Developing Models for Multi-Talker Listening Tasks using the EPIC Architecture: Wrong Turns and Lessons Learned

David E. Kieras and Gregory H. Wakefield

University of Michigan

Report No. TR-EPIC-17

July 30, 2014

This work was supported by the Office of Naval Research, Cognitive Science Program, under grant numbers N00014-10-1-0152 and N00014-13-1-0358, and the U. S. Air Force 711 HW Chief Scientist Seedling program.

Introduction

The Problem

Both EPIC (Meyer & Kieras, 1997a,b; Kieras & Meyer, 1997; Kieras, in press) and the closely related GLEAN architecture for simulating human performance (Kieras, Wood, Abotel, & Hornof, 1995; Kieras & Knudsen, 2006) have basic facilities for presenting simulated audio signals and speech input to the simulated human, and generating simulated speech output from the human which can be sent to the simulated task environment or other simulated humans. Other cognitive architectures such as ACT-R (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004) and Soar (Laird, 2012) have no better facilities for modeling auditory effects. In various models developed with EPIC and GLEAN over many years, basic auditory architecture facilities have demonstrated their value, such as simulating telephone operator tasks in Kieras, Wood, and Meyer (1997), and CIC team members interacting via speech in Santoro, Kieras and Pharmer (2004). However, the very simple form of these facilities does not address many critical issues that the future design of military human-computer systems will have to deal with. The basic problem is that the important phenomena specific to audition, such as spatial localization, masking, and segregation of auditory streams of both speech and non-speech sounds are not represented in extant cognitive architectures. These phenomena have to be addressed because current and future military task settings, such as CIC watch-standing, involve multiple simultaneous speech inputs. Adequately representing auditory phenomena in predictive models of human performance requires representing audition and speech far more comprehensively and accurately than it has been so far.

Purpose of this Report

The purpose of this report is to supplement our formal publications which necessarily do not cover some of the failures and lessons learned from a research effort. Thus this report provides a

historical description of our efforts to model multi-talker listening tasks by exploring extensions to the EPIC auditory components, constructing and testing many models, and working with our AFRL collaborators, Eric Thompson, Nandini Iyer, and Brian Simpson, who collected additional experimental data on the 2-channel listening task that greatly informed the modeling work.

First will be presented the 2-channel listening task that provided the initial test bed for this work, followed by the narrative of the modeling effort.

A note on terminology: this reports follows the common practice in the psychoacoustics literature of referring to human speakers as *talkers* apparently to avoid confusion with the acoustic transducers known as "speakers."

A Basic 2-channel Listening Task and Data Set - Brungart (2001)

We surveyed key work by Brungart and co-workers on basic multiple-channel speech processing (e.g. Brungart & Simpson, 2005). This led to a decision that our first model should be of a basic 2-channel speech processing task, whereupon we realized that most of the studies in the literature omitted some key empirical facts concerning the "fate" of the "unattended" material.

The Paradigm

However, one of Brungart's earliest studies (Brungart, 2001) provided what appeared to be an ideal first study to model. This study compared performance in a two-talker task using the well-known CRM corpus, which consists of sentences of the form "Ready <call sign> go to <color> <digit> now" recorded by eight different talkers. Two such sentences are presented simultaneously to the subject; the call signs, colors, and digits used in the message are selected to be different. The task is to click on the colored digit on a display that corresponds to the message with a designated *target* call sign (always "baron" in this study). The two messages differed in relative loudness (SNR) and three levels of talker speech similarity: different genders, different talkers of the same gender, and the same talker. The message with the designated call sign is the *Target* message; the other is the *Masker* message. Every trial contains a Target message and a Masker message.

The Results

Unlike almost all studies of multiple talker processing, this study included some detailed data about the incorrect responses; the results show not only the proportion of correct responses as a function of talker similarity and SNR, but also a separate breakout for the color and digits, and whether they were correct (from the target message), from the masker message, or neither one. These results are shown in Figure 1, which for each talker similarity condition in the upper panel shows the proportion of responses designating a color that matches the Target (blue curve), Masker (red curve), or Neither (green curve) message , and correspondingly for the digit responses in the lower panel. Also shown in each panel is the *Both-Correct* responses (black curve) in which both color and digit are correct; this curve is identical in both panels. Most of the graphs in this report are coded the same way, so some study of this graph will save time later.

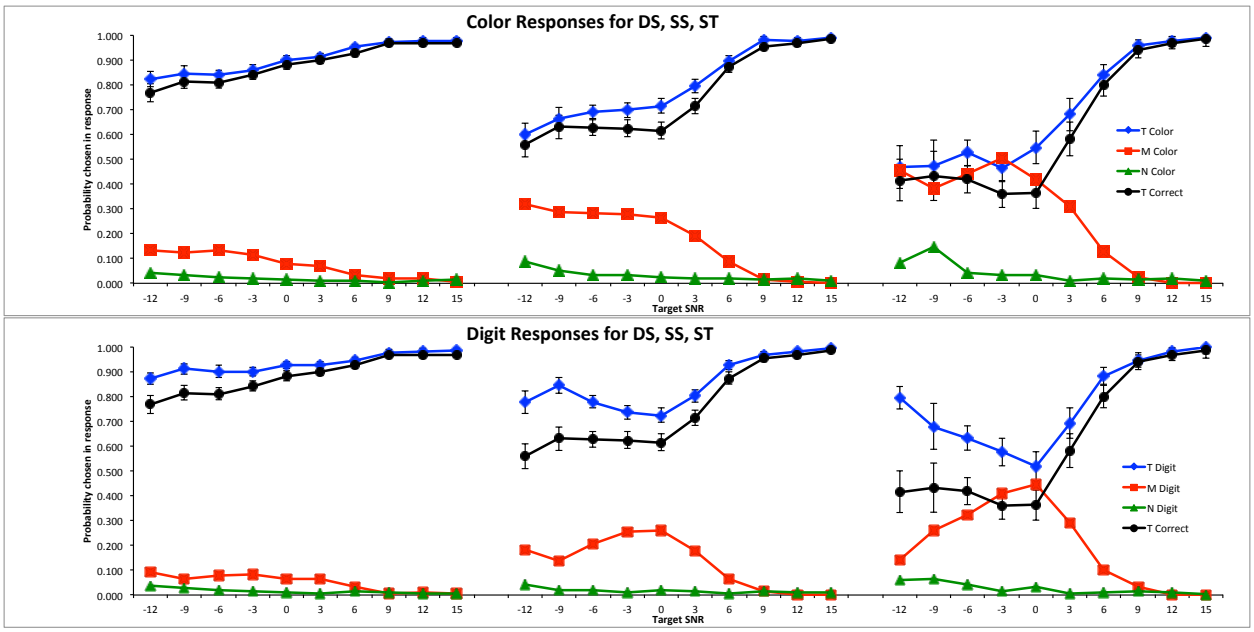

Figure1. *The proportions for each response choice in the Brungart (2001) data as a function of signal-to-noise ratio (SNR) of the target message relative to the masker message. The vertical axis is probability of response; the horizontal axis is SNR in dB. Color responses are in the upper panel; digit responses are in the lower panel. The three talker conditions are shown left-to-right in the three subpanels: Different Sex (DS), Same Sex (SS), and Same Talker (ST). Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker. The confidence intervals are those reported in Brungart (2001) which were only for the target and both-correct responses.*

Key Phenomena

The basic effects are as follows: with increasing SNR, the Both-Correct and Target color and digit responses are chosen more often, and Masker and Neither content are chosen less often. The overall performance when the messages are delivered by different-sex talkers (DS) is greater than for same-sex (SS) talkers, which is turn is greater than when the two messages are from the same talker (ST). This means that the two message streams are easier to segregate when the talkers differ, especially by sex.

A key empirical fact is that the incorrect responses were almost always from the masker message, which places a basic constraint on the cognitive-architectural processes in a model, in that it implies that Masker message content was being perceived and remembered, and chosen as a response, rather than being simply filtered out. Brungart's (2001) interpretation was that these results, combined with the fact that Neither responses were less likely than Masker responses, mean that *informational* masking rather than *energetic* masking was primarily at work. That is, the results support the idea that the interference between the messages was less about obscuring individual words due to energetic masking, and more about obscuring which stream they come from, a form of informational masking. Words are not simply swamped by a louder word from the other stream and become undetectable. Rather, they can still be perceived and recognized; the

problem is that their relationship to their source stream can be lost or jumbled, leading to an incorrect response.

Another key empirical fact is that the color words (upper panel in Figure1) were recognized rather differently than the digit words (lower panel); note the differently-patterned effect of SNR. This is a result that neither Brungart (2001) nor subsequent researchers have attempted to explain. Most striking is the fact that in the ST condition, and to some extent in the SS condition, digits are recognized better at lower SNRs compared to medium SNRs. For colors, the corresponding effect is that the curves flatten out below 0 dB SNR; the color words are disproportionately detectable at low compared to middle SNRs, compared to digits, which are detected better at low SNRs.

Introduction to the Models

General Approach

The focus of the work in this report was to account for the Brungart (2001) 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 considerable more theoretical precision; most importantly, the strategy that the subject follows to perform the task is directly represented, and this turns out to be vastly important to accounting for the specific effects in this data.

At the top level, the significance of the models presented in this report 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 sound 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, 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 for a set of information-processing stages that has a cognitive-strategic middle which controls a back-end motor system, resulting in a model that performed the entire task end-to-end.

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, in press), 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

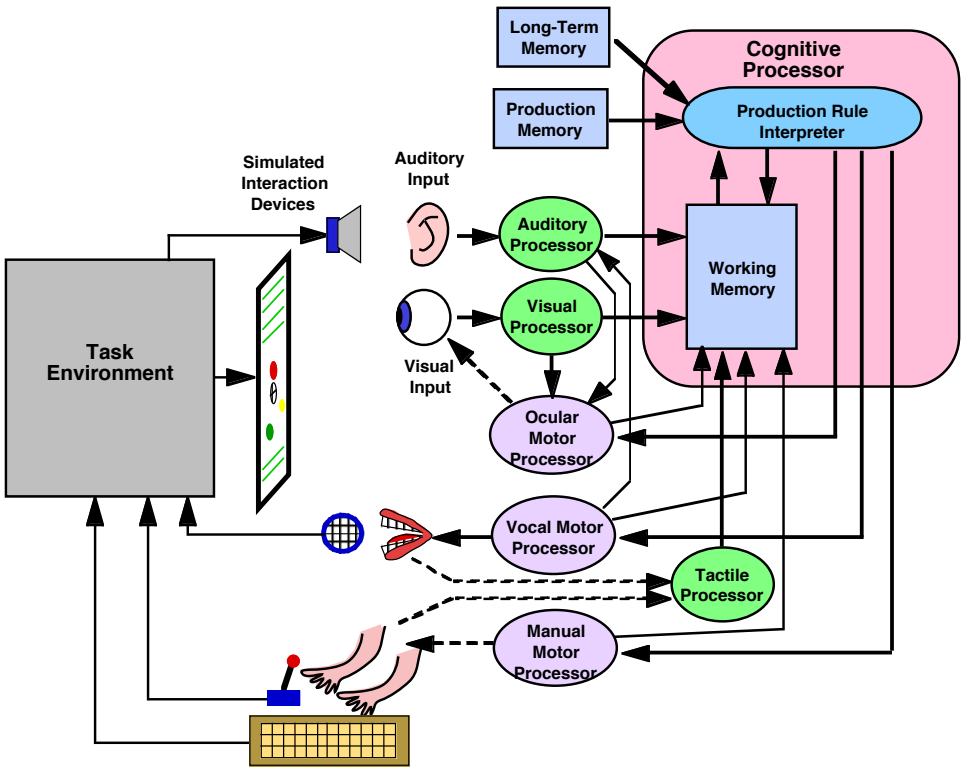

***Figure 2**. The EPIC architecture in simplified form. The simulated environment, or device, is on the left; the simulated human on the right.*

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 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: **E**xecutive **P**rocesses **I**nteract with and **C**ontrol the rest of the system by monitoring their states and activity.

See Kieras (in press) 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 report describes the work to develop the auditory system in more detail to support modeling of multichannel speech processing.

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 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.

Common Features of EPIC Models for Multi-talker Tasks

The basic problem in constructing a computational cognitive model is choosing how much detail about which components of the human cognitive system to represent. Because current architectures represent only a tiny fraction of the human system, some parts of the system must be "black-boxed" – represented with components specified only in terms of the input-output behavior. This decision allows the modeler to focus on aspects of human performance that have not yet been studied without attempting to solve very difficult problems that other researchers are fully engaged in. A good choice of black-boxing decisions can thus allow good progress to be made. It must be kept in mind, of course, that these decisions could be wrong or misleading.

In addition to the black-box decisions, the models developed in this work share a basic approach and some common features. The following list contains the black-box decisions, approaches, and features shared by all of the models to be presented; Figures 3 and 4 accompany this description.

1. *Input is pre-parsed.* The input speech message is already parsed into word units: the beginning and end time of each word in the message is known and specified in the input to the simulated auditory system. The perceptual counterpart of these word units are termed *word objects* in the architecture. We made this decision to provide a simpler starting point than beginning with phonemes that are even smaller units. Because the syntactical structure of the messages is known and fixed, the model strategy can simply "count" word objects to determine which word objects correspond to the call signs, colors, and digits, and which are

the filler words.

2. *Attributes are detected.* Each word object has a set of attributes that can be recognized and supplied to the rest of the architecture. The most important such attribute is the *content* or semantic meaning of the word (e.g. “ready” or “green”). If the conditions are right, the perceptual process can recognize the content of the word and supply it to the rest of the architecture. While the recognition process is completely black-boxed, we assume that the probability that the content will be recognized can be described by a psychophysical detection function whose shape and parameters will be chosen to fit the data. Additional attributes used in various models are stream-related attributes such as the mean loudness and pitch of the word. Determining the attributes required to distinguish the streams in this two-channel listening task was a goal of the modeling work.

3. *Basic stream concept.* The models assume that in addition to word objects, the perceptual system perceives *stream objects* that represent the perceived source of sounds, such as the individual talkers in the multi-channel listening task. The process of how streams are identified is black-boxed, so the models assume that the number of streams corresponds to the number of simultaneous messages. The models propose ways in which the perceptual

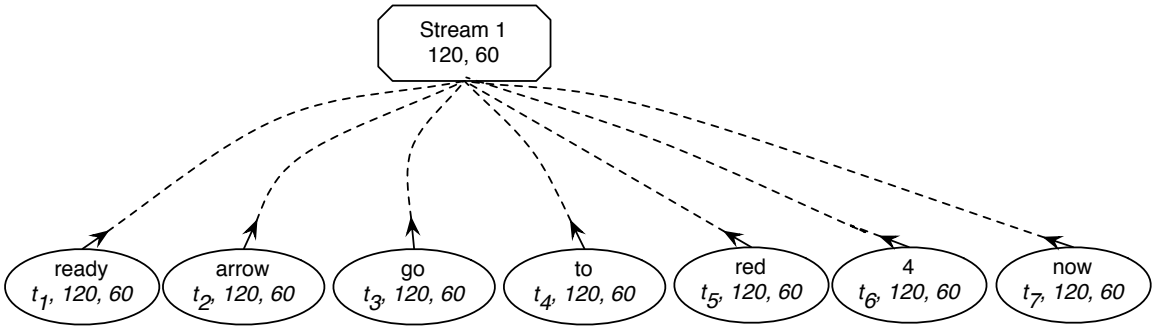

Figure 3*. Basic representational concept, shown for the CRM message “ready arrow go to red 4 now”. A stream object is at the top of the figure, and a set of word objects appear at the bottom in time-sequence order. Each word object has attributes of content, a time stamp, mean pitch (in Hz), mean loudness (in dB), and stream assignment shown by the arrows. In this example, a stream-tracking perceptual process has assigned each word object to the same stream object Stream1, which has attributes of mean pitch and loudness of the words assigned to it.*

process assigns each word object to one of the stream objects. Figure 3 illustrates the basic representational concept for the most important approach in our models.

4. *Masking model.* When two messages are presented simultaneously, these models assume that the words are aligned in time, and that masking interference effects happen on a word-by-word basis. That is, the two call sign words can interfere with each other, such that the content of only one (or neither) is detected, but the outcome does not affect the probability that the next pair of words in the messages will suffer interference. The masking mechanism itself is represented symmetrically in the psychophysical detection function - the probability of the content of *word1* being recognized is a function of the signal-to-noise (SNR) ratio of that *word1* compared to the other *word2*. In turn, the probability of *word2*'s content being recognized is simply the same function at the SNR of *word2* compared to *word1*. Figure 4

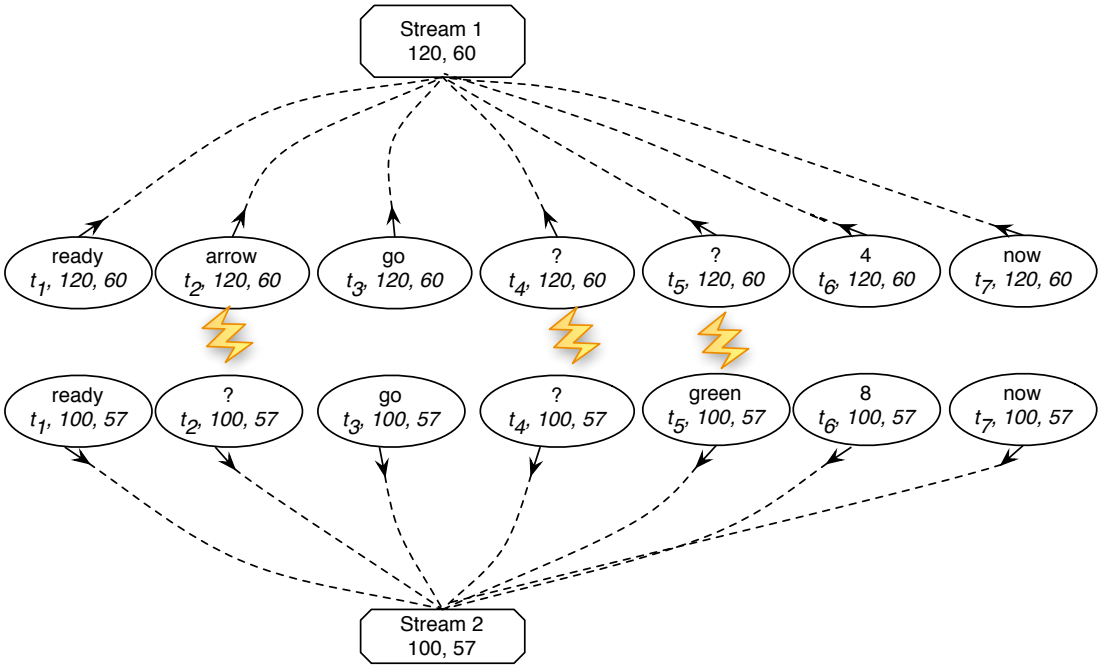

Figure 4*. Conceptual example of effects of masking at the individual word level, extending the example in Figure 3. Two CRM messages have been presented, and each word object was correctly assigned to the correct stream. However, masking between corresponding words in the two messages has resulted in the content being unavailable for some of the words. However, the response strategy can infer the correct response in this case.*

illustrates this independent masking of content between corresponding pairs of words as the "lightning bolts", resulting in failures to detect the content of one or both of the words, represented in the diagram as a "?" for the content attribute of the word object.

5. *Basic task strategy.* The basic strategy is to identify the target stream by noting which stream the target call sign content is associated with, and then choose a color content and a digit content from word objects that are associated with that target stream, and use these to produce the response. If some of the needed attributes are undetected due to masking, then other strategy options come into play.

6. *Inference of missing information.* In the two-talker case, there are only two possible streams, so it is possible to make inferences to compensate for missing information. For example, if one has not heard the content of the call sign for the target stream, but has for the masker stream, then one can infer that the other stream must be the target stream. As long as at least one of the call sign contents is recognized, and one of the color or digit contents is detected, at least part of a correct response can be produced, or part of a known-incorrect response avoided, most of the time. Furthermore, if no streams have been identified as target or masker, a choice of color and digit from the same stream is a better guess than a color and digit from different streams. In the example shown in Figure 4, some of the word content is undetected due to masking. Stream1 has been assigned the word objects whose contents are "ready arrow go ? ? 4 now", and Stream2 has "ready ? go ? green 8". Despite the missing

information, the strategy can infer that Stream2 must be the target stream because Stream1 has the masker call sign content, and this means that the correct response is the “green 8” assigned to Stream 2.

7. *Using and avoiding masker content in the response.* Because the Brungart (2001) paradigm is a forced-choice procedure, an important strategy issue is how to choose a response color or digit if the detected and inferred information does not completely specify a correct response. One option, called *avoid maskers*, would be to avoid using in the response content known to be from the masker stream, making a pure guess from the non-masker alternatives instead. Another option, called *use maskers*, is to prefer using masker content in a response if target content is not available, making a pure guess only if neither target nor masker content has been detected. This "use what you heard" strategy is not necessarily irrational, as will be discussed more in the context of the specific models. In the most important model in this report, these two options on using/avoiding maskers can be implemented by simply enabling or disabling two rules for color choice, and two for digit choice, in otherwise identical strategies. Furthermore, strategies might use different combinations of these (and other) options depending on the situation.

