible observation state that does not fall under the two-talker strategy. On some trials, under very low TMR conditions, it is possible that a target stream may simply not be heard. In that case, should the listener follow the task strategy of the experiment (I'm supposed to be hearing two talkers) or do they opt for some other strategy (I'm supposed to report the color and number I heard)? The dominance of masker intrusions in the replication data (which is also very clear in the original data) suggests the listeners are making responses well outside the instructions they were told to follow.

## 6. DISCUSSION AND CONCLUSION

Among the standard cognitive architectures, EPIC supports detailed modeling of those sensory modules central to the performance of a given task. It has been successfully applied to the modeling of human performance in visual tasks and has also been used to explore combined visual and auditory tasks, typical of multimodal watch stations. Over the years, EPIC has developed a sophisticated model of the human visual system. The present work

[2]Ten parameters are required for the content detection psychometric functions (3 means X 3 talker conditions + 1 variance). Four parameters are required for the stream tracker (mixing parameter (1), additive random term (1), perturbation probability (2)).

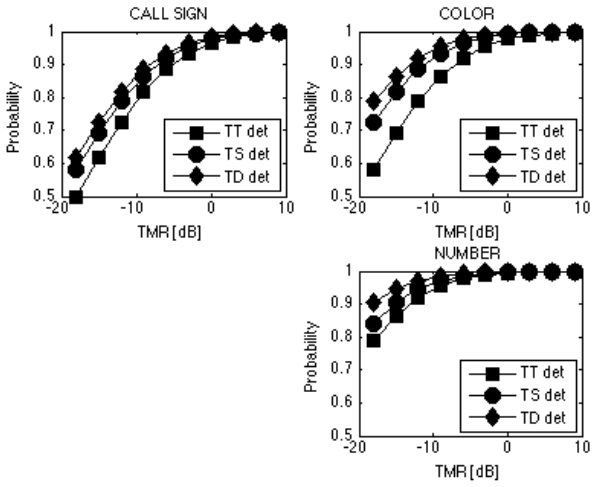


Figure 6: Best-fitting psychometric functions are shown for call sign, color, and number content detection. The mean of each function was allowed to vary freely in the optimization, but a single variance term was fit to all detection functions.

focuses on the need for a similarly detailed model of the human auditory system.

The construct of an auditory stream, a notion that has a rich history of discussion in the psychoacoustic literature, is introduced formally as an object in EPIC's working memory. To support the production rules necessary to execute a two-talker CRM listening task, multiple auditory streams are proposed, each of which comprises an ordered list of words. To maintain such streams, it is necessary for the auditory processor to update each list by appending new words to the old. Production rules are cyclically applied in a manner that operationalizes what would be called the receiver (ideal or otherwise) in classical signal detection theory, but for far more complicated types of listening tasks.

Careful enumeration of the rules necessary to solve the task is one of the fundamental requirements in using EPIC to model complex human behavior. Such enumeration not only guarantees that a task can be successfully completed by the simulation, it also specifies what must be known about the sensory front-end. In this sense, EPIC provides a means to work backwards, from the task specification to the sensory input, to determine what must be known minimally about the sensory black box. In the present example, modeling the complex processes starting with the front-end and moving forward to predict the outcome of a two-talker listening task requires a detailed understanding of speech processing, energetic masking, and informational masking. A survey of the literature suggests that our understanding of these has yet to provide adequate predictive power. In the application of EPIC to the present problem, two questions arise which the auditory black box must model: what is heard and who is speaking. Working out the consequences of these questions leads to the need to address word content detection and word assignment. By fixing the production rules, the present work demonstrates that the two-talker CRM data can be used to extract the underlying psychometric functions for content detection and streaming errors which operationally determine the success or failure of the listener's response. What pro-

cesses are involved in the system that produces such psychometric functions is an interesting question, but an answer is not required in order to model the two-talker CRM task.

The Switch model was entertained as a first step towards a stimulus-driven tracking model of auditory stream formation and support. Systems that engage in multi-target tracking are apt to commit assignment error in how data is fused to each target track. From working with the simple switch model, the need for suboptimal production rules was revealed. Using these, a corpus-driven stream tracker was shown to account for the human data. This is not to say that the true auditory stream tracker computes average pitch and level on entire words! However, the finding does provide a guide to anyone interested in more detailed modeling of such auditory processing and stream assignment as any proposed model must generate the same type of results as produced by this more elementary stream tracker. It is important to understand the limits of EPIC modeling: the present work rests on the success of some as of yet poorly understood process by which auditory input is parsed into multiple pitches, multiple intensities, and multiple words. That said, the success of the EPIC approach is that predictions can be made about performance on a different corpus without waiting for a complete understanding of auditory perception in multi-source environments.

## 7. ACKNOWLEDGMENT

Funding was provided by grants from the Office of Naval Research (N00014-10-1-0152, N00014-13-1-0358) and the U. S. Air Force 711 HW Chief Scientist Seedling program. The authors thank the reviewers for their excellent suggestions, and apologize in advance for not having the space to discuss several of the more interesting, broader issues raised.

## 8. REFERENCES

[1] R. Bolia, W. Nelson, M. Ericson, and B. Simpson, "A speech corpus for multitalker communications research," *Journal of the Acoustical Society of America*, vol. 107, pp. 1065–1066, 2000.

[2] D. E. Meyer and D. E. Kieras, "A computational theory of executive cognitive processes and multiple-task performance: Part 1. basic mechanisms," *Psychological Review*, vol. 104, pp. 3–65, 1997.

[3] ——, "A computational theory of executive control processes and human multiple-task performance: Part 2. accounts of psychological refractory-period phenomena," *Psychological Review*, vol. 104, pp. 749–791, 1997.

[4] ——, "Precis to a practical unified theory of cognition and action: Some lessons from computational modeling of human multiple-task performance," in *Attention and Performance XVII*, D. Gopher and A. Koriat, Eds. Cambridge, MA: M.I.T. Press, 1999, pp. 15–88.

[5] D. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 109, pp. 1101–1109, 2001.

[6] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sounds*. Bradford Press, MIT Press, 1990.

[7] D. E. Kieras, S. D. Wood, and D. E. Meyer, "Predictive engineering models based on the EPIC architecture for a multimodal high-performance human-computer interaction task," in *ACM Transactions on Computer-Human Interaction*, 1997, pp. 230–275.

[8] D. E. Kieras, D. E. Meyer, S. Mueller, and T. Seymour, "Insights into working memory from the perspective of the epic architecture for modeling skilled perceptual-motor and cognitive human performance," in *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, A. Miyake and P. Shah, Eds. New York: Cambridge University Press, 1999.

[9] S. Mueller, "The roles of cognitive architecture and recall strategies in performance of the immediate serial recall task," Ph.D. dissertation, The University of Michigan, Ann Arbor, MI, 2002.

[10] D. E. Kieras and D. E. Meyer, "Predicting performance in dual-task tracking and decision making with EPIC computational models," in *Proceedings of the First International Symposium on Command and Control Research and Technology*, National Defense University, Washington, DC, 1995, pp. 314–325.

[11] J. A. Ballas, D. E. Kieras, and D. E. Meyer, "Computational modeling of multimodal i/o in simulated cockpits," in *Proceedings of the Third International Conference on Auditory Displays*, S. P. Frysinger and G. Kramer, Eds., Xerox Palo Alto Research Center, Palo Alto, CA, November 1996, pp. 135–136.

[12] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2913–2922, 2003.