8. *Extremely late selection model of attention.* All of the detected attributes of each word object are supplied to the auditory perceptual store where the production rules in the cognitive processor can examine them. Thus rather than there being some form of selective-attention "filter" or "bottleneck" that governs what is supplied to it, cognition has available all of the content that makes it through the auditory perceptual process, and selects a response based on that information. In effect, "selective attention" is represented only in the sense that the cognitive strategy *selects* what portion of the available perceptual information will be used to determine the response. Thus the problem in processing multichannel speech does not reflect an “attentional” limitation, but rather is due to perceptual-level interference effects on content recognition and stream assignment that produce incomplete or incorrect perception of the messages.

Overview of Modeling Work

The rest of this report is basically a narrative of the model development and its interaction with empirical data collection by our AFRL collaborators.

Our first model, the *Stream ID Detection Model*, was based on the concept of black-boxing the process by which the listener associates a word with a stream, reducing it to a simple detection process. That is, the model assumes that the listener detects not only the content of each message word, but also its *stream ID* which is an attribute that represents whatever characteristics would be used by the listener to identify which stream the word came from. SNR is assumed to affect this detection just it affects the detection of word content. While this model produced a good fit to the data, it did not scale to three- and four-channel data, which led us to question some of its assumptions and spurred the development of a more sophisticated model.

Our second model, the *Stream Tracking Model,* incorporates a separate perceptual

mechanism that assigns message words to streams using acoustic properties of the words. Along the way, we discovered problems in understanding the effects in the data, with the resultant difficulty of accounting for them in a model. Mixture strategies were required which suggested that subjects were not following a single strategy as might be the case if the Brungart(2001) methodology did not constrain the subjects' task strategies. In response, the AFRL group collected new data using a variation of the Brungart(2001) paradigm that imposed a bit of strategy control. This resulted in data with much less problematic effects, which could be well-accounted for by the stream tracking model. However, with this complete data set we were able to identify some distinctly different strategies used by individual subjects even with the improved paradigm. This led to AFRL collecting an additional data set with much more thorough control of subject strategy. The final data set is well fitted by a straightforward stream tracking model that shows promise of scaling to three- and four-talker tasks.

Stream ID Detection Model

Basic Concept

The basic hypothesis of the stream ID detection model was that speech processing requires two recognitions for each word in a speech message: the *content* of the word itself (i.e. recognizing that a sound was the word green), and the *stream* that the word was from (which requires discriminating which talker uttered the word). Thus each word object has the attributes of time stamp, content, and stream ID, and the content and stream ID are recognized independently, depending on the SNR and the talker differences (e.g. same talker vs different genders, etc.). All recognitions are assumed to be independent of each other. Figure 5 illustrates this very simple representation concept.

During the presentation of the messages, auditory memory fills with a mélange of representations of words with/without recognized content and with/without recognized stream identification. The task strategy implemented by the cognitive processor makes the inferences described above, then sorts through the available information to choose what response to make. Errors in the response are due to either undetected content or undetected stream IDs, with effects

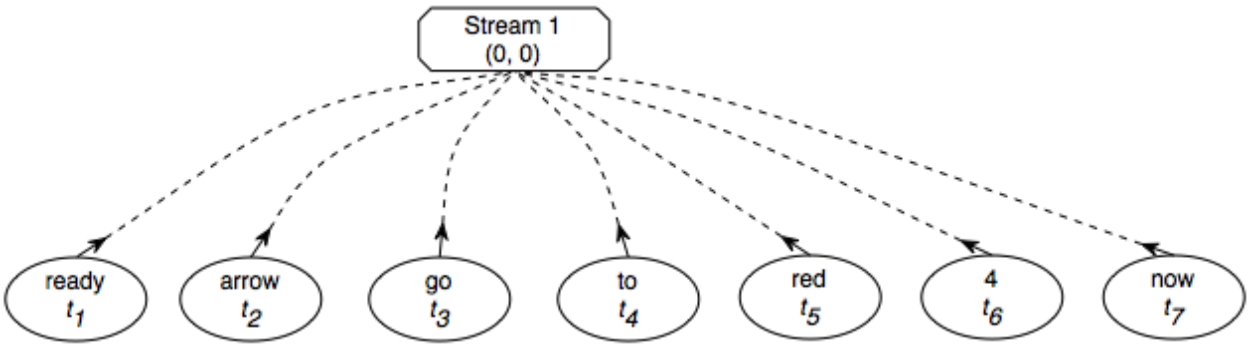

Figure 5*. Conceptual illustration of the representation in the stream ID detection model. Each word object has attributes of content, time stamp, and the identification (symbolic name) of the associated stream, shown by the arrows. The stream has only an (unused) x-y location attribute.*

that differ depending on which of the content words is affected. For example, if both call sign word contents are undetected, then all other content and stream IDs can be correctly detected, but the strategy will be unable to identify the target or masker streams, and so must make a guess between them. In the example shown in Figure 6, the word-by-word masking has not only obscured the content of some of the words, but also the stream membership of some of them. Stream 1 has the content "ready arrow go ? ? ? now"; Stream 2 has "ready ? go ? green 8", and the content items "? 4 now" are not perceived to be in either stream. However, in spite of the lost information, the task strategy can infer that Stream 2 must be the target stream, since Stream 1 has masker call sign content, and in this case "green 8" both have perceived stream ID for Stream 2, so they must be the correct response content.

Model Fitting Issues

To fit this model to the data, we had to choose the form of psychometric function for the content and stream recognition, estimate the parameters for these, and also we also had to experiment with different task strategies for making inferences and choosing responses. Because the stream IDs were detected, the strategy had to take into account that each word object could have three possible states for the associated stream ID: Stream1, Stream2, or none.

Despite the apparent degrees of freedom in the model, it proved to be very difficult to fit the data. An early result was that subjects were not optimal – in particular, they appeared to choose

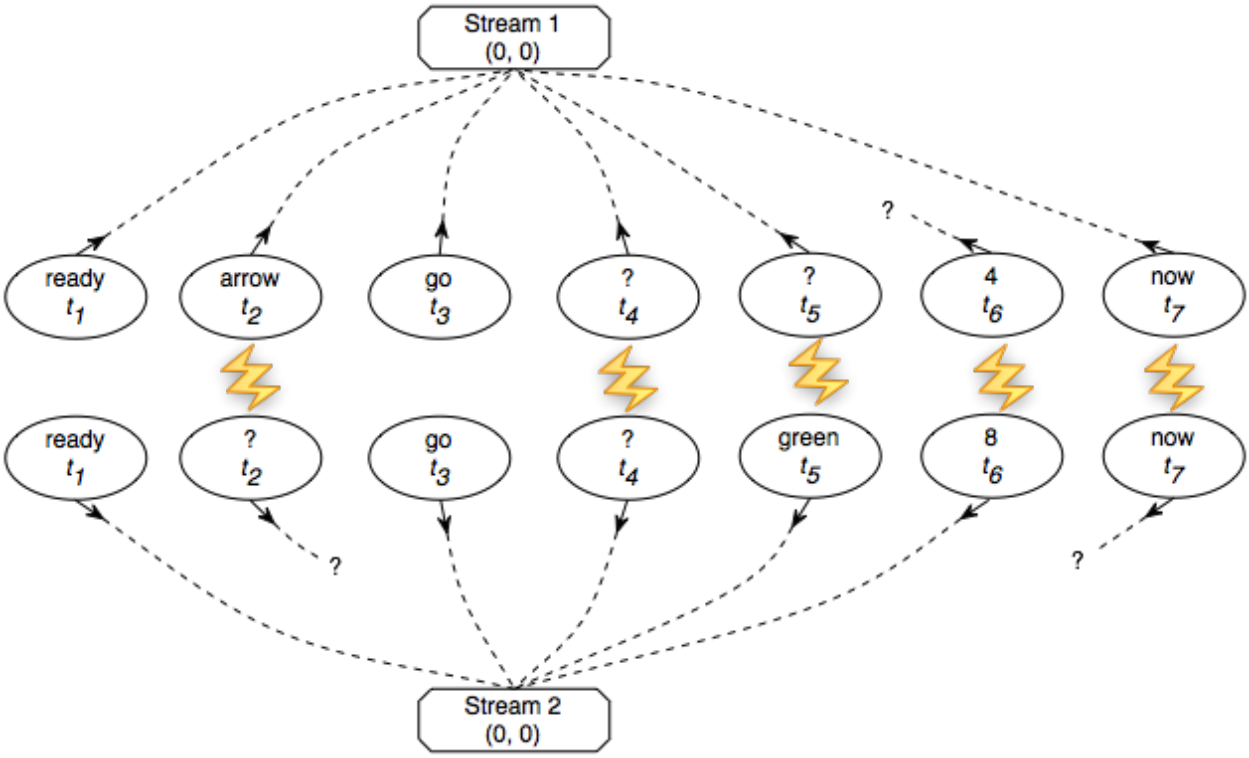

Figure 6*. Illustration of the effects of masking in the stream ID detection model. Masking effects can result in a failure to detect content (shown as a ? for the content), or a failure to detect which stream is associated with a word object (shown by a ? for the destination of the stream ID arrow).*

the Masker stream content more often than they should have. In terms of model-fitting, this was a difficult result; the optimum strategy is a good initial choice for a model strategy because it is usually a unique strategy that can be identified on the basis of formal analysis, rather than by iteration for goodness of fit. The fact that subjects were suboptimum meant that we had to iterate through a large joint space of model strategies, hypotheses about the form of the psychometric functions, and specific values for their parameters.

We devised a novel approach that solved the model-fitting problem by being willing to accept a strikingly unacceptable intermediate result. We broke the problem into two steps: (1) choosing a strategy that fit the data without any assumptions about the form of the psychophysical functions; (2) choosing psychophysical functions and their parameters that fit the data when used in conjunction with the good-fitting strategy from step 1. We performed step 1 by iterating over different strategies while assuming individual values for the probability of identifying content and stream for each content word at each SNR value in each talker-difference condition (a total of 180 probability values, 60 for each talker condition). A version of the model was coded in MATLAB, and its efficiency at optimization was used to quickly search the space of probability values to maximize the goodness of fit to the 210 available data points for a particular task strategy, represented as decision rules that mapped the space of content/stream recognition for each content word (4096 combinations) into a response decision. The task strategy and probability values were then put into the EPIC version of the model and a Monte-Carlo run was performed to verify the MATLAB model.

Despite the large number of "parameter values" – the individual probability values – only a small number of plausible task strategies could fit the data. Or to put it differently, an incorrect strategy could not hide behind the large number of individual probability values – there were too many simultaneous constraints imposed by the regularities in the data. For example, the optimum strategy cannot be made to fit the data even when it has 180 estimated "parameters." A simple version of the suboptimal use-masker strategy is closer to the data in some sense, but still clearly incorrect. The strategy that fits the data is one that uses maskers only when a key aspect of the stream identification is known to be unreliable.

Once we identified a good-fitting strategy, we then proposed reasonable psychophysical functions and fit them to the individual parameter estimates, in the process discovering that additional constraints could be imposed on these function parameters to reduce the total number of parameters involved. We verified the results by Monte-Carlo runs of the EPIC model using the functions and parameter values.

This two-phase approach enabled a fast search of an otherwise prohibitively complex space. It should apply to other models which combine a strategic component with parametric functions.

Results

We explored a complex space of task strategies and perceptual functions efficiently, and we arrived at a model for the Brungart (2001) data that fits remarkably well; these are the predicted points in Figure 7.

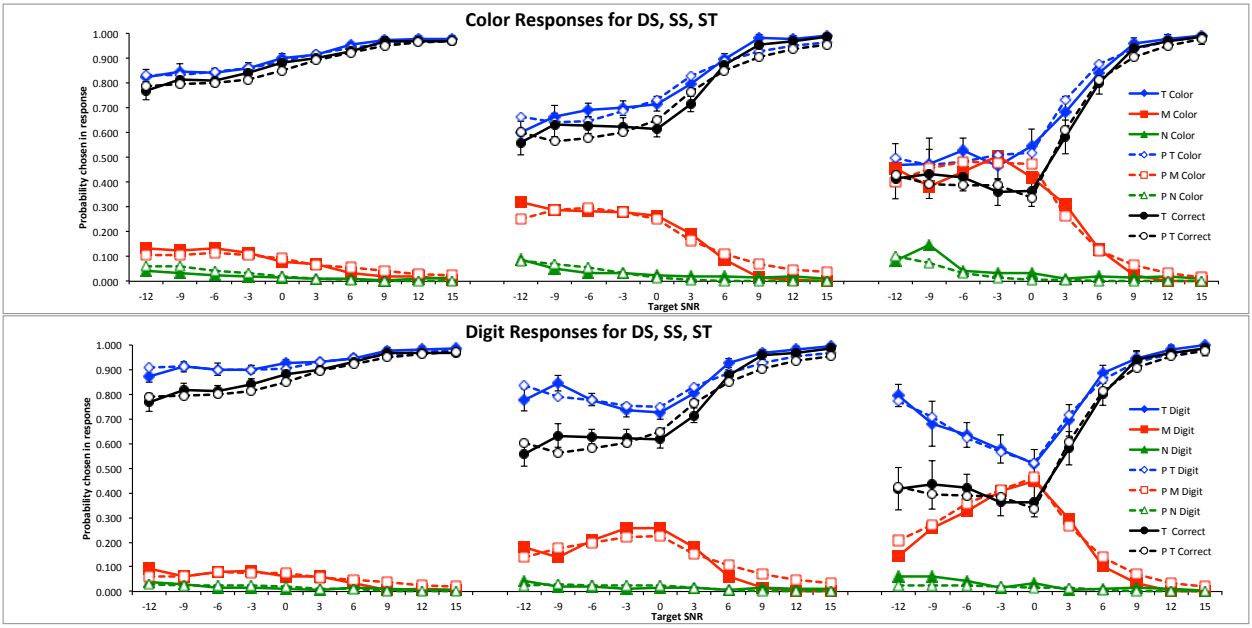

Figure 7*. Stream ID detection model results. Observed (solid lines and points) and predicted (dotted lines and open points) probabilities for each response choice in the Brungart (2001) data as a function of signal-to-noise ratio (SNR) of the target message relative to the masker message. The vertical axis is probability of response; the horizontal axis is SNR in dB. Color responses are in the upper panel; digit responses are in the lower panel. The three talkers conditions are shown left-to-right in the three subpanels: Different Sex (DS), Same Sex (SS), and Same Talker (ST). Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker. Note the excellent fit of this model. V9bSubOptPairX_JointV16_subopt1b_PP5a.xlsx*

See Appendix 1 for details on the final strategy, but it follows the common strategy features described above, with the key feature being that it adopts a use-masker option depending on whether the target call sign content was actually detected, or *observed*, meaning that the target stream ID is directly known rather than being inferred. In other words, the strategy says that if the Target Color or Digit content is missing, and the target stream is inferred, and there is color or digit content known to be from the masker stream, then use it in the response. The rationale for this apparently suboptimal strategy is that if the target call sign content was not in fact detected, then the assignment of stream IDs to target vs masker could be unreliable - maybe what you thought was the masker was actually the target. This means that "using what you heard" could be a better choice than making a pure guess from the other alternatives.

The perceptual detection functions in this model are shown in Figure 8 for each talker condition and content word (call sign, color, digit). The stream detection function (upper panels) is a two-sided exponential function with a probability floor centered at SNR 0, which captures how relative SNR contributes to identifying a stream both when it is the quieter as well as the louder of the two streams. The content detection function was a familiar ogive with a probability floor that relates increased SNR to increased probability of recognition (bottom panels in Figure 8). These functions were well-behaved across the talker and SNR manipulations, being strictly

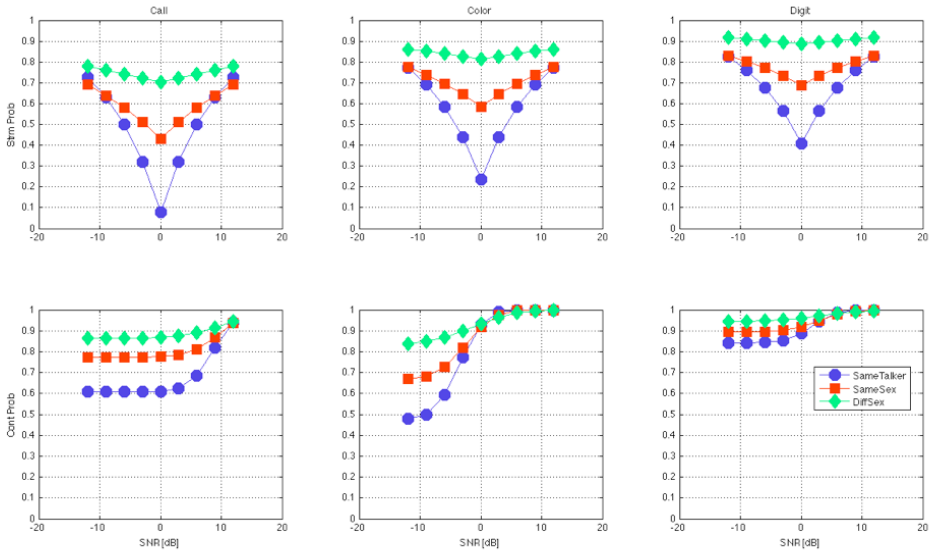

***Figure 8**. Psychophysical functions used in the model to fit the Brungart (2001) data. The horizontal axis is the SNR in dB; the vertical axis is the probability of recognition. Reading from the top in each graph, the talker conditions are color-coded as green, red, and blue for Different Sex, Same Sex, and Same-Talker respectively.The upper panels show the probability of identifying the stream correctly for the call sign word, the color word, and the digit word, from left to right. These functions are an exponential probability function with a minimum probability symmetrical around an SNR of zero. The exponential parameter was constrained to the same value for the three words in each talker condition, so only the minimum probability varied as a function of word and condition. The lower parameters show the probability of identifying the word content correct for the three words. These are Gaussian functions with a minimum probability whose variance parameter was constrained to the same value for the three words in each condition, so only the mean and minimum probability varied as a function of word and condition.*

ordered from Different Sex to Same Talker, and show improved detection of stream identification through the course of the utterance, which agrees with a common interpretation of stream formation phenomena.

In summary, the model for 2-channel listening combines psychoacoustic components for perception with cognitive components to implement the inference and decision procedures involved in the task, and provides an excellent account of the data.

Problems with the Stream ID Detection Model

Plausibility of detection functions

Although this model fit the data very well, there are some problems. First, it does not contain any mechanism for how streams were identified or could "build up" over time, which has long been a concern in psychoacoustics. Rather, we had "finessed" this process with a black-box treatment. Second, the black-boxing of stream ID detection has a couple of odd features in the model. If the stream ID is detected, it is always detected veridically — it is not possible to misidentify the stream ID of a word object. Also, the two-sided detection function is unusual in that it claims that the Stream ID can be detected better as its SNR gets more negative, unlike

almost every other psychophysical detection function ever proposed. While the symmetry makes sense in terms of accounting for how a stream can be identified if it is the quieter of the two, this seems unnecessary because in this experiment, any time one stream is quieter than the other, the other is automatically louder, meaning that if the louder stream can be identified, the identity of the quieter one follows. So why can't the stream ID attribute be detected according to a more conventional detection function? A related concern is that “floored ogive” content detection functions are somewhat unconventional as well; a prototypical detection function ranges from 0 to 1 over the full SNR range.

Problems with predicting three- and four-talker effects

The most serious problem was that the model did not scale to additional talkers. We discovered this using the data from Brungart, Simpson, Ericson, & Scott (2001) which is basically the same CRM paradigm except two or three masker talkers were included to yield a total of three and four talkers. We used the reported subset of the data in which the masker talkers were all of the same type - i.e. different talkers of different sex from the target talker, different talkers of the same sex as the target talker, and the same talker as the target talker. Unfortunately, in these papers, the data is much less detailed; only the proportion of completely-correct responses (both color and digit are from the target message) is reported. For ease of comparison, Figure 9 shows Both-Correct responses from Brungart et al. (2001) for two-, three- and four-talkers (corresponding to one, two, and three maskers), where the two-talker data is the same as reported in Brungart (2001). In other words, adding a second masker causes performance to fall off a cliff; a third masker makes it only somewhat worse. Accounting for this sudden drop in performance is a clear challenge.

Extending the model and the architecture to multiple maskers

We extended the stream ID detection model to multiple maskers with two changes:

1. We modified the masking model to handle the case of more than two word objects masking each other. For each word object, an effective SNR was calculated by summing the total masking power in the other word objects so that having two maskers of the same loudness decreased the SNR of the target by 3 dB. The content and stream ID detection

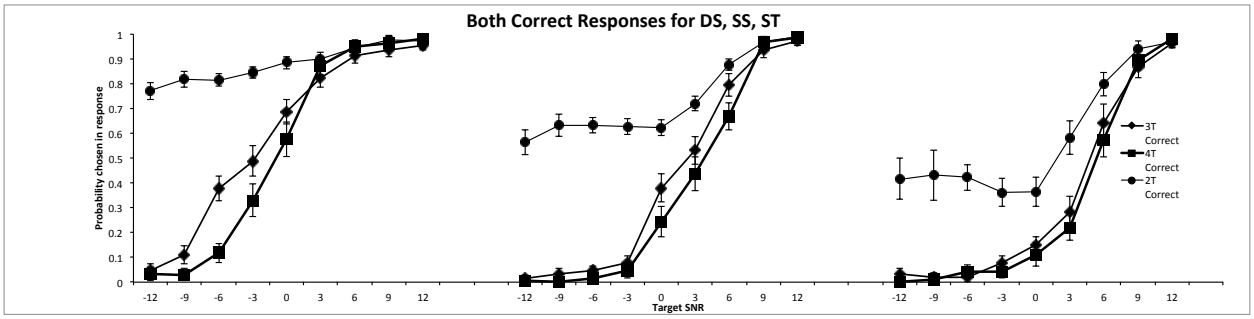

Figure 9. *Observed proportions of both-correct responses as a function of SNR for two, three, and four talkers from Brungart et al. (2001). The three talkers conditions are shown left-to-right in the three panels: Different Sex (DS), Same Sex (SS), and Same Talker (ST). The highest performance is for two talkers; three and four talkers produce much lower, and similar, performance.*

functions used this effective SNR to determine whether content or stream ID was masked for a particular word object. This effective SNR for multiple maskers is retained in all of our multiple-masker models described in this report.

2. We expanded the production rules to deal with multiple maskers. This was relatively simple because the change is mostly in the rules for inferring target vs. masker content when some of the content or stream IDs are missing. For example, if the target stream could be identified from its call sign, then all other streams could be tagged as masker streams even if their call signs were unheard. If all masker call signs were heard and associated with a stream ID, than any "odd" stream had to be the target even it its call sign was unheard. However, notice that relative to the two-talker case, such inferences are weaker. For example, if there are three talkers, and both the target call sign and one of the masker call signs are unheard, then it is not possible to infer from the known masker stream which of these other two streams was the target. But if only two streams are involved, then the status of one can always be inferred from the status of the other.

Model failure

This expanded model was run in the three- and four-talker conditions and compared to the data reported in Brungart et al. (2001). What we discovered was that the model did not come close to fitting the data, and was in fact fundamentally wrong - the two-sided stream detection function caused the model to perform very well at low SNR in the three- and four-talker case, seriously contradicting the data. If we changed the stream detection function to be a more conventional normal-ogive function, the model could fit the three- and four-talker data, but was then fundamentally wrong with respect to the two-talker data.

In more detail, Figure 10 shows what happened when the expanded production rule strategies were used with the two-talker psychometric functions for content and stream ID detection and applied to three- and four-talkers. The observed Both-Correct proportions are the black solid points and lines; the prediction values are the black open points with dotted lines. Since only the Both-Correct data was reported, the predictions for target responses, masker response, and neither responses cannot be compared to the data.

The problem is obvious - the observed Both-Correct proportion drops substantially at negative SNRs, but the predicted performance stays high, apparently due to the high sensitivity of stream ID detection at low SNR produced by the double-sided stream ID detection function.

We investigated whether this was in fact the problem by fitting the three- and four-talker data assuming a "normal" single-sided gaussian detection function for the stream ID. As shown in Figure 11, this produced a good fit for three- and four-talkers, but the same function applied to the two-talker data was completely wrong. Keep in the mind that production rules are accounting for the strategy differences implied by less ambiguous two-talker case where inferences about missing information are stronger. So the problem in moving from a single masker to multiple maskers was that the fundamental form of the stream ID masking function appeared to be different - this is not a simple parameter value difference, but a fundamental difference in how the data could be modeled, even though we had the full power of EPIC's cognitive inferential

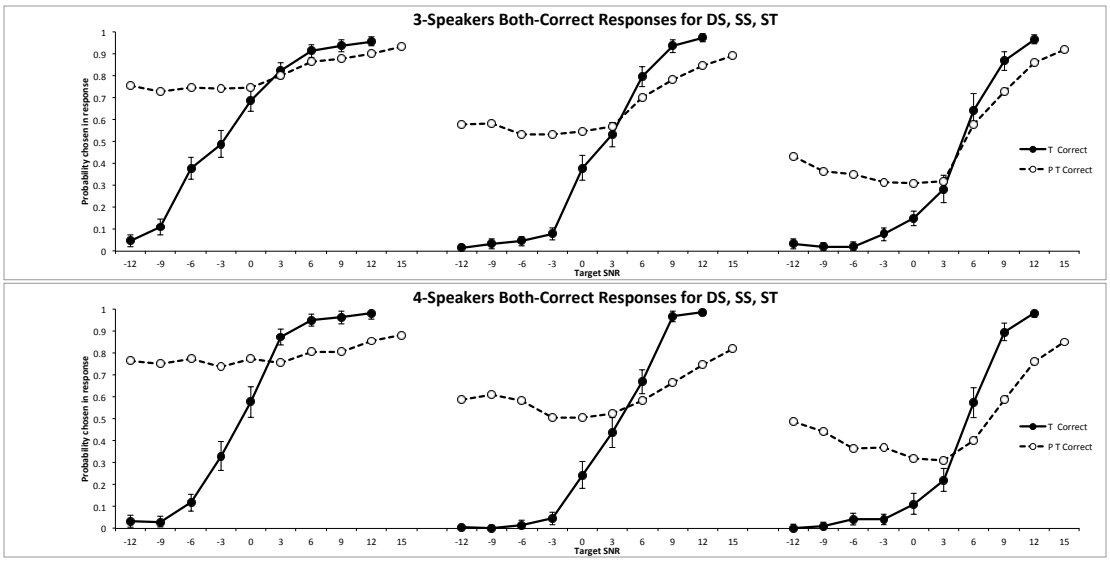

Figure 10. *Observed (solid lines and points) and predicted (dotted lines and open points) probabilities for both-correct response choice in the Brungart et al. (2001) data. The predictions are clearly incorrect, especially below 0 SNR. MSV5b_3S_r1, MSV5b_4S_r1*

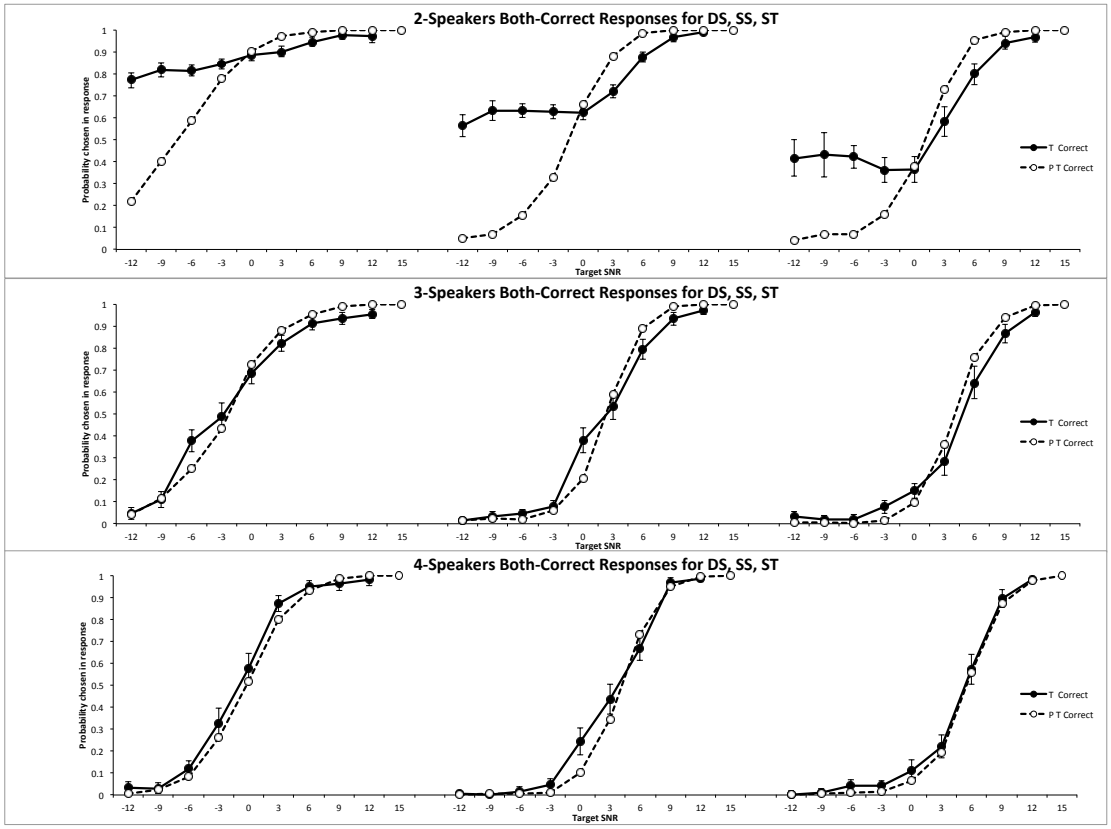

Figure 11. *The predictions are close to the data for three and four talkers, but seriously discrepant for two talkers. MSV5b_2S_r5, MSV5b_3S_r10, MSV5b_4S_r10*

machinery available to apply all of the information from the speech input. The difference between one masker and two maskers was not being captured by our modeling approach. This made us suspect that the underlying theoretical structure of our model for the one-masker case was incorrect, even if it could be made to fit the data very well by an unusual detection function.

Because we doubted the validity of some of the basic assumptions, we decided not to attempt to publish our two-talker stream detection model results, even though a model of this precision and explanatory power was unprecedented in this field. Rather we proceeded to rethink the fundamental structure of the model, in particular, the decisions we had made in black-boxing the low-level speech perception processes.

Stream Tracking Model

Stream Tracking Concept

Recapitulation

The problems with black-boxing have already been alluded to above. In practical terms, in order to construct a cognitive architecture and model for a complex psychological phenomenon in a reasonable amount of time, we have to decide what to black-box and what to represent in finer detail. Our original concept for black-boxing the perceptual process was that each word object had two independent perceptual attributes: content and stream ID. Content was the recognized word, and the stream ID was a stand-in for whatever perceptual attributes allowed one to distinguish one talker or message from another. We assumed that these attributes were either detected in veridical form, or not detected at all; that is, there was no possibility that a stream ID would be detected but it would be the wrong stream ID.

The difficulty with our black-box assumption became more apparent when we considered the case of the Same-Talker condition at SNR 0. The fits from our two-talker model said that under these conditions the stream ID would be undetected most of the time. However, it seems intuitively obvious that a listener in this case would be able to tell quite easily that the two messages were from the same talker, which implies that the stream ID attributes were being detected, but the different words simply could not be reliably assigned to one of the two messages! In this case, a bit of analysis shows that there is some tendency for messages from the same talker to differ in loudness and pitch contours, but two messages from the same talker will show considerable word-to-word variation in loudness and pitch. Thus, the listener in this condition is likely to detect the stream-relevant acoustic properties of the words, but this information is ambiguous in deciding which words come from which message stream.

Streams inferred from the acoustics

This led us to a different black-box analysis of stream ID attributes. We assumed that stream ID would be an inferred or perceived property that is based on acoustic attributes of the sound that are always detected. Two sound attributes that would be relevant in this type of study are the pitch and loudness of the sounds coming from the different talkers. For example, if the two talkers are different sexes, words from the female talker are likely to have higher pitches than words from the male talker. So the perceptual system could assign all higher-pitched words to

one stream, and the lower-pitched words to the other stream. Likewise, if one of the talkers is speaking more loudly than the other, the perceptual processor could assign all of the louder words to one stream, and the softer words to the other. If the two talkers differ consistently in these attributes, the stream assignment of the words will be consistent and correct. However, if the two messages are similar in loudness or pitch, then it is possible that at some point in the message, the stream assignment will "flip" and the wrong words will get assigned to the streams.

Stream Tracking Mechanism

We changed the auditory representation of the model to accommodate the different black-box model. Instead of a stream ID attribute, each word object carries its mean pitch and loudness as stream attributes. We assumed that these stream attributes are *always* detected, but that the assignment of word objects to streams is a perceptual process subject to error. Each stream object is thus a percept, and carries the loudness and pitch information accumulated for all the word objects that have been assigned to that stream.

Figures 12 through 14 illustrate the stream tracking process. As shown in Figure 12, the first words in the two messages result in two word objects that are simultaneously present; a stream percept is created for each word and initialized with the pitch and loudness information for its word. The next pair of words starts a process in which each word is assigned to its closest-matching (on some metric) stream percept, which in turn is updated to reflect the loudness and pitch of its newly assigned word. Thus as the two messages are presented, each word gets assigned to a stream based on how well that word corresponds to the words that have already

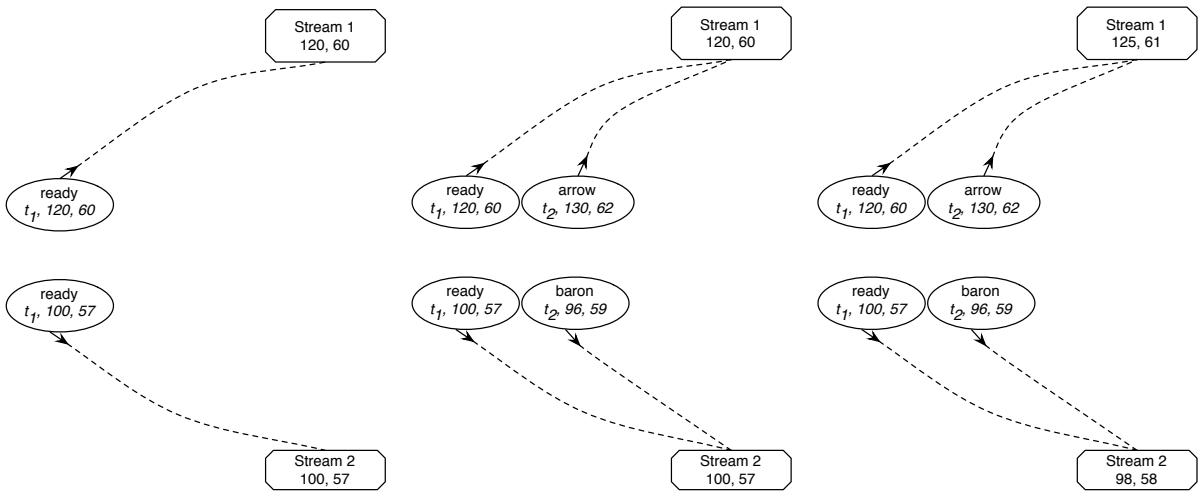

Figure 12*. Illustration of stream tracking as each word in the two messages are presented, assuming that content is always correctly detected.The left panel shows the state after the words "ready" have appeared and the two stream objects have been initialized with the mean pitch and loudness of the two words, associating the top word with Stream1 and the bottom word with Stream2.The middle panel shows the two call sign words appearing, adjacent to the "ready" words in the same message. The call signs have been assigned to the stream object that most closely matches their pitch and loudness. The right panel shows the stream objects have been updated to reflect the mean pitch and loudness of the two words assigned to them.*

been assigned to the stream. In this example, the call sign words are correctly assigned to the same stream as the initial "ready" words in the same message.

As shown in Figure 13, if a word turns out to match the wrong stream better than the right stream, then an error can result; in this example, the color words are assigned incorrectly, leading to a "switch" in the stream assignment, and bogus updates of the stream percepts.

Finally Figure 14 illustrates how the stream assignment can switch back to the correct stream if the next digit words match the correct current stream percepts better than the incorrect ones. This produces a situation in which the colors and digits from each message are assigned to different streams, meaning that the color will not match the target content, but the digit will.

This new model seems appealing, but it is rather more complex than our original model that was based on the detection of simple stream IDs; in particular, it must be determined what algorithms should be used to compare words to streams and update the stream information, and how the loudness and pitch information for words should be represented in the model.

An Abstract Model of Stream Tracking

Rather than plunge ahead with an effort to apply a full stream-tracking model, we explored the feasibility of the idea with a much simpler mathematical simulation model, which we called the *abstract stream-switch model.* The purpose was to check that our stream tracking concept could indeed account for the data, and at the same time give us an advance reading on what characteristics a stream tracker would require in order to account for the observed performance data. The basic idea for this model was to assume that the stream tracker starts with the correct assignment of call sign words to streams, and then at the color word can *switch* to the wrong stream or *stay* with the correct stream, and then switch or stay again at the digit word. In the

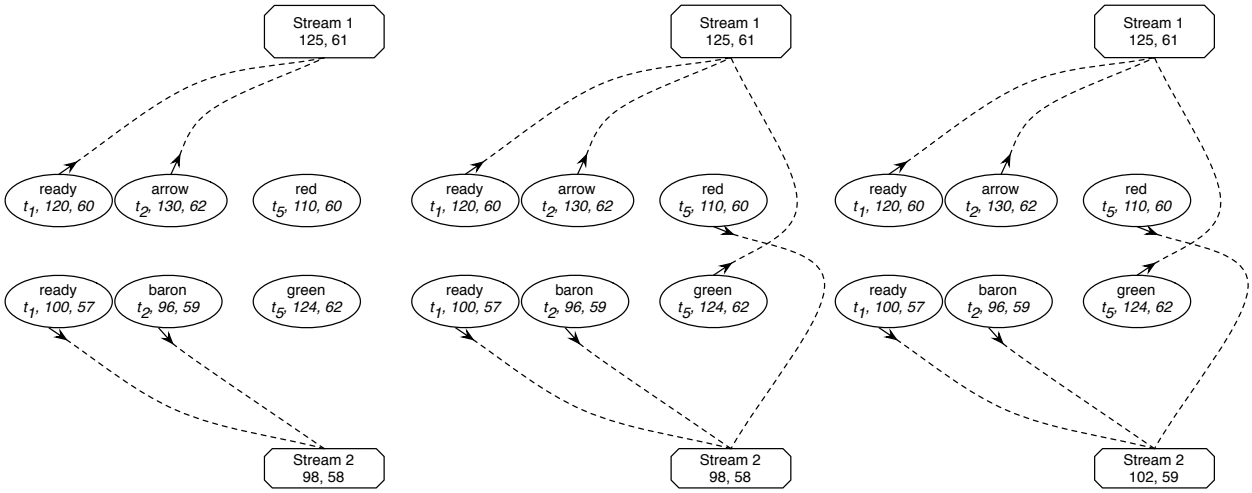

Figure 13*. Continuation of Figure 12, with the "go to" filler words omitted for brevity. In the left panel, the two color words have appeared, but their pitch and loudnesses are a better match to the other stream's current values, and so as shown in the middle panel, they have been assigned to the wrong stream object. The perceived stream of the messages has "switched" from the initial assignment. The right panel shows the updated values of the stream objects after the incorrect assignment.*

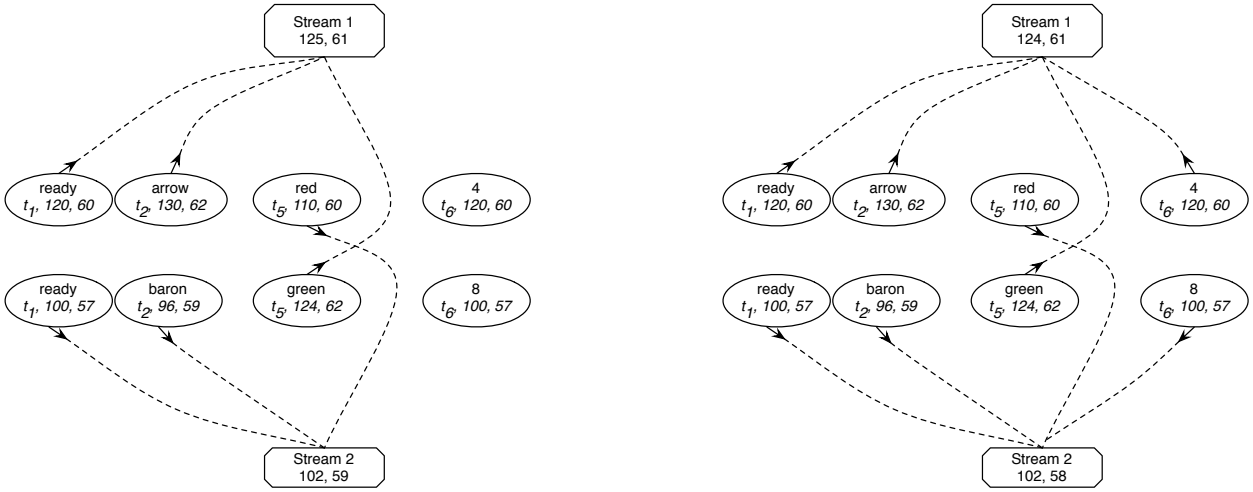

Figure 14*. Continuation of Figure 13. In the left panel, the two digit words have appeared, and their pitch and loudness matches the correct streams best, and so the stream assignment "switches" back to the correct streams for the digit words, as shown in the right panel; the updated stream values will be used to assign the last word "now" (not shown). At this point, Stream1's content is "ready arrow green 4" and Stream2's content is "ready baron red 8." The response will thus have the incorrect color and the correct digit.*

meantime, the content of the color and digit words is either recognized or not, depending on a separate pair of detection functions. The model then uses the response rules from the complete EPIC model to choose its responses depending on the stream ID and content available for each color and digit word from each stream. Using optimization routines in Matlab and Monte Carlo simulations, we estimated the probability functions for content recognition and stream-switching that fit the data, and we also compared hypotheses about task strategy as they would be represented in the production rules.

The abstract stream-switch model added an important feature of mixing the avoid-masker and use-masker strategies: if target digit content was missing, then with some probability, the avoid-maskers strategy was used, or the use-masker strategy was used. The limitation of the mixture strategy to missing target digit content was another innovation; previously our guessing strategies had applied uniformly to both color and digit content. However, on average, the digit responses are more accurate than the color responses, even though there are more possible digits than colors. Some of our earliest work on the project showed that the digit words in the corpus were less synchronized in time than the color words, which would make the digits easier to recognize than the colors. In addition, because there are more digits than colors, the digit content is more diagnostic of a correct response than color content. All of these differences justify treating missing target digit content differently from missing target color content in the guessing strategy.

Mixture models are used occasionally in production-system cognitive models to deal with the fact that in most psychological experiments, the subject's task strategy is neither assessed, trained, nor controlled, and so the data might reflect an unsystematic amalgam of different strategies. If only aggregated data is available, it is not possible to determine whether this

strategy variation is within- or between-subject. Our assumption in this third model is that the variation is effectively within subject in that the mixture decision is made independently on each trial on which the guessing strategy is triggered. The fit optimizer adjusted the mixture probability along with the content detection and stream-switch probabilities.

The result for the abstract stream-switch model was a very close fit to the data using reasonable probability functions, shown in Figure 15. These results demonstrated that we should indeed be able to construct a model that accounts for the difficult and puzzling two-talker results, and also set some requirements for what the model will have to include. First, a mixture model for the guessing strategy is required that is triggered by missing target content in the most diagnostic case, the digit content. It is possible that a fixed mixture probability will account for the data reasonably well, which would mean that only a single parameter value for this factor will be needed. In addition, the EPIC architecture model should be able to fit the data with reasonable and well-behaved content detection functions. These functions will need to have different parameters for the different talker conditions that reflect similarities in the talkers. A key assumption in this abstract model was that the stream-switch probability is symmetric for the color and digit words; that is, the probability of a stream switch was the same for the digit word as for the color word - it did not depend on the state of the stream.

Most importantly, the behavior of the stream switch/stay probability function needed to fit the data suggests two important requirements for the stream tracking mechanism:

- The stream assignment is generally reliable when the SNR is different from zero. This means that the effects of the different talker conditions mainly involve whether the streams can be segregated when the loudnesses are similar. In the case of Different Sex, the consistent pitch

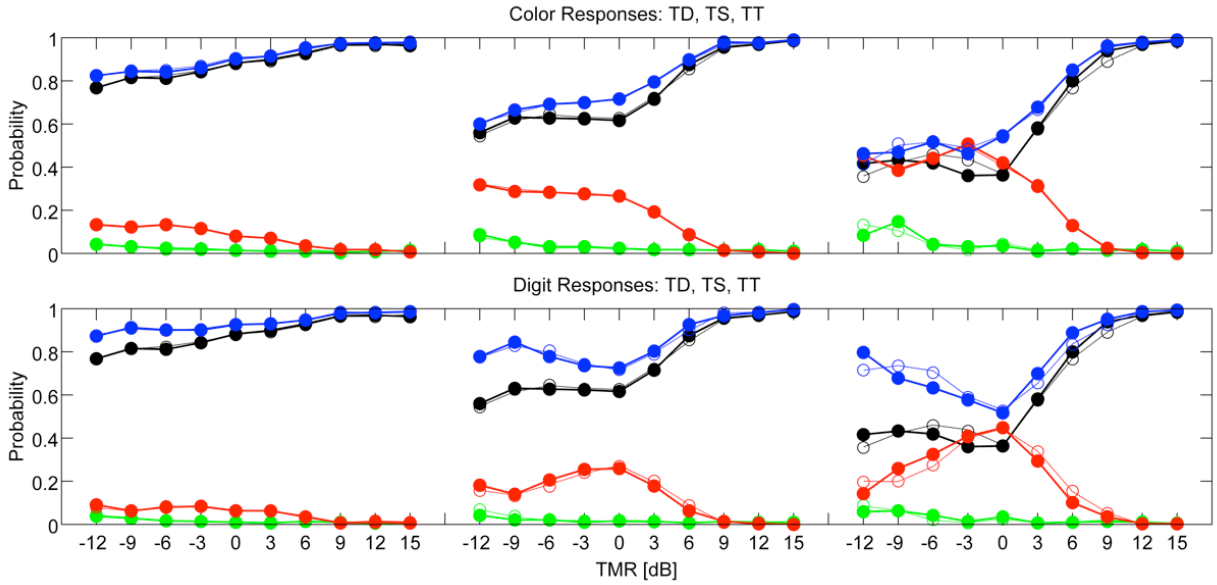

Figure 15*. Observed (solid points and lines) for the Brungart (2001) data and Predicted (open points, dotted lines) values produced by the symmetrical stream-switch abstract model. The fits are extremely close, and the estimated parameter values are reasonable.*

difference suffices, and results in a very high stream assignment accuracy. In the case of Same Talker, the stream assignment accuracy is very poor, as expected, because all of the speech characteristics are similar. For the Same Sex condition, the stream assignment accuracy is intermediate between the other two conditions; if a stream attribute can be identified that produces assignment accuracy corresponding to the stream "stay" probability function in the model, then this condition can be accounted for.

- The probability of a stream assignment switch between color and digit is relatively low, as shown by the high digit "stay" probability in the above figures. This means that whatever attributes are used to identify and assign the stream, they are stable going from color to digit, and more variable going from call sign to color.

A New Two-Talker Dataset: Replication 1a

The importance of the strategy mixture in the abstract model inspired our collaborators at AFRL to conduct a new experiment, named *Replication 1a* here; they made the complete dataset available to us. This experiment followed the same paradigm as the Brungart (2001) experiment, but with two changes: Rhe SNR range was changed to include more negative values than Brungart (2001), going from -18 to +9 dB instead of the original -12 to +15 dB. This would determine whether stream identification performance did indeed drop off at sufficiently negative SNRs, which was not clear in the Brungart (2001) results.

The more critical change was in the feedback given to subjects. The original experiment did not provide feedback to the subjects on whether their responses were correct. As a result of our discussions regarding the importance of task strategy, and how subjects could develop a strategy during the experiment, the AFRL group provided subjects with feedback on each trial. This one methodological change made a substantial difference in the outcome of the experiment, eliminating the most puzzling effects visible in the Brungart (2001) results. These results are shown in Figure 16.

Notice how first, the curves for color responses (top panel) and the curves for digit responses (bottom panel) follow very similar patterns, unlike the large, and puzzling discrepancies in the original Brungart (2001) results (see Figure1). This is extremely important, because trying to make colors and digits behave differently in the Same-Sex (SS or TS) and Same-Talker (ST or TT) conditions has been the primary obstacle in trying to fit any of the models we have considered. The second new result is that at the new, more negative SNR values, accuracy drops off, as would be expected when the target is presented at a sufficiently low level, relative to the masker, that it can no longer be clearly heard. The data from Brungart (2001), like these results, show that subjects can pick out the target stream based on its being softer than the masker stream, but did not go far enough to see the accuracy drop when the target stream becomes so quiet that it is likely masked by the masker stream. The new results thus provide a better anchor at the low SNRs.

Verification of stream switch effects

Because we had the full dataset available, we were also able to make a direct test of a key assumption in our abstract stream-switch model, namely that probability of switching streams

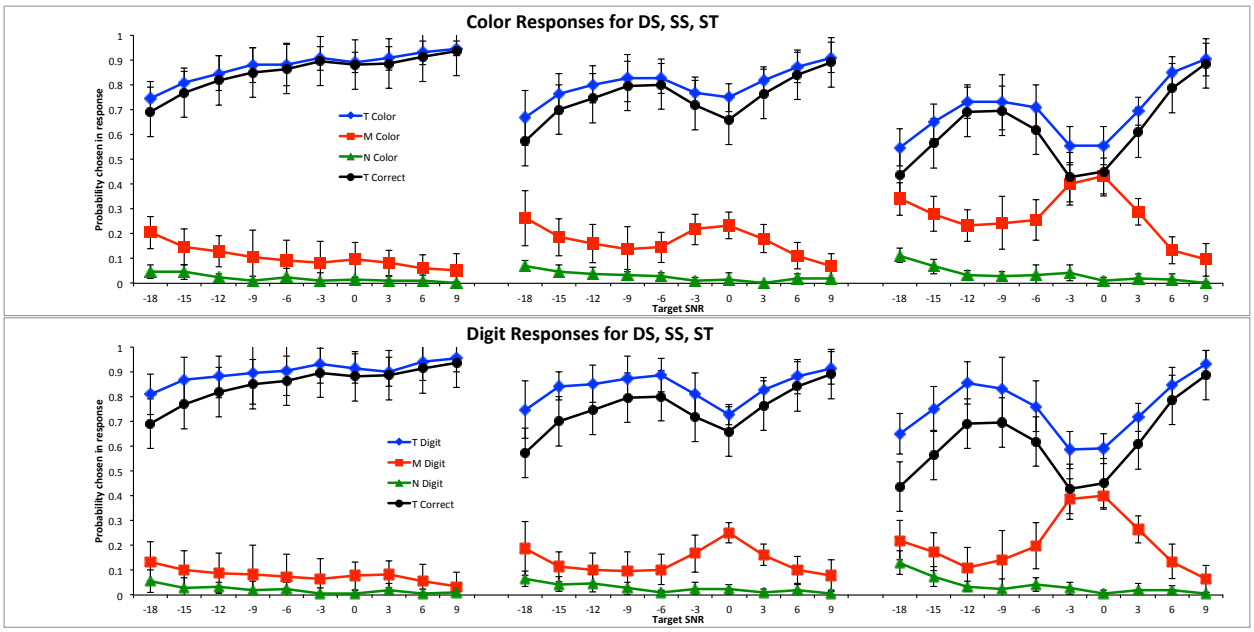

Figure 16. *The Replication 1a results. Probability of responses as a function of SNR for Different-Sex, Same-Sex, and Same-Talker conditions. Top panel is for color responses, bottom panel for digits. Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker. Error bars shown are 95% confidence intervals for the proportions over the pooled responses from all subjects.*

does not depend on the stream state, meaning that the probability of switching to the other stream on the digit word would be the same as the probability of switching to the other stream on the color word. Figure 17 shows for the Replication 1a data, the observed conditional probability of *staying* in the same stream for the digit as for the color as a function of talker condition and SNR. The blue curves show the probability P(target digit | target color) that the digit response will be the target response given that the color response was from the target; the red curves show the probability P(masker digit | masker color) that the digit response will be the masker response

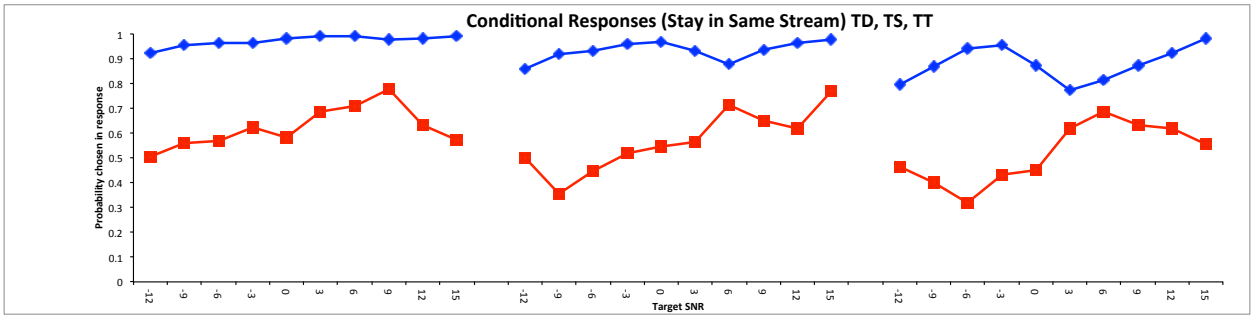

Figure 17. *Asymmetrical stream switch probability: Conditional probabilities for digit responses staying in the same stream as color responses, as a function of SNR and talker condition, in the Replication 1a data. The blue (top) curve shows the probability that the digit response will be from the target given that the color response was from the target. The red (bottom) curve shows the probability that the digit response will be from the masker given that the color response was from the masker.*

given that the color response was from the masker. Notice how the target stream *stay* probability is generally higher than the masker stream *stay* probability. Rephrasing this: If the color is chosen from the target stream, the digit is almost always chosen from the target stream. But if the color is chosen from the masker stream, the digit is often chosen from (switches to) the target stream.

Since this result contradicted our original assumptions, the abstract stream-switch model was modified to allow the switch/stay probability to be asymmetric, and assumed a strict avoid-maskers response strategy. The resulting fit, shown in Figure 18, was highly encouraging.

Notice that this asymmetry corresponds directly to how we might expect stream trackers to work. Namely, if the tracking of the target stream was working well enough to allow the listener to identify the target color correctly, then it is probably working well enough to also allow the digit to be identified. But if the tracking is going poorly enough that the color is chosen from the masker instead of the target, there is still some chance that the digit will be chosen from the target.

A Stream-Tracking Model for the Replication 1a Data

Because the color-digit asymmetry had caused so much difficulty in modeling the Brungart (2001) data, we decided to try our full stream-tracking model on the Replication 1a experiment data before revisiting the Brungart (2001) data.

Using stream attributes from the corpus

A key idea of the stream-tracking model developed here is that it is driven by the properties of the speech signals, which can be computed from the actual speech messages rather than estimated as parameters to fit the task performance data. For example, rather than estimating the

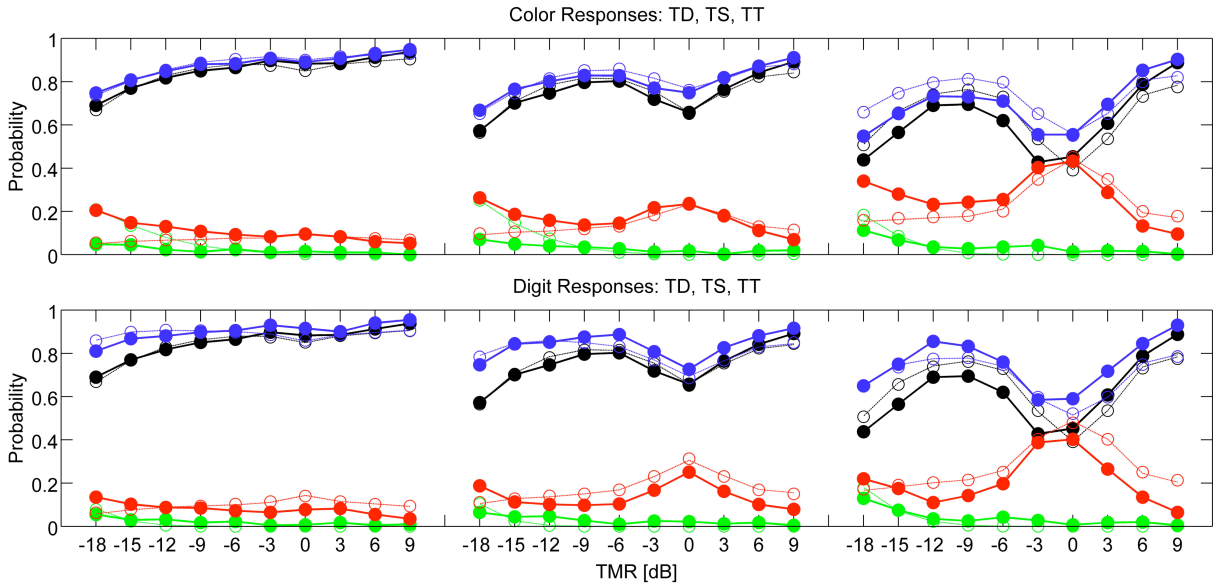

***Figure 18**. Asymmetrical abstract stream-switch model results. The Replication 1a observed values (solid points and lines) and predicted values (open points and dotted lines) from an abstract stream-switch model that allows asymmetry between stay probabilities for colors and digits.*

probability that a female voice stream could be distinguished from a male voice stream, we compute the fundamental pitch of male and female utterances in the CRM corpus, and supply these pitches to the model. If the sequence of pitches of the female utterance are always greater than that of the male utterance, then the stream tracking mechanism will correctly assign the word content information to the two acoustic sources. Thus rather than additional collection of free parameters for the distribution of the pitch and loudness attributes of the words, we used the actual properties of individual words in individual utterances. The simulated experimental environment samples a pair of utterances for each trial in the task, and then supplies the actual pitch and loudness values of each word to the simulated human's auditory system.

Calculating stream attributes for the CRM corpus

Calculating the mean pitch and loudness values for each word in each utterance in the corpus requires a *segmentation* of each utterance - the start and stop time of each word. A complete segmentation had not been done for the CRM corpus, and performing one requires a substantial effort because a skilled human listener must make judgements on the segment boundaries. In order to explore whether our stream-tracking model was viable, we performed a "quick and dirty" segmentation of the corpus that took advantage of the fact that the utterances were fairly consistent in duration and timing, corresponding to six "beats" for the segments [ready][call sign] [goto][color][digit]. Thus we divided the duration of each utterance into six equal-length segments, and computed the mean pitch and loudness of each such segment. We maintained our assumption that the word/segments of the pair of utterances were aligned in time.

Since the stream-tracking model tracks the pitch and loudness of the segments appearing in the streams, it is useful to examine the pitch and loudness of utterances in the corpus in terms of the individual talkers. Figure 19 in the left panel the mean pitch (Hz) of each of the six segments for individual talkers. The female talkers are all in the upper group, with mean pitches around 200 Hz, in a pattern that generally drops from beginning to end. The male talkers are in the lower group, around 100 Hz, with a flatter pattern. The interesting result here is that there is a small but relatively consistent difference in average pitch between talkers of the *same* sex, which will certainly be important in the same-sex condition (SS or TS).

The mean loudness for each segment for each talker is shown in the right-hand panel of Figure 19, where a 60 dB baseline was assumed. While there is an overall trend toward loudness decreasing about 3 dB in the course of the utterance, there is little or no difference between either genders or talkers in the loudness statistics. This is not surprising because the corpus was normalized to have the same overall loudness, so that the experimental manipulation of SNR (e.g for the 6 dB SNR condition, +6 is added to the baseline for the target message and 0 for the masker) will thus be the source of stream loudness differences in the task.

Auditory model specifics

For each trial, the experiment simulation samples two utterances and then supplies EPIC's auditory system with the content, loudness, and pitch of each segment. As before the loudness is in dB, with the SNR condition decibels being added to the baseline of 60 dB and the result added to the loudness of the target segment. The pitch was converted to semitones ($12 \cdot \log_2$(*pitch in*

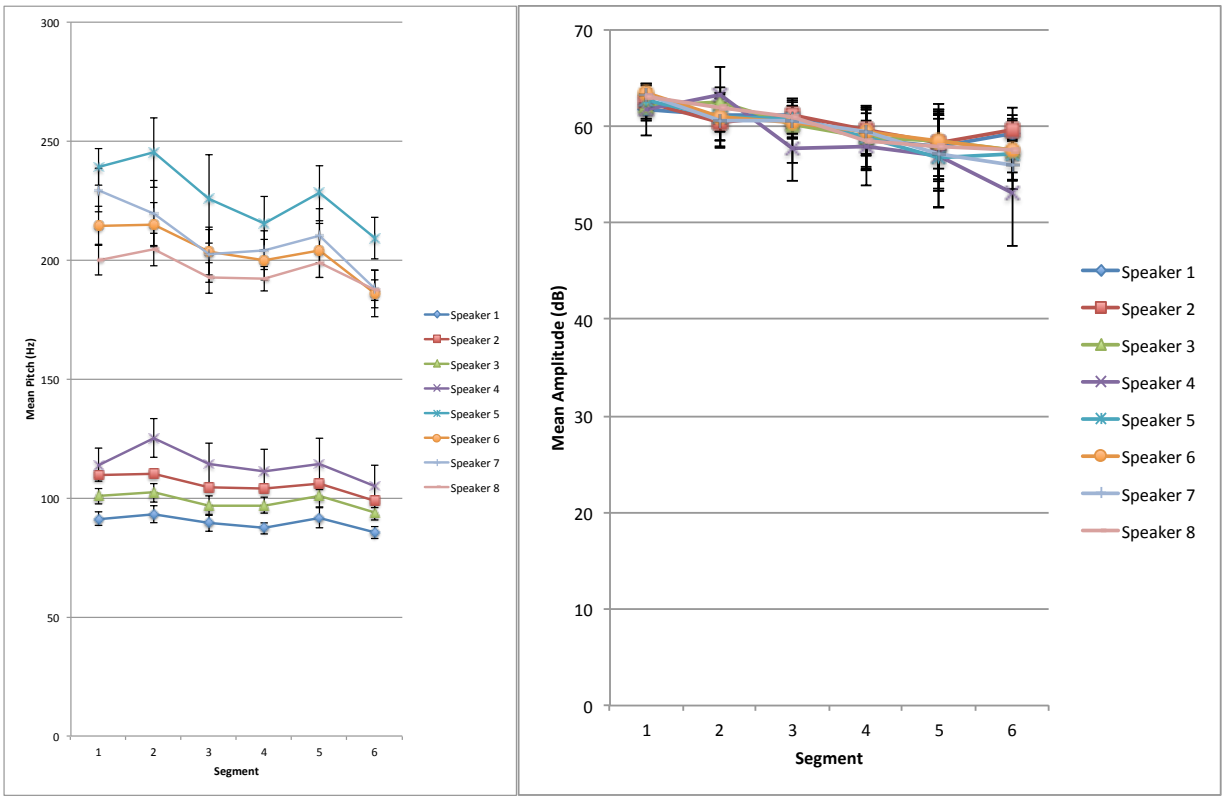

Figure 19*. Corpus statistics for each segment for each talker. Segments are shown on the x-axis, with a separate curve for each talker. The left panel shows the mean fundamental pitch in Hz; the right panel shows the mean loudness in dB. Error bars show the standard deviation of the data points contributing to each mean.*
crmstats_v15_corrected_edited crunch.xlsx

Hz)) which provides a logarithmic scale for pitch similar to the logarithmic decibel scale normally used for loudness. The pitch difference between two word objects was capped at 4 semitones in EPIC's auditory processor. The effective SNR used in content detection functions was simply the loudness of the word segment compared to the total loudness power in the other streams present.

In this model, the content detection function is a gaussian function with the mean and standard deviation estimated separately for each content word segment type (call sign, color, digit) in each talker similarity condition.The call sign content detection function is used for the non-content segments [ready], [goto], and [now], but their content plays no role in the model's stream tracking or strategy, although their loudness and pitch will be used by the stream tracking.

The stream perception model in the EPIC auditory processor uses an averaging minimum-distance stream tracking algorithm, defined in a way that allows two, three, or four streams to be tracked. There is a tracker for each stream. Each tracker accumulates the mean pitch (in semitones) and mean loudness (dB) of the word segments that have already been assigned to that stream. The tracker predicts that the pitch and loudness of the next, or new, word segment will be the same as the current means. The stream perception mechanism then calculates the prediction

error between each stream tracker and each new word segment as the cartesian distance between the (pitch, loudness) points, with pitch weighted by a parameter λ, loudness by $(1 - \lambda)$, where λ lies between 0 and 1. The new word segments are then assigned to streams so as to minimize the total distance between all words and their assigned streams. The stream trackers are then updated to include their newly assigned word segments, and the resulting means used to predict the next segment.

The stream perception model included a noise component to take account of the fact that even in the easiest Different-Sex (DS) condition, the peak performance is only about 95% correct. After determining the minimum-distance assignment, the stream perception process compares the maximum and minimum total distance; if the difference is less than a threshold value θ, the assignment is chosen at random. If there are only two streams, this means that a different assignment would be chosen half the time. If this threshold check does not discard the minimum assignment, then with a small probability α the minimum assignment is discarded in favor of a different randomly chosen assignment - a sort of "jitter" in the process.

This stream tracking mechanism as described works for more than two streams, and was used in the three- and four-talker models mentioned below.

Model strategy

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 to the cognitive processor, which is running a strategy implemented in production rules.

The strategy used to fit the Replication 1a data was a *use-what-you-heard* Use-Masker strategy whose details are in Appendix 2 and can be summarized as follows:

During the processing of the utterance, if call sign content is detected, tag its stream as the Target or Masker stream accordingly. If not, infer the Target/Masker status from the other stream if its call sign content was detected. Then tag the Target/Masker status of each color and digit word, based on its stream. Then if there are color-digit pairs from the same stream, tag them as same-stream-pairs.

When it is time to choose a response, the following rules are used for both choosing the color and choosing the digit: 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.

Optimizing parameters in corpus-driven models

With the introduction of a corpus-driven stream tracker, a trial-by-trial stochastic element appears in the MATLAB versions of the models that precludes exact and deterministic expressions for characterizing performance. In lieu of exact expressions, any parameter optimizer must rely on the accuracy of a Monte Carlo simulation which, in the present case, was

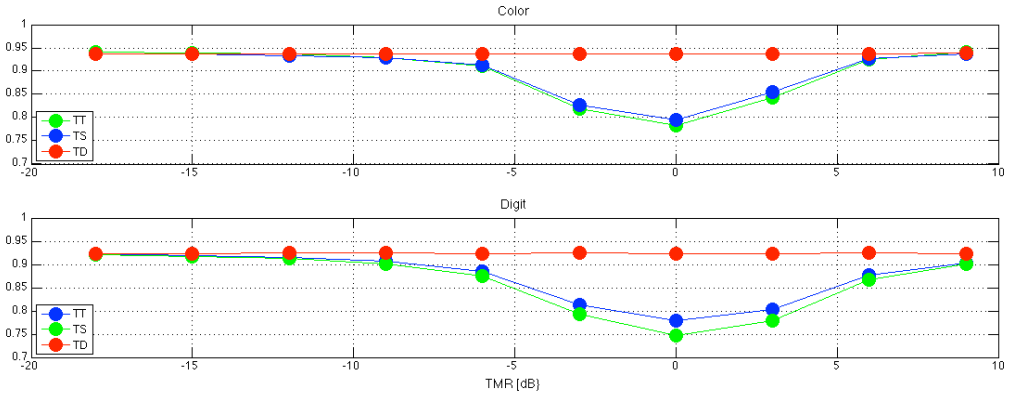

Figure 20*. Probabilities for staying in the same stream are shown for the simple stream tracker with parameter values of 0.75 (lambda, the mixture of loudness and pitch), 0.04 (alpha) and 0.35 (theta). Residual switching appears for the TD cases due to the choice of alpha. For TS and TD, color appears to be slightly more robust to switching than digit. Minima for all four cases occur at 0 dB TMR (SNR), the functions are symmetric about the minimum, and asymptote beyond absolute TMR values between 6 and 9 dB. In general, increasing the value of alpha lowers the asymptote, while increasing the value of theta broadens the switch function in the neighborhood of 0 dB TMR.*

found to require tens of thousands of trials per run. Thus, the new techniques was devised to improve the speed of the MATLAB model fitting. A two-phase method was developed in which a model of average stream behavior was used in place of the segment-by-segment and trial-by-trial outcomes from the stream tracker model (see Figure 20). A deterministic performance model was derived using this average stream tracking model as a component. This model, in turn, was used in optimizing the content detection parameters. Final refinement of these parameters was accomplished in the second phase of the method by manual adjustment based on Monte Carlo simulations. Overall, this approach aided in fitting the data, but left open the possibility that certain regions of the parameter space were not be fully explored due to the decoupling of the stream tracking and content detection subsystems. A more satisfactory computational approach would be simple grid search.

Results

This model applied to the Replication 1a experiment data produces very good fits with reasonable parameter values throughout, as shown in Figure 21, which shows the predictions from the EPIC implementation using 3000 trials per talker/SNR condition. The parameter values were first determined by search using the MatLab implementation of the model, and then verified in the EPIC implementation. The parameter values are listed in Table 1. In choosing these values, we attempted to use the same standard deviations for content detection functions, and same tracking process parameters for the different talker conditions, so that the effects of talker condition were isolated to the difference in content detection function means. Within these constraints, we could account for the data with a relatively small set of parameters. Thus the major effects in the data are being produced by the acoustic properties of the input as sampled from the segmented corpus.

As a summary measure of goodness of fit, $r^2 = .99$ between predicted and observed value for the Target and Masker color and digit probabilities (blue and red), and $r^2 = .96$ for the both-correct probabilities (black). These both are excellent results; only a few of the predicted values

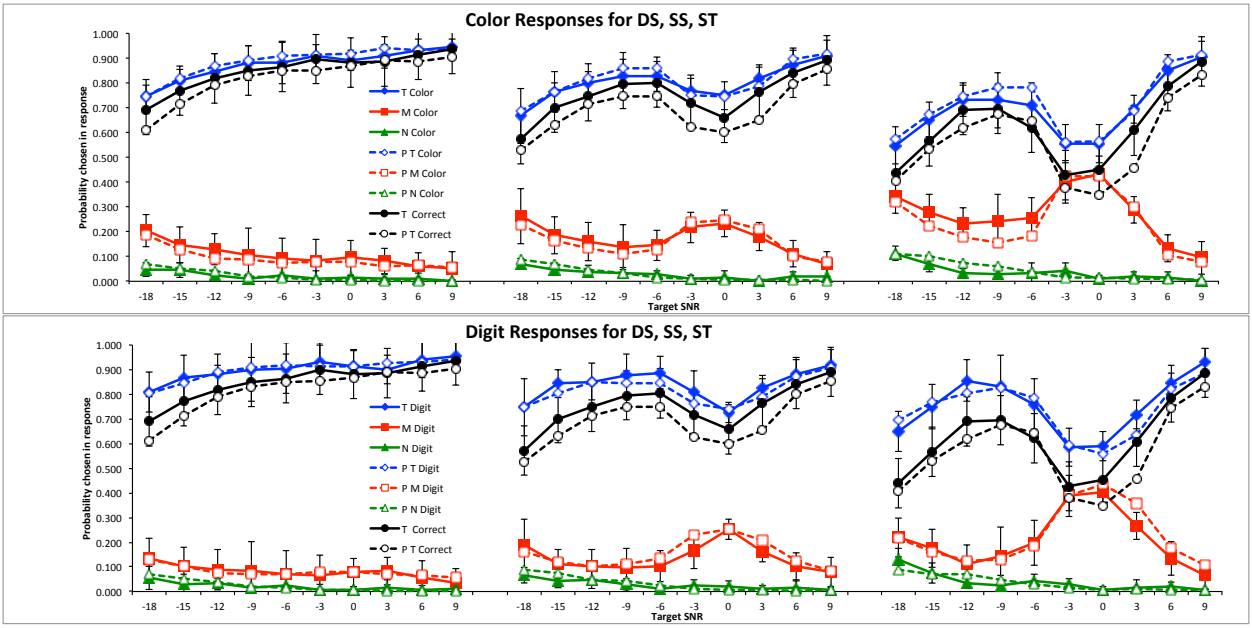

Figure 21*. Observed (solid points and lines) and Predicted (open points and dotted lines) probabilities of responses for the Replication 1a Experiment. Top panel shows color responses, bottom panel shows digit responses. First plot on the left is for Different Sex (DS) condition, middle is for Same Sex (SS) condition, and rightmost is for Same Talker (ST) condition. Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker. Plotted points are the means of each subject's proportion of responses; error bars are 95% confidence intervals for these means.*
NMA_crm15_V7aUse_P061813_r5.xlsx

Parameter	DS Condition	SS Condition	ST Condition
Call sign content detection	$\mu = -21, \sigma = 10$	$\mu = -20, \sigma = 10$	$\mu = -18, \sigma = 10$
Color content detection	$\mu = -26, \sigma = 10$	$\mu = -24, \sigma = 10$	$\mu = -20, \sigma = 10$
Digit content detection	$\mu = -31, \sigma = 10$	$\mu = -28, \sigma = 10$	$\mu = -26, \sigma = 10$
Stream tracking distance pitch weight	$\lambda = 0.75$	$\lambda = 0.75$	$\lambda = 0.75$
Tracking distance difference threshold	$\theta = 0.35$	$\theta = 0.35$	$\theta = 0.35$
"Jitter" probability	$\alpha = 0.04$	$\alpha = 0.05$	$\alpha = 0.05$

Table 1*. Parameter values used in fitting the model in Figure 21.*

are outside the confidence intervals in the data. However, there is a clear tendency for the both-correct points to be generally under-predicted; this problem will be discussed further below.

Since the stream-tracking model for Replication 1a fit the data very well, we decided to undertake a complete segmentation of the CRM corpus while continuing work with the "quick and dirty" six-beat segmentation described above. The complete segmentation and its application in models will be the a subject of a separate report. The models in this report all use the six-beat segmentation.

Scaling the Stream Tracking Model to Multiple Maskers

As summarized above, our first model approach, the stream ID detection model, failed to scale to three- and four-talker cases, where there is more than one masker message. We rejected the stream detection approach and developed the stream tracking model as a result. In considering how the stream tracking model should be applied to the multiple-masker case, certain issues arise. First, as mentioned above, performance in the two-talker case is much better than in the three- and four-talker case under "equivalent" measures of SNR (Brungart, Simpson, Ericson, & Scott (2001)). There are two possible explanations for this *multiple-talker deficit* with respect to the EPIC architecture and our models.

First, there is a difference in the inferences that can be made if some of the call sign content goes undetected. If there are only two talkers, and the call sign is detected for only one of them, then the Target/Masker status of the other can be strongly inferred. In the case of two or more maskers, inferences can still be made, but they are not as strong: For example, if the Target call sign is detected, then the other streams can be inferred to be Maskers. But if the Target call sign and one of the Masker call signs are undetected, then the status of those two streams is ambiguous. However, a model using the Table 1 parameters that takes this inferential difference into account and makes all of the possible inferences to deal with the ambiguity optimally, does not result in acceptable fits to the three- and four-talker data. Therefore, something else must be going on.

The second explanation is that the deficit is due not to issues of stream tracking, but to content detection becoming much less sensitive as the number of talkers increases. Our explanation for this effect is not some notion of limited processing resources, which is a common default explanation for any human limitation. Rather, the explanation is that it becomes impossible in signal-processing terms to separate out the individual formants of the competing words if the pitch/loudness are similar; the combined speech signal is ambiguous, *even for an ideal analyzer*. Thus the less sensitive and steeper content detection function is due to the formant structure of the competing vowels becoming very ambiguous when additional masking vowels are present, becoming in effect, like masking noise. This might be similar to the effects reported in the literature on speech reception thresholds in the presence of noise, where the interfering effect is less if the noise masker is modified to be speech-like in a way that might improve stream tracking (cf. Hopkins and Moore, 2009).

If this hypothesis is correct, it should be possible to fit the model to multiple-talker data by simply making the content detection functions less sensitive in the presence of multiple maskers. Since Brungart, Simpson, Ericson, & Scott (2001), reported only both-correct responses, our model predictions are limited to that metric. We modeled only the subset of the data corresponding to homogenous maskers: compared to the target, the maskers are either the different sex, same sex, or same talker. We used the very same stream tracking and task strategy as those used for the Figure 21 model. The only change was to the parameters of the content detection functions.

Results

Figure 22 shows the fit for both three- and four-talker data. This fit was obtained by adding 18 dB to the mean of the detection functions from the Figure 21 model, and decreasing the standard deviations from 10 to 3 dB. This single uniform change to the detection functions suffices to account fairly well for the multiple masker data; no other alterations in either the stream tracker or the production rules is necessary.

The fit of this model clearly needs further work, but compared to the total failure of the stream ID model to scale to multiple maskers, this is an extremely encouraging result.

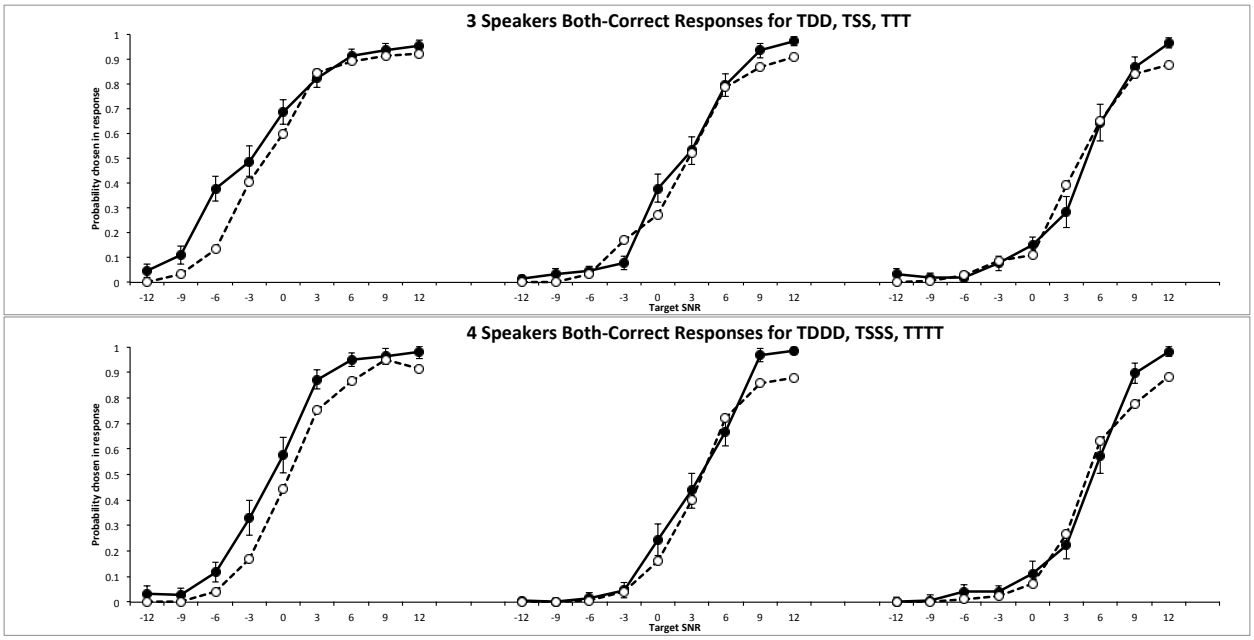

Figure 22*. Observed (solid points and lines) and Predicted (open points, dotted lines) probabilities of a both-correct response for a three-talker task (top panel). The talker conditions left-to-right are two maskers of different sex than the target (TDD), two maskers of the same sex as the target (TSS), and two maskers who are the same talker as the target (TTT). The bottom panel shows the corresponding results for a four-talker task.*
3S_NMA_crm15_V7aUse_P061813Add18SD3_r1.xlsx
4S_NMA_crm15_V7aUse_P061813Add18SD3_r2.xlsx

Can We Account for the Original Brungart Data?

What strategy are subjects using?

The success of our current stream tracking model in accounting for the Replication 1a results and the three- and four-talker results is encouraging. but what about the original Brungart (2001) results? The fact that response-correctness feedback makes such a difference in the effects strongly suggests that the effects in the original data should be mainly a matter of the task strategy; the same perceptual and stream tracking mechanisms should hold.

The main strategy issue is that in the original study, while subjects were told to report the color and digit from the message with the target call sign, they were not told how to proceed if they didn't hear the target call sign or were unsure about which color or digit word went with it - there was no defined "fallback" for them to follow, and they were not allowed to say "I don't know the correct answer." Because they were given no feedback about the correctness of their response, they had no defined basis for inferring a strategy that worked well when the message content was unclear. This, presumably, left subjects free to develop an otherwise outlandish or idiosyncratic strategy, and quite likely a mixture of strategies, as suggested by the abstract stream-switching model we earlier developed for this data.

Outlandish strategies fit the data

We experimented with task strategies for the Brungart (2001) data, which were outlandish compared to the more reasonable strategies developed so far, in that they used the *loudness* of the content words, not just the content, in deciding how to respond, and furthermore, made decisions differently for color content than for digit content. We developed two related strategies.

The first strategy, Mixture Model 1, can be summarized from the subject's point of view as follows: *"When the genders are different, it's easy. But when they are the same, I can hear the digits fine, but the colors are a mess. You didn't tell me what to do here. So if I can't tell which goes with the target, I'll flip a coin to decide to go with the louder or just guess."* When translated into production rules, the strategy is rather convoluted; it is not at all optimal, treats digits and colors very differently, and mixes avoid-maskers and use-what-you-heard response rules.

The rules can be summarized as follows: Tag a stream as Target or Masker if that call sign content was detected, but infer only the Target stream from the Masker streams, not vice-versa. If digit content from the Target stream is available, use it in the response; if not, pick any perceived digit content at random; if no digit content is available, make a pure guess of the digit. For the choice of color, a mixture of strategies is used based on the perceived gender of the talkers: If the genders are different, and color content from the Target stream is available, pick it for the response. If the genders are the same, then with probability 0.5 decide to use *the loudness of the color word* in choosing the response. If the loudness is not being used, then color content from the Target stream is selected if available. If the loudness is used, then select the louder color word if it is not tagged as from the Masker stream; if it is from the Masker, pick any available color content at random; if no color content is available, make a pure guess of the color.

This strategy was used with the exact same parameters and stream tracking model as the model for the Replication 1a Experiment data shown in Figure 21 above; only the task strategy (production rules) was changed. The resulting predicted and observed values are shown in Figure 23. While the fit is imperfect in terms of many predicted values lying outside the confidence intervals, the summary measure of goodness of fit is not bad at all, being the same as in Figure 21, namely r^2 = .99 between predicted and observed value for the Target and Masker color and digit probabilities (blue and red), and r^2 = .96 for the both-correct probabilities (black). The basic problem with the model is that the both-correct (black) curves are generally under-predicted.

Mixture Model 2 explores how the same-stream-pair could be given priority in a generally similar model that falls back on loudness in a mixture strategy. It can be summarized from the subject's point of view as follows: *"Sometimes I can hear both digit and color loud and clear and other times I might hear the digit, but the colors are a mess. I'll try to listen for the target stream as best I can, but, if stumped, I'll flip a coin to decide to go with the louder pair, or the louder color, or just guess."*

The specifics of this strategy are also rather convoluted, but can be summarized as follows: There are two levels of using loudness in the strategy. First flip a coin to decide to use loudness for both color and digit; if so, choose the louder color and louder digit if the content is available.

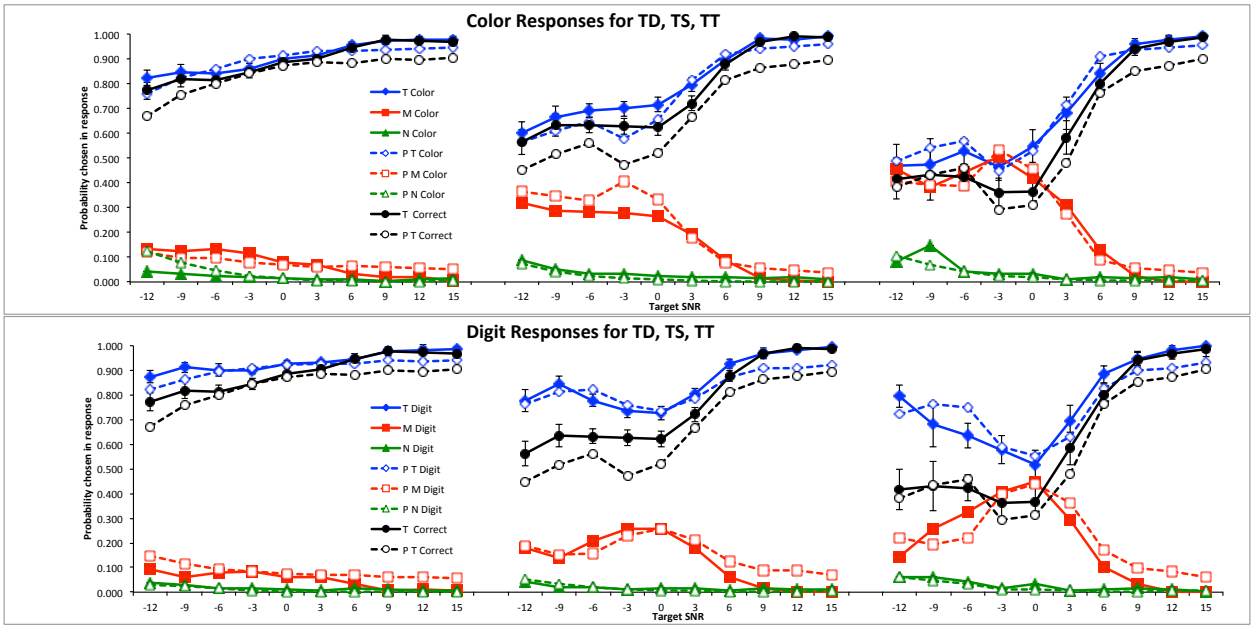

Figure 23. *Observed (solid points and lines) and Predicted (open points and dotted lines) probabilities of responses in the original Brungart (2001) data. Predicted points are from Mixture Model 1. Top panel shows color responses, bottom panel shows digit responses. First plot on the left is for Different Sex (TD) condition, middle is for Same Sex (TS) condition, and rightmost is for Same Talker (TT) condition. Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker.*
B2001NMA_crm15_V7a7Use_P061813_r4.xlsx

If not, flip a coin to decide whether to use loudness just for color. If so, choose the louder color if its content is available. The probability of using loudness was estimated for each condition. It is rare for TD (0.03, 0.06 for the two coin flips), more common in TS (0.10, 0.12), and most often in TT (0.12, 0.28). If not using loudness, use the color (or digit) tagged as from the Target stream. If no color (or digit) has been selected by the above rules, make a guess but do not use a color (or digit) that is tagged as being from the Masker stream.

The Mixture Model 2 strategy was also used with the exact same parameters and stream tracking model as the model for the Replication 1a Experiment data shown in Figure 21 above; only the task strategy was changed. The resulting predicted and observed values are shown in Figure 24. This fit is better in that the prediction of the both-correct (black) curves is visibly better than in Mixture Model 1. However, the summary goodness of fit is the same at $r^2 = .99$ between predicted and observed value for the Target and Masker color and digit probabilities (blue and red), and only slightly better at $r^2 = .97$ for the both-correct probabilities (black), and more parameters were estimated. But the fact that both-correct is better predicted does suggest that a strategy that gives preference to (color, digit) pairs from the same stream might fit both the Replication 1a Experiment data and the Brungart data better than our current independent-choice models.

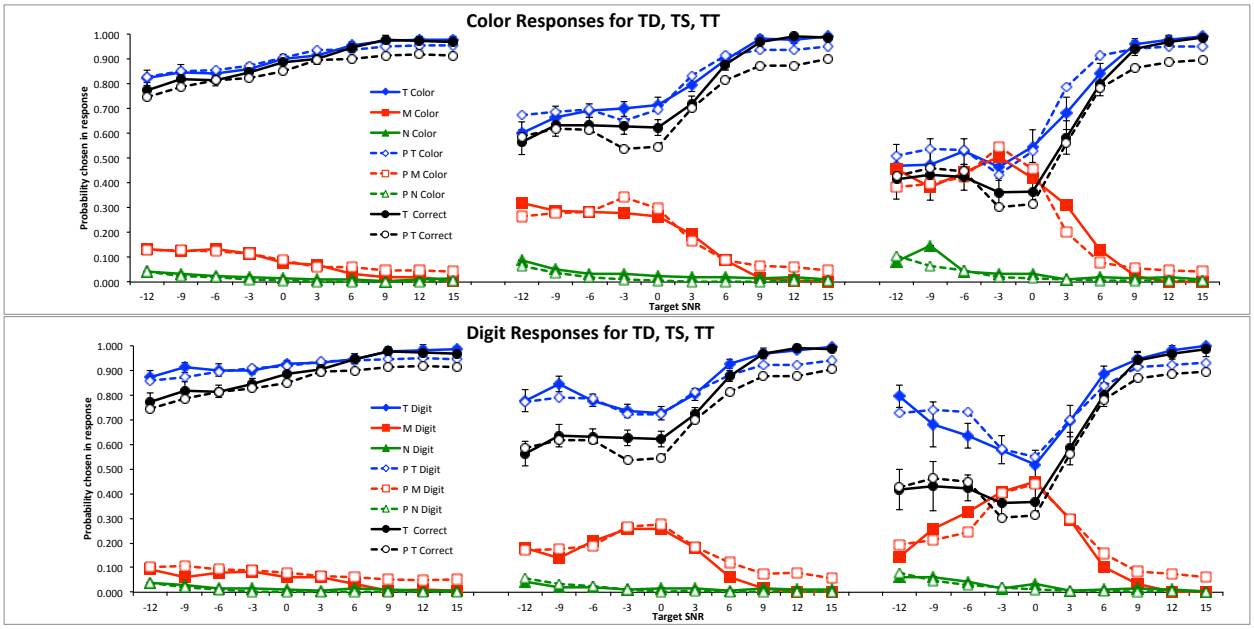

Figure 24*. Observed (solid points and lines) and Predicted (open points and dotted lines) probabilities of responses in the original Brungart (2001) data. Predicted points are from Mixture Model 2. Top panel shows color responses, bottom panel shows digit responses. First plot on the left is for Different Sex (TD) condition, middle is for Same Sex (TS) condition, and rightmost is for Same Talker (TT) condition. Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker..*
B2001NMA_crm15_V7a6Use_UWYH8o_P062213.xlsx

Additional Data Reveals Individual Strategies

The models using the Replication 1a data set were described in Wakefield, Kieras, Thompson, Iyer, and Simpson (2014). Shortly thereafter, AFRL sent us a larger version of that dataset that included 8 additional subjects for a total of 18; this is labeled *Replication 1b* here. But to our horror, in this supposedly larger and more stable dataset, there were many more Masker intrusions at low (negative) SNR, especially in the Same-Talker condition, than in the earlier Replication 1a dataset.

See Figure 25, which follows the same color-codes and conventions as the previous graphs, but the 1a data is shown as solid points and lines, and the 1b data by the open points and dotted lines. The 1b performance is generally lower than 1a, and there are also many more Masker responses.

Our Replication 1a model did not fit the Replication 1b data very well - the model wasn't making masker responses often enough. This led us to wonder why the larger 1b dataset included so many more Masker intrusions than 1a. This was always something of a puzzle: Many Masker choices would be expected around 0 SNR because there the two messages are most difficult to distinguish in SS and ST. But at very negative SNR, the Masker choices increase even though the Masker message should be "loud and clear" - there should be no doubt in the subject's mind that this is the Masker message, so why are they choosing to respond with the Masker so frequently?

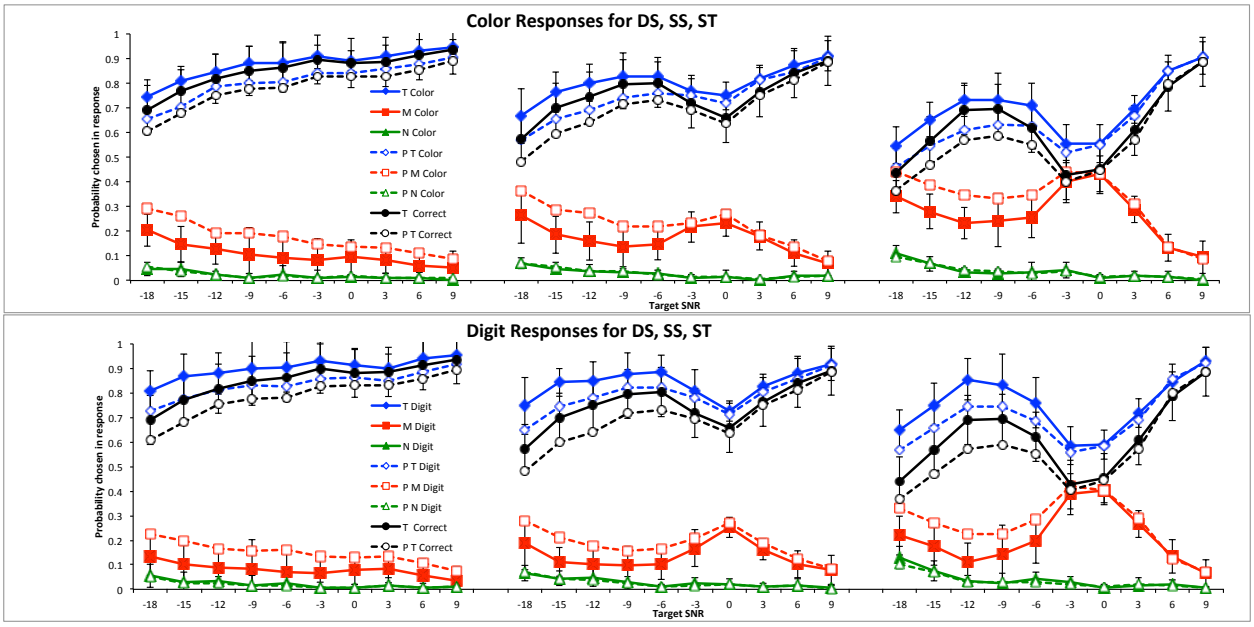

Figure 25. *The solid points and lines are the Replication 1a data used in the model fit shown in Figure 2. The Replication 1b data are shown with open points and dotted lines. Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker. Note that despite the larger sample size, overall performance (blue and black curves) in Replication 1b is substantially lower than in 1a, and Masker responses (red curves) are chosen far more often in Replication 1b.*

Strong effects of individual differences in strategies

We performed a preliminary analysis of the Replication 1b data at the level of individual subjects that suggested that some subjects were choosing Maskers at negative SNR much more often than others. We followed up with a full analysis of individual subjects, classifying them by the average proportion of Maskers they chose for trials with SNR ≦ -6 (this gives a total of 30 proportions per subject to average together). Figure 26 shows that three groups are apparent: those choosing Maskers less than 20% of the time; those choosing maskers between 20% and 50% of the time, and a small group who chose maskers more than 60% of the time!

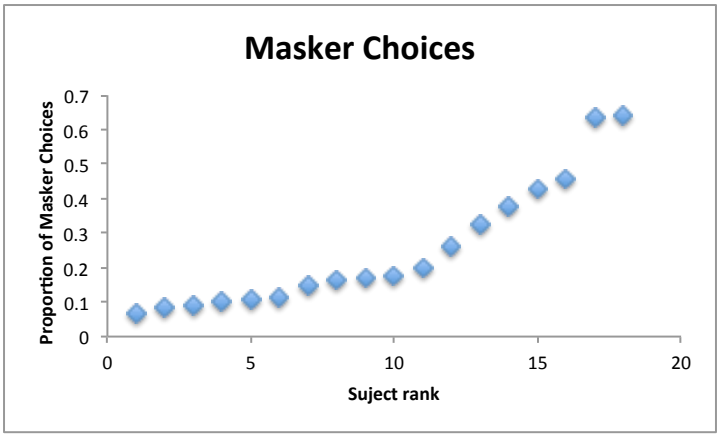

Figure 26. *Subjects ranked by the average proportion of Masker choices they made for SNR conditions ≦ -6. Some subjects chose Maskers extremely often!*

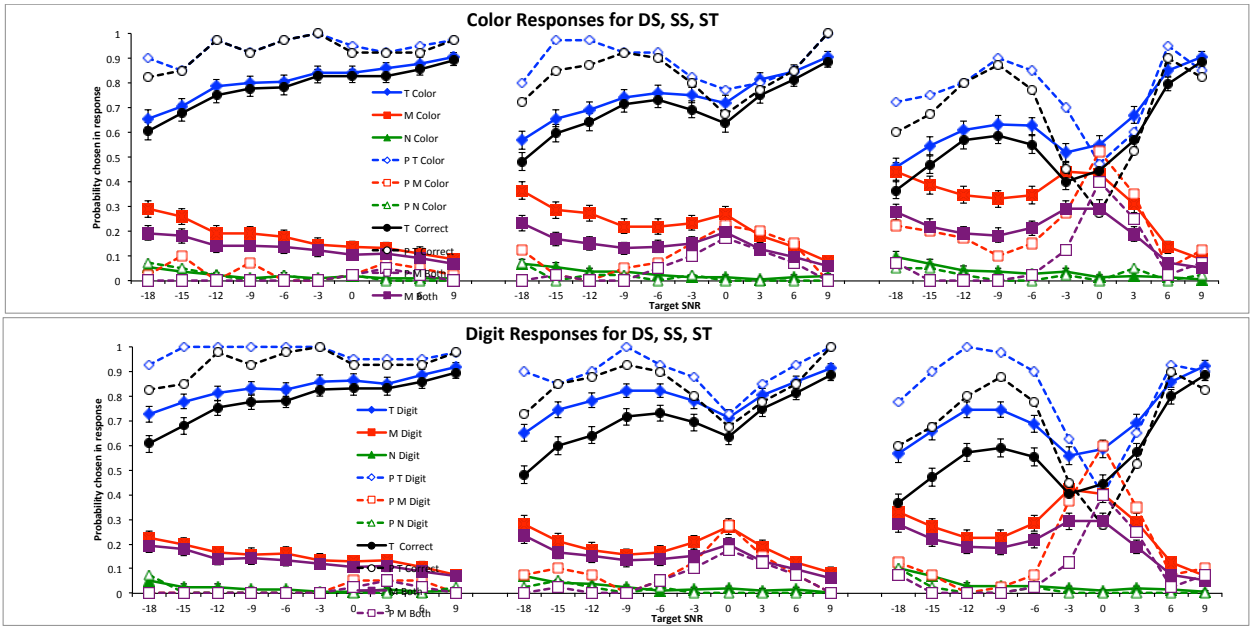

Figure 27. *Solid points and lines show the Replication 1b aggregate data for the whole group of 18 subjects (mean proportion of choices in each condition). Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Purple curves are Both-Masker responses (both color and digit from the Masker message, the opposite of Both-Correct). Green curves are for neither Target nor Masker. Open points and dotted lines show the data for an individual subject who had a very low rate of Masker choices at SNR ≦ -6. This subject seems to be following an optimal Avoid-Maskers strategy, but note also the very high correct performance of this subject compared to the group mean.*

A few example subjects make clear what is happening. Figure 27 shows a subject whose Masker choice rate was only 6.4%. The plot follows the same color coding conventions as the earlier graphs, with the solid points and lines with confidence intervals showing the group aggregate data, but the open points and dotted lines show individual subject proportions for each response choice. The purple curves are new in these plots; they show the proportion of *Both-Masker* responses, where the subject chose both color and digit from the Masker message. Notice how this subject, in general, produces far fewer Both-Masker responses than the group average, clearly following something like an Avoid-Maskers strategy - if the Masker message is loud and clear compared to the Target, this subject simply does not choose it. In addition, notice how this subject's performance on the Target and Both-Correct responses is much higher than average. It is possible that his or her perceptual apparatus is much better than average at content detection or stream tracking.

In contrast, consider the subject whose performance is shown in Figure 28, and who chose Masker responses at negative SNR an average of 63.8% of the time. It is striking how this subject's Masker choices (red and purple curves) are essentially a mirror-image of their Target choices (blue and black curves), and so their data resembles the group average only at positive SNR. At negative SNR, it is as if they are doing a entirely different task from the group altogether - perhaps something like "report the loudest message you hear" rather than trying to consistently identify the Target and reject the Masker. This type of strategy is not one that we

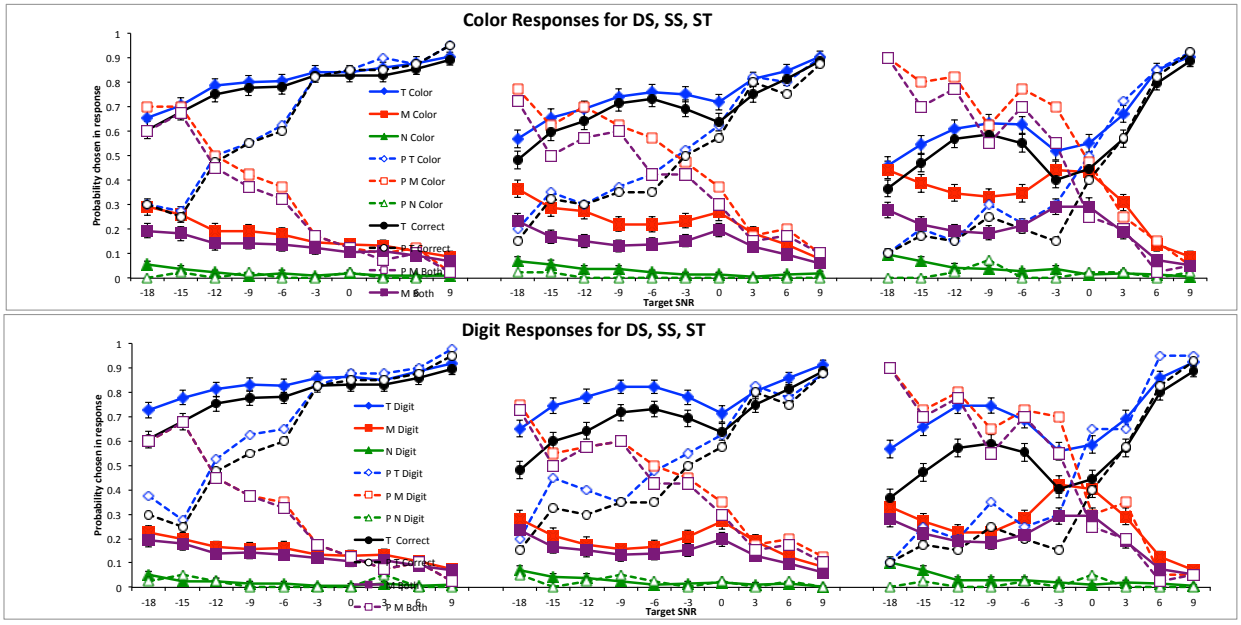

Figure 28. *Solid points and lines show the Replication 1b aggregate data for the whole group of 18 subjects (mean proportion of choices in each condition). Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Purple curves are Both-Masker responses (both color and digit from the Masker message, the opposite of Both-Correct). Green curves are for neither Target nor Masker. Open points and dotted lines show the data for an individual subject who had a very high rate of Masker choices at SNR ≦ -6. Note the mirror-image relationship between Target responses (blue and black) and Masker responses (red and purple). This subject seems to be choosing the loudest response rather than rejecting the Masker, a completely different strategy than the ones used in the model, or intended in the task instructions.*

tried to use in our models, because it is completely contrary to what the task is supposed to be - which poses the problem of why some subjects behaved this way. Also, note that such strategies are possible in our models - in fact our last models for the Brungart (2001) data (see Figures 23 and 24) would "report the loudest" under some conditions, which we now see might actually have been the case by some of the subjects.

As a final example, Figure 29 shows the results for a subject who chose Masker responses 19.9% of the time. Their data resembles the group average fairly well, showing an increase in Masker responses as SNR becomes negative, resembling the range that our Use-Maskers strategy tends to cover.

Using the individual subject results, we were able to determine that the subjects added to the Replication 1a dataset to produce the 1b dataset were mostly "heavy Masker users", which resulted in a far higher level of Masker responses in the 1b aggregate data than in the 1a data. Thus, these results make two points: (1) Individual subjects are differing not just in basic perceptual parameters (e.g. sensitivity of content detection)which might be expected, but are following radically different strategies. (2) Both the original instructions and the additional feedback in Replication 1 did not produce consistent subject strategies. In fact, the original and Replication 1 instructions simply told subjects to "ignore" the Masker message, but the procedure *required* subjects to make a response, even if they hadn't heard the Target message. Thus the instructions are ambiguous concerning how they should respond in this case. Clearly, at least some subjects think it is appropriate to "ignore" the Masker message by choosing it when they didn't know the Target content.

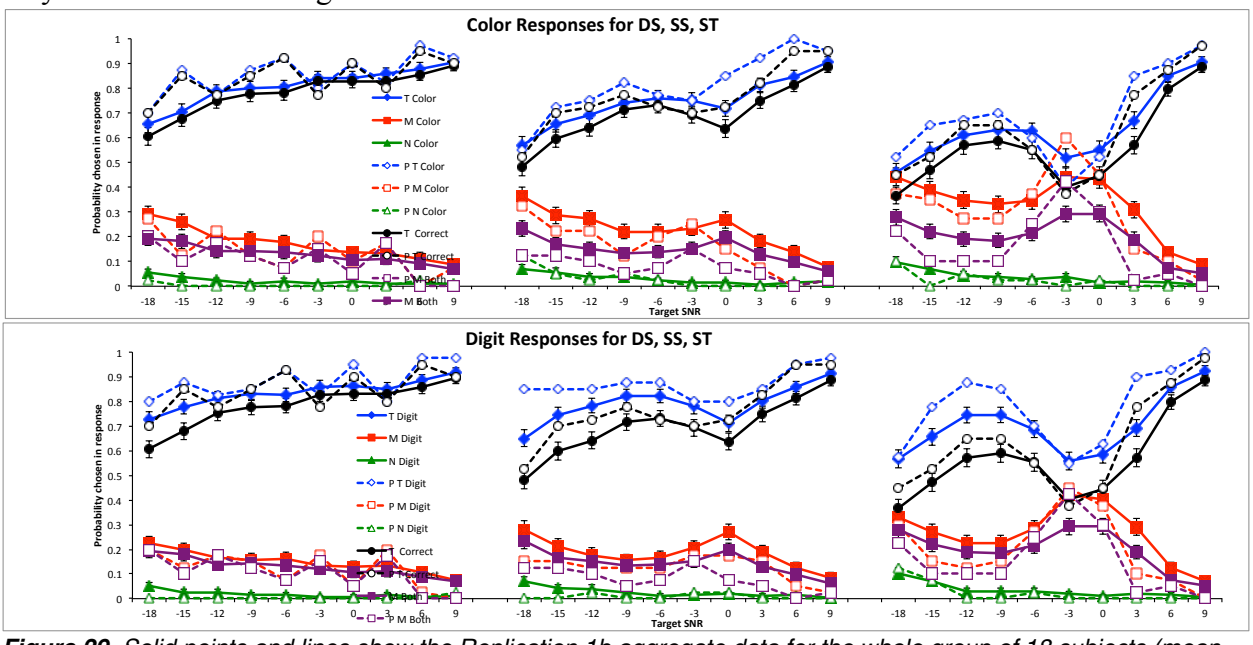

Figure 29. *Solid points and lines show the Replication 1b aggregate data for the whole group of 18 subjects (mean proportion of choices in each condition). Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Purple curves are Both-Masker responses (both color and digit from the Masker message, the opposite of Both-Correct). Green curves are for neither Target nor Masker. Open points and dotted lines show the data for an individual subject who had an average rate of Masker choices at SNR ≦ -6. This subject seems to be following a form of Use-Maskers strategy.*

Most of the subjects appear to be using strategies that are within the range we had explored (e.g. Avoid-Maskers vs Use-Maskers) or that we could implement. We could fit models with different strategies to individuals or subsets of the subjects reasonably well. However, the fact that the experimental methodology is producing less uniform behavior than it could is a clear problem.

A New Replication with Performance Incentives for Strategy Control

Using payoffs to stabilize strategy choices

After discussion of the individual subject data, the AFRL group collected a new dataset, called Replication 2 here, in which they followed some "best practices" in human performance experimentation to reinforce the task instructions with an explicit payoff scheme. Specifically, on each trial the subjects were awarded +1 point each for a correct (Target) color or digit response, -2 point each for a Masker color or digit response, and 0 points for a color or digit response that was Neither Target nor Masker. Thus on each trial they could make up to 2 points, or lose as many as 4 points from their cumulative total; after each trial they were shown their cumulative points with comparison to other subjects and were given a bonus if they achieved 1200 points or more by the end of the experiment. This payoff scheme had the effect of making explicit what the instructions intended; namely that if they did not know the Target color or digit, they should not choose something that they know is the Masker, but choose something else as a guess instead.

AFRL provided the Replication 2 data set; it is shown in Figure 30. The rate of Masker choices in the Same-Sex and Same-Talker conditions is much lower than in either the original Brungart (2001) or the Replication 1b dataset; all of the 10 subjects chose Masker responses at SNR ≦ -6 less than 18% of the time, which is much better than in Replication 1b. However, notice how there remains a tendency for subjects to choose the Masker response more often at the lowest SNRs - this is especially pronounced in the Same-Talker color responses. Nevertheless, compared to the original Brungart (2001), and Replication 1b data, the frequency of Masker responses is greatly reduced. More importantly, analyzing the data by individual subjects shows similar patterns of effects for almost all of the subjects, a pattern consistent with a Use-Maskers strategy. This means this data should be easier to interpret and fit with a single set of parameters and strategy.

Models for Replication 2

We applied our current stream tracking model to these data, readjusting the parameter values to maximize the fit to the Target, Masker, and Neither Color/Digit observed data, and we compared the Avoid-Maskers and Use-Maskers strategy to see if the Replication 2 dataset actually resulted in an strategy to not respond with Maskers. The strategies used for the comparison are fully described in Appendix 2.

Figure 31 shows the fit for the Avoid-Maskers model. In terms of overall goodness of fit metrics, the fit for the Target/Masker color/digit data (Blue and Red points) is very good: r^2 = 0.98; the fit for the Both-Correct data (Black curves) is not as good, but still impressive: r^2 = 0.95. However, it is clear that the strategy is wrong; notice how far too many Neither responses

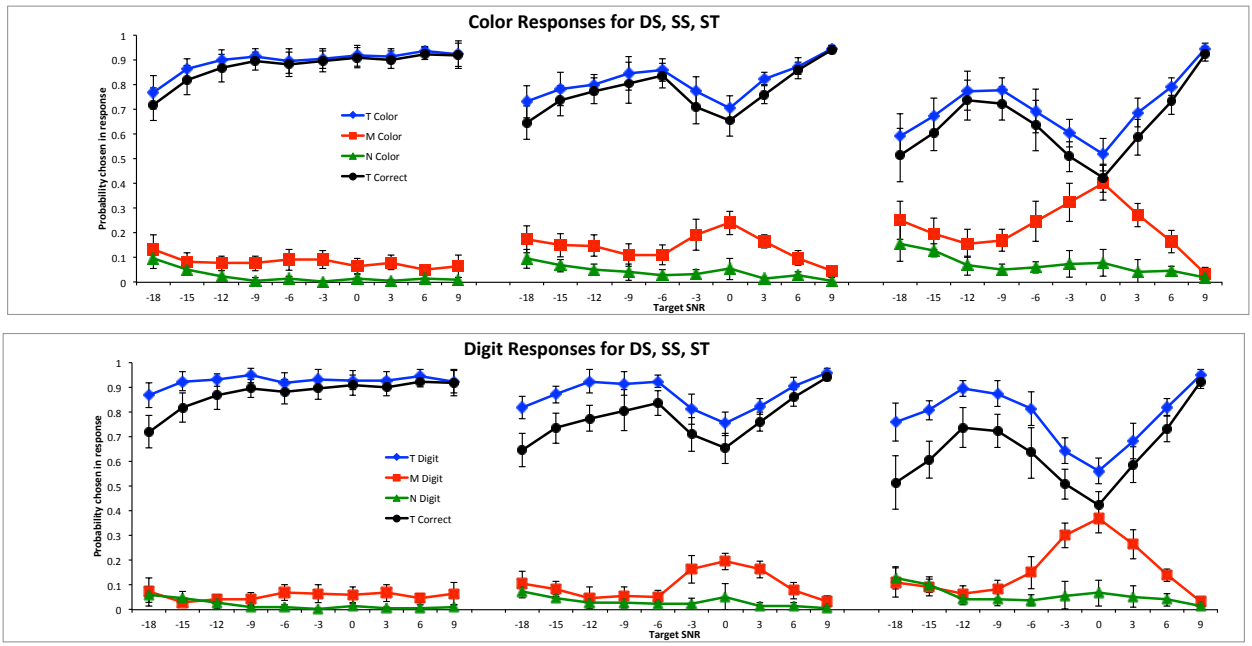

Figure 30. *The Replication 2 data set aggregate. Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker. Each data point is the mean of the proportions produced by individual subjects, and the error bars are the 95% confidence intervals for these means. Although Maskers are chosen less often than in previous experiments, they are still being chosen somewhat more often as SNR becomes very negative.*

(green curves) and far too few Maskers are predicted at negative SNRs, especially in SS and ST, especially for color responses. So this model is clearly wrong, even if the fit is nominally good by the usual metrics.

Figure 32 shows the fit for our standard Use-Maskers strategy (the same as described above and in Appendix 2) in which known Masker content is used in the response if Target content is not available and the Target stream identity was inferred rather than its call sign content perceived. The parameter values are shown in Table 2. In this case we did not try to constrain the parameters to have the same values across talker conditions. This fit is slightly better, $r^2 = 0.99$, for the Target/Masker Color/Digit performance, and the Both-Correct is slightly worse at $r^2 = 0.94$. But all of the trends are correctly captured, and almost all of the predicted points are within the confidence intervals of the observed data. Thus it seems clear that subjects are still following a strategy of responding with a known Masker under certain conditions.

Why are subjects still choosing maskers?

There may be some residual failure to follow the instructions, even as reinforced with the payoff scheme, but perhaps there is a genuine perceptual problem underlying the Masker choices, now that we can be more confident that most subjects are trying to perform the intended task. A key point is that these studies are done in a mixed-block design; on each trial the stimulus can be any one of the 30 combinations of talker condition × SNR level. So if the subject did not perceive the call signs, he or she has no way of knowing whether the louder message is the Target or the Masker. The fact that this is most pronounced in Same-Talker Color, where the two

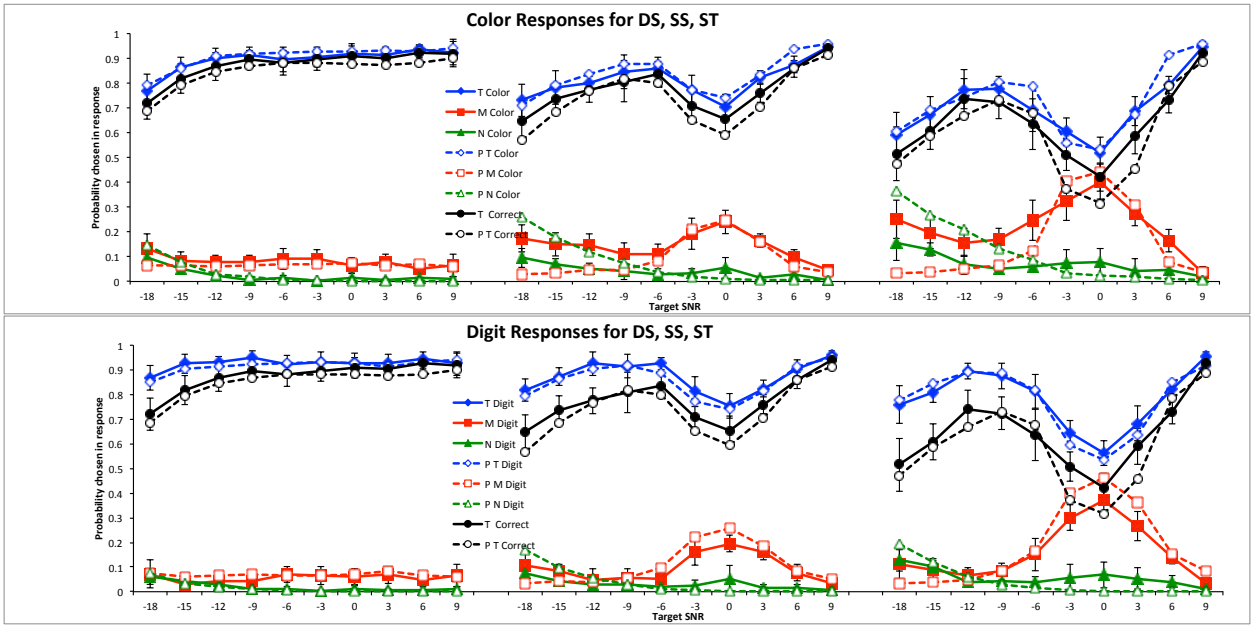

Figure 31. *Fit of the stream tracking model with Avoid-Maskers strategy to the Replication 2 aggregate data. Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker. Note the misfit especially in the SS and ST conditions at negative SNR where the model is producing too many Neither responses and too few Masker responses.*
Jun02RepAll_V7aAvoid_Trk43ab_Fitv23ab

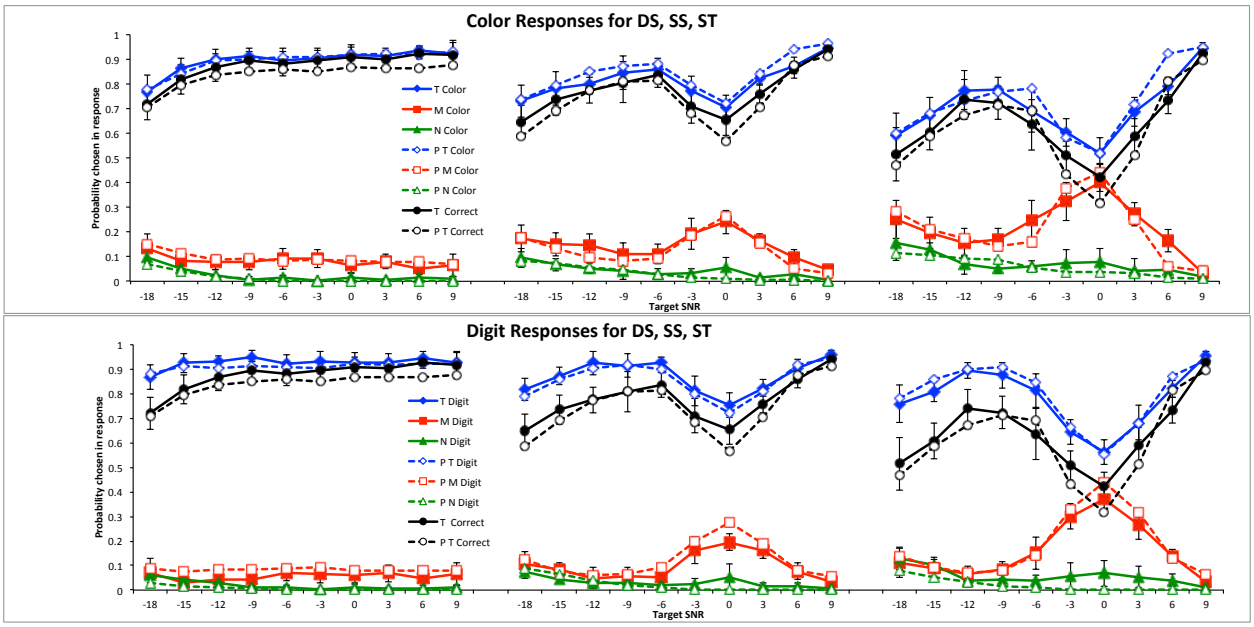

Figure 32. *Fit of the stream tracking model with Use-Maskers strategy to the Replication 2 aggregate data. Reading from top to bottom in each graph, blue curves are for Target responses. Black curves are for both color and digit being from the Target, and are the same in the top and bottom panels. Red curves are for Masker responses. Green curves are for neither Target nor Masker. Except for a tendency to under-predict the Both-Correct responses (black curves), the fit is excellent.*
Jun02RepAll_V7aUse_UWYH_Trk43ab_Fitv23ab_r2

Parameter	DS	SS	ST
Callsign content detection μ, σ	-21, 6	-22, 15	-19, 15
Color content detection μ, σ	-24, 6	-25, 12	-20, 13
Digit content detection μ, σ	-30, 6	-25, 6	-25, 6
Stream tracking pitch weight λ	0.60	0.60	0.65
Stream tracking threshold θ	0.45	0.45	0.55
Stream tracking "jitter" α	0.05	0.02	0.02

***Table 2**. Parameter values for the fit shown in Figure 32.*

messages are clearly hardest to distinguish, suggests that it might not be a strategy failure, but rather that the streams can be hard to distinguish in these conditions, and the masking interference on the content of *both* messages might still occur even at very negative SNRs. When we look at individual subject performance, the best-performing subjects rarely choose Maskers at SNR ≦ -6 (less than 5% of the time), but most subjects display some tendency to do so (10% - 17% of the time), especially in the Same-Talker condition with the Color responses.

How well does the stream tracker work?

The quality of this data and the model fit encouraged us to consider more carefully something that had been apparent earlier and is clear in Figure 32: the Both-Correct (black) data points are being consistently under-predicted even with the Color/Digit responses are being very closely predicted. The discrepancy is not large, but the consistency of the under-prediction is puzzling. Furthermore, it could not be removed by parameter adjustments, because color and digit are already being closely predicted, nor by modifications to the model strategy. That is, there is no way to additionally favor responding with *both* target color and target digit because the model strategy is already using all of the information available about the target content.

Rather than a strategy problem, the under-prediction of Both-Correct responses could be due to the stream tracking. As discussed above, our simple stream tracker predicts a general effect that we can see in both Replication 1 and Replication 2 datasets: If a subject has chosen the Target color, they are far more likely to choose a Target digit - they tend to "stick" with the Target stream. This effect is asymmetrical - if they choose the Masker color, their choice of digit shows that they are roughly equally likely to "stick" with the Masker stream or "switch" to the Target stream (see above). This asymmetry would be expected if the stream tracking mechanism is working well; if it has correctly identified the Target stream by the time of the color word, then it is more likely to stick with it for the digit word.

The quality of stream tracking can be examined by considering the conditional probability of choosing the Target digit given that the Target color was selected. Figure 33 shows this predicted and observed conditional probability of choosing the Target, Masker, or Neither digit given that the Target color was chosen, as a function of talker similarity condition and SNR. The blue

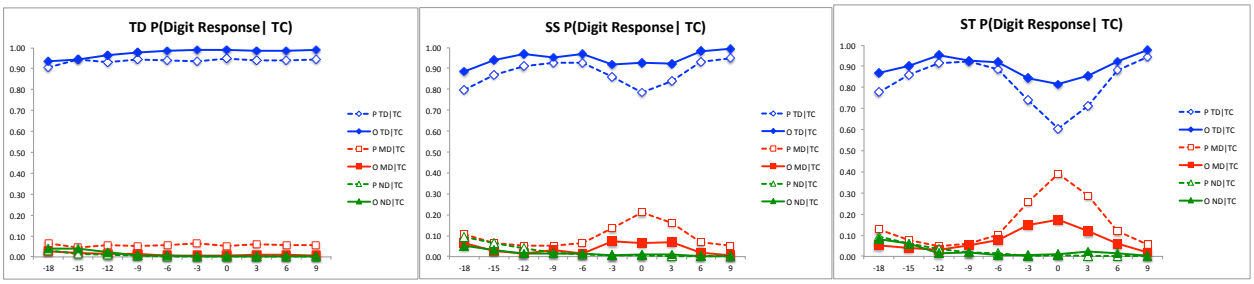

***Figure 33**. The observed and predicted conditional probabilities of choosing the Target, Masker, or Neither digit given that the Target color was chosen. Observed are solid points and lines; predicted are open points and dotted lines.Top most blue curves are for Target digit, middle red curves are for Masker digit, and bottom green curves are for Neither digit. The Target curve being much higher than the others means that if the correct stream was identified for the color response, it is very likely to lead to a correct identification for the digit response. Note that the predicted conditional probabilities are consistently too low for the Target, and consistently too high for the Masker.*
Jun02RepAll_V7aUse_UWYH_Trk43ab_Fitv23ab_r2

curves are the probability of choosing the Target digit; the red curves are the probability of choosing the Masker digit, and the green curves are for choosing a digit that is Neither Target nor Masker, all conditional on having chosen the Target color. Perfect tracking and reliable content detection would result in the blue curve being at 1.0. In the data, the blue Target digit curves are very high; the only place they approach 0.5 is in the Same-Talker case in the vicinity of SNR 0, where the streams are very hard to distinguish - this is due to the asymmetry noted above. But the key result is that the predicted conditional probability for Target digit are always low compared to the data, and the predicted conditional probability for Masker digit is always high compared to the data. This means that our stream tracking model is not as "sticky" as it should be - it is not following the streams as well as humans do, and that is why the Both-Corrects are under-predicted even though the separate Target color/digit predictions are very close.

Why is performance on colors worse than digits?

Another persistent issue in the data is that performance on the color words is inferior relative to the digit words. Right from the beginning of this project, it was clear in the original Brungart (2001) data that color words were different somehow from digit words, because color and digit responses produced very different patterns in the data. With the more stable Replication 1 and 2 data, colors and digits have similar trends in the data, but now we can see clearly that subjects choose Target color at lower rate, and Masker color at a higher rate, than they did for digit.

Because a simple explanation in terms of the relative number of colors and digits and guessing strategies predicts the opposite effect, the colors must simply be less distinguishable than digits. One explanation might be in terms of onset time differences, as mentioned above. Preliminary results from the complete corpus segmentation are that color onsets show relatively little variation compared to digit onsets, meaning that color words in the two messages are more likely to overlap in time than digits, which is consistent with the notion that colors would be more likely to mask each other, and thus be less distinguishable than would digits.

Onset time effects alone, however, may not fully account for degraded performance in recognizing the target color. We did some preliminary work with computing the confusion matrices for colors and digits, and confirmed an observation made by the AFRL group that

"Red" and "White" are often confused, while the digits seem quite distinguishable. We gathered results from a number of studies concerning "vowel capture" phenomena in which there is an asymmetry in masking effects of vowels. This research suggests that the color-digit effect may be related to the acoustics of vowel spaces. That is, in terms of the formant structure, the vowels in red and white are actually fairly similar, while those for blue and green are quite different from each other and with both red and white. This might be why we see the confusion matrix results for color. If the digits lack a similar source of confusion, it might help explain the color-digit difference.

Conclusions

1. The auditory architecture and the model for two-talker listening now include explicit mechanisms for auditory stream perception and tracking. These mechanisms rely on the acoustic properties of the speech input itself, with only a few parameters for the stream tracking mechanism itself.

2. Unlike the stream ID detection model, the stream-tracking model shows promise in scaling from the two-talker task to three- and four-talker tasks.

3. The proposed stream-tracking mechanism predicts an asymmetry in the stream "switch" behavior that was verified in the Replication 1a and 2 data. However, the current stream tracking mechanism does not track the streams as well as subjects do.

4. The model for the two-talker task has been verified with empirical data of very high quality (Replication 2). The difficulties with modeling the Brungart (2001) dataset were problems of unconstrained strategy effects in the data rather than a problem in the model itself - an instructive case of the data being "wrong" rather than the model!

5. On the more negative side, we have learned to be more cautious of the published data in this field - we will have to be especially careful to consider whether the task strategy was sufficiently controlled before using the data. In mitigation of this somewhat grim conclusion, it is fair to point out that the only reason why we were able to eventually identify this issue was that the Brungart (2001) study reported the Masker and Neither responses in addition to the correct responses, and this is also why we chose to model those results. Future studies in this paradigm should exert better control over subject strategies along the lines of those used in Replication 2.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y . (2004). An integrated theory of the mind. *Psychological Review* 111, (4). 1036-1060.

Brungart, D.S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am*. 109 (3), 1101-1109.

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 talkers. *J. Acoust. Soc. Am*. 110 (3), 1101-1109.

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

Kieras, D. E. (2004). The EPIC Architecture: Principles of Operation. Retrieved from http://www.eecs.umich.edu/~kieras/docs/EPIC/EPICPrinOp.pdf

Kieras, D.E., Ballas, J.A., & Meyer, D.E. Computational models for the effects of localized sound cuing in a complex dual task. (EPIC Tech. Rep. No. 13, TR-01/ONR-EPIC-13). Ann Arbor, University of Michigan, Electrical Engineering and Computer Science Department. January 31, 2001.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. Human-Computer Interaction, 12(4), 391-438.

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

Laird, J. (2012). *The Soar Cognitive Architecture*. Cambridge, MA: MIT Press.

Meyer, D. E., & Kieras, D. E. (1997a). 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. (1997b). A computational theory of executive control processes and human multiple-task performance: Part 2. Accounts of Psychological Refractory-Period Phenomena. *Psychological Review*. 104, 749-791.

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*.(pp. 15-88) Cambridge, MA: M.I.T. Press.

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.

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.

Appendix 1

StreamID Detection Model Strategy

(Brungart2talkerV9bSubOptPairX.prs)

At the top level, the strategy is to first listen to the messages, and then choose and make a response, and then repeat for the next trial.

Listening to the messages

Overview. The strategy assumes that the messages have a fixed structure, and it thus essentially "counts" words to know what message role each word plays. For example, the strategy can ignore the content of the first word (it is always "ready"), but then the strategy expects the next word to be a call sign, and so tags the word object as a call sign word, but its content (e.g. baron) and StreamID may or may not be available due to masking. Then the strategy can ignore the content of the next two word objects ("go" and "to), but then tag the next one as a color word, whose content (e.g. red) and StreamID may or may not be available due to masking. The strategy knows that the next word object is a digit word, and will treat it similarly. The content of the final word ("now") can be ignored, but its appearance means that the messages are complete and the strategy can go on to the next phase.

During listening to the messages, the strategy adds tags to word objects and other symbols. For example, if a call sign word has content that matches the Target call sign (baron), the strategy will tag that call sign word as the Target call sign word, and if that word object has a StreamID that is available (didn't get masked), the strategy will also tag that StreamID as the Target stream. The strategy could then tell that later words with the same StreamID are from the Target stream.

The production rules are written to not add tags to items that already have the same tags; for brevity, this condition is not described below. In addition, because of how the EPIC production system rules work, there has to be many separate rules that match individual combinations of data. In the prose description that follows, redundancy and duplication has been eliminated.

Call sign processing

Tag "observed" stream identification. When the call sign words have appeared, the strategy waits for any content of these words to be recognized. If a call sign word's content is available and is the Target call sign or a Masker call sign, and the word's StreamID is available, the strategy tags that streamID as the Target stream or Masker stream accordingly, with the qualifier "observed" to indicate that the strategy actually had the call sign content and streamID available to fully identify the source stream, as opposed to inferring it (described next). However, it greatly simplifies the remainder of the strategy on many occasions if it can treat both observed and inferred stream identifications as the same. So at this point the strategy also tags these observed stream identifications with the "inferred" qualification.

Tag "inferred" stream identification. This operation may not succeed in tagging both source streams; for example, one of the StreamIDs or call sign contents might not be available due to masking. However, the strategy can often infer the missing information. For example, say the

StreamID of the target call sign has been correctly tagged because the call sign content was available, but the content of the other call sign word is not available, but its StreamID is. That StreamID must be the Masker stream. Several other cases are possible depending on which content is/is not available and which StreamID is/is not available. In summary, if the content and the Target/Masker call sign word category is available for one of the call sign words (and its stream ID may or may not be available), but the content and the Target/Masker call sign word category is not available for the other call sign word, and that other call sign word's stream ID is available, then tag that other stream as the other Target/Masker stream, qualified as "inferred."

Color word processing

Tag from StreamID. When the color words have appeared, the strategy tries to associate them with streams regardless of whether their content is available, based only on whichever StreamIDs are available for the color words and the observed or inferred status of the Target and Masker StreamIDs. If the Target/Masker StreamID is tagged as observed, and a color word has the same stream, then the strategy tags that color word as Target or Masker qualified with both "observed" and "inferred." But if the Target/Masker StreamID is known as inferred only, and a color word has the same stream, then tag that color word as Target or Masker qualified with "inferred."

Cross-infer from StreamID. Again, the previous step may not have resulted in tags for both color words. If so, the strategy tags words in a stream using information from the other stream: If one of the color words is tagged as inferred Target/Masker, then tag the other color word as the opposite source qualified with "inferred". But if one of the two streams has been known and tagged as Target/Masker inferred, but not the other, and the known stream is not associated with either Color word (because it was not available for the relevant word), but one of the Color words has a different StreamID available, then tag that Color word as an inferred Target/Masker color word, with the opposite stream from the known stream.

Digit word processing

When the digit words have appeared, the strategy performs the same tagging and inference steps on them as it does for color words, resulting in digit words with the corresponding Target or Masker tags.

Tag Pairs

Once all of the content words have appeared, the strategy marks pairs of (color word, digit word) that are from the same stream. This allows the strategy to implement an optimization that if no stream got tagged as being the target stream (either directly or inferred), then a color and a digit from the same stream is a better guess than a color and digit from different streams.

Tag pairs using StreamID. The strategy tags color-digit word pairs that have the same StreamID as being same-stream pairs. If one color word has a StreamID, and the other does not, and one digit word has a different StreamID, and the other has none, then tag the color word with a StreamID and the digit word without a StreamID as a Same_stream_pair, and the other color word and digit word as a Same_stream_pair.

Cross-infer pairs. If a Same_stream_pair has been tagged, and there is a different color word and different digit word, the strategy tags them as a Same_stream_pair, even if they don't both have a StreamID available.

Choosing and Making a Response

Tag content

The strategy first applies the stream tagging for word objects to tag the content of the color and digit words with their source streams: If a word is tagged as Target/Masker color/digit observed/inferred, and has available content, it tags that content with the same Target/Masker color/digit observed/inferred information. These tagged content items will be used to choose the actual response. Because content might be masked, it is possible that a word will be tagged with a stream but have no content available to use in the response. For example, a color word object might be tagged as Target observed, but if its content is not available, there will not be any color content tagged as Target observed.

Choose response information

At this point all possible information has been extracted from the input messages, so the strategy is use it to pick the best response. The strategy chooses which color or digit to use in the response, and which to avoid in case is it necessary to guess from the alternatives on the display. The following rules are applied:

Use target: If a color/digit is tagged as Target content inferred, tag it as to-be-used in the response.

Note: The following two rules apply to the Use-Masker and Avoid-Masker versions; only one of these rules can be enabled at a time. The two versions of this strategy are otherwise identical.

Avoid maskers: If Target Color/Digit content is missing, (nothing is tagged as Target Color/Digit Content inferred), but the corresponding Color/Digit word is tagged as Target observed, and there is some content that is tagged as Masker Color/Digit Content inferred, then tag that content as to be avoided in the response.

Use what you heard, even if it is a masker: If Target Color/Digit content is missing, (nothing is tagged as Target Color/Digit Content inferred), and NO Color/Digit word is tagged as Target observed, and there is some content that is tagged as Masker Color/Digit Content inferred, then tag that content as to-be-used in the response.

Use a pair: If neither color nor digit has been tagged as to-be-used, and there is a color and digit tagged as Same_stream_pair_content, choose a pair at random if there is more than one, and tag the color as use_color and the digit as use_digit.

Use a singleton: If neither color nor digit has been tagged as to-be-used, and there is a color/digit word with color/digit content, and it is not tagged as a Target/Masker color/digit inferred, nor part of a Same_stream_pair, tag the content as to-be-used in the response, randomly choosing if there is more than one color or digit content present that meets these conditions.

Pick response object

At this point the content to be used and avoided has been tagged, so the last task in the strategy is to pick an object on the display to click on. As a simplification of the model, EPIC's visual parameters are set to eliminate a need to move the eyes around to determine the colors and digits of the display objects, since this visual search would play no role in the phenomena of interest. The following rules are used to pick the object to click on:

- If there is an object whose color and digit are both tagged as use, pick it.
- If not, pick an object at random whose digit is tagged as use and whose color is not tagged as avoid.
- If no digit tagged as use, then pick an object at random whose color is tagged as use, and whose digit is not tagged as avoid.
- If neither color nor digit is tagged as use, then pick an object at random whose color and digit are not tagged as avoid.

Finish trial

After the object has been clicked on, the strategy waits for the objects to disappear, cleans up working memory by discarding all of its tag information, and then waits for a new trial to start.

Appendix 2

Stream Tracking Model Strategy for Replication 1a and Replication 2

BrungartMSV7aUse.prs, BrungartMSV7aAvoid.prs

At the top level, the strategy is to first listen to the messages, choose and make a response, and then repeat for the next trial. The strategy as described works for more than two talkers, so at a few places in the strategy, the number of talkers is taken into account. Also, for brevity, this description includes both the Use-Masker and Avoid-Masker strategies - there is only one point of difference, in the response selection rules.

Listening to the messages

Overview. The strategy assumes that the messages have a fixed structure, and it thus essentially "counts" words (segments) to know what message role each word plays. For example, the strategy can ignore the content of the first word (it is always "ready"), but then the strategy expects the next word to be a call sign, and so tags the word object as a call sign word, but its content (e.g. *baron*) may or may not be available due to masking - in this model, a word object always has a Stream ID, but it may have been mis-assigned by the stream tracking perceptual process. The strategy can ignore the content of the next word object, which is the segment that combines "go" and "to", and tag the following one as a color word, whose content (e.g. *red*) may not be available due to masking. The strategy knows that the word object that follows a color word is a digit word, and will treat it similarly. The content of the final word ("now") can be ignored, but its appearance means that the messages are complete and the strategy can go on to the next phase.

As the messages are heard, the strategy adds tags to word objects and other symbols. For example, if a call sign word has content that matches the Target call sign (baron), the strategy will tag that call sign word as the Target call sign word, and tag that word's StreamID as the Target stream. The strategy can then tell that later words with the same StreamID are from the Target stream.

The production rules are written to not add tags to items that already have the same tags; for brevity, this condition is not described below.

Call sign processing

Tag "observed" streams identification. When the call sign words have appeared, the strategy waits for any content of these words to be recognized. If a call sign word's content is available and is the Target call sign or a Masker call sign, the strategy tags that word as the Target or Masker call sign word, and that word's StreamID as the Target stream or Masker stream accordingly. For the Target stream, the strategy also tags the StreamID as an "observed" Target stream to indicate that the strategy actually had the call sign content to fully identify the source stream, as opposed to inferring it (described next).

Tag "inferred" stream identification. The previous operation may not succeed in tagging both source streams; for example, the content of one call sign might not be available due to masking. However, the strategy may infer the missing information. For example, say the StreamID of the target call sign has been correctly tagged because the call sign content was available, but the content of the other call sign word is not available, but its StreamID is present because the stream tracker always assigns one. That StreamID must be the Masker stream. If there is more than one such stream, as in the three- or four-talker case, they can all be tagged as Masker streams. However, if Target call sign content is not available, the target StreamID can only be inferred if *all* of the other streams (one, two, or three in number, depending on the number of talkers) have already been tagged as Masker streams. Any call sign words identified by inference are tagged as Target/Masker call sign word, and any streams so identified are tagged as Target/Masker stream, and also tagged as Target/Masker stream inferred.

Color word processing

Tag from StreamID. When the color words have appeared, the strategy tries to associate them with streams regardless of whether their content is available, based only on the observed or inferred status of the Target and Masker StreamIDs. If the Target/Masker StreamID is tagged, and a color word has the same stream, then the strategy tags that color word as Target or Masker.

Digit word processing

When the digit words have appeared, the strategy performs the same tagging and inference steps on them as it does for color words, resulting in digit words with the corresponding Target or Masker tags.

Tag pairs

Once all of the content words have appeared, the strategy marks pairs of (color word, digit word) that are from the same stream. If no stream got tagged as being the target stream (either directly or inferred), then the strategy that maximizes the probability of correctly responding both target color and digit chooses a color and a digit from the same stream is a better guess than a color and digit from different streams.

Choosing and Making a Response

Tag content

The strategy first applies stream tagging for word objects to tag the content of the color and digit words with their source streams: If a word is tagged as Target/Masker color/digit, and has available content, it tags that content with the same Target/Masker color/digit information. These tagged content items will be used to choose the actual response. Because content might be masked, it is possible that a word will be tagged with the Target stream but have no content available to use in the response. For example, a color word object might be tagged as Target color word, but if its content is not available, there will not be any color content tagged as Target color content. Finally, if there is a Same_stream_pair whose color and digit word content is available, the strategy tags that color and digit content as Same_stream_pair content.

Choose response information

At this point all necessary information has been extracted from the input messages, so the strategy is to make use of it to pick the best response. The strategy chooses which color or digit to use in the response, and which to avoid in case is it necessary to guess from the alternatives on the display. The following rules are applied:

Use target: If a color/digit is tagged as Target content observed or inferred, the strategy tags it as to-be-used in the response.

Note: The following two rules apply to the Use-Masker and Avoid-Masker versions; only one of these rules can be enabled at a time. The two versions of this strategy are otherwise identical.

Use-Masker Version: Use what you heard, even if it is a masker: If the Target stream was inferred (rather than observed), and Target Color/Digit content is missing, and there is some content that is tagged as Masker Color/Digit Content, then tag that content as to-be-used in the response.

Avoid-Masker Version: Do not use known masker content in the response: If Color/Digit content is tagged as being from the Masker, then tag that content as to-be-avoided in the response.

Use a pair: If there is no stream tagged as the Target stream, and neither color nor digit has been tagged as to-be-used, and there is a color and digit tagged as Same_stream_pair_content, choose a pair at random if there is more than one, and tag the color and digit as to-be-used.

Use singletons: If there is no stream tagged as the Target stream, and there is no Same_stream_pair content, and there is a color/digit word with color/digit content that is not already tagged as to-be-used, tag the content as to-be-used in the response, randomly choosing if there is more than one color or digit content present that meets these conditions.

Pick response object

At this point the content to be used and avoided has been tagged, so the last task in the strategy is to pick an object on the display to click on. As a simplification of the model, EPIC's visual parameters are set to eliminate a need to move the eyes around to determine the colors and digits of the display objects, since this visual search would play no role in the phenomena of interest. The following rules are used to pick the object to click on:

- If there is an object whose color and digit are both tagged as use, pick it.
- If not, pick an object at random whose digit is tagged as use and whose color is not tagged as avoid.
- If no digit tagged as use, then pick an object at random whose color is tagged as use, and whose digit is not tagged as avoid.
- If neither color nor digit is tagged as use, then pick an object at random whose color and digit are not tagged as avoid.

Finish trial

After the object has been clicked on, the strategy waits for the objects to disappear, cleans up working memory by discarding all of its tag information, and then waits for a new trial to start.

Appendix 3

Stream Tracking Model Strategy for Brungart 2001, Mixture Model 1

BrungartMSV7a7Use.prs

At the top level, the strategy is to first listen to the messages, and then choose and make a response, and then repeat on the next trial. In a few places in the strategy, the number of talkers is taken into account.

Listening to the messages

Overview. The strategy assumes that the messages have a fixed structure, and it thus essentially "counts" words to know what message role each word plays. For example, the strategy can ignore the content of the first word (it is always "ready"), but then the strategy expects the next word to be a call sign, and so tags the word object as a call sign word, but its content (e.g. baron) may or may not be available due to masking - in this model, the Stream ID for a word object is always available, but may get mis-assigned by the stream tracking perceptual process. Then the strategy can ignore the content of the next word object, which in this model combines "go" and "to", and tag the following one as a color word, whose content (e.g. red) may not be available due to masking. The strategy knows that the word object after a color word is a digit word, and will treat it similarly. The content of the final word ("now") can be ignored, but its appearance means that the messages are complete and the strategy can go on to the next phase.

During listening to the messages, the strategy adds tags to word objects and other symbols. For example, if a call sign word has content that matches the Target call sign (baron), the strategy will tag that call sign word as the Target call sign word, and tag that words's StreamID as the Target stream. The strategy can then tell that later words with the same StreamID are from the Target stream.

The production rules are written to not add tags to items that already have the same tags; for brevity, this condition is not described below.

Call sign processing

Tag "observed" streams identification. When the call sign words have appeared, the strategy waits for any content of these words to be recognized. If a call sign word's content is available and is the Target call sign or a Masker call sign, the strategy tags that word as the Target or Masker call sign word, and that word's StreamID as the Target stream or Masker stream accordingly. For the Target stream, the strategy also tags the StreamID as an "observed" Target stream to indicate that the strategy actually had the call sign content to fully identify the source stream, as opposed to inferring it (described next).

Tag "inferred" stream identification. The previous operation may not succeed to tagging both source streams; for example, the content of one call sign might not be available due to masking. However, the strategy can often infer the missing information.

If Target call sign content is not available, the target StreamID can only be inferred if all of the other streams (one, two, or three in number, depending on the number of talkers) have already been tagged as Masker streams. Note that in this strategy, Masker streams are NOT inferred from the Target stream. Any call sign words identified by inference are tagged as Target call sign word, and any streams so identified are tagged as Target stream, and also tagged as Target stream inferred.

Color word processing

Tag from StreamID. When the color words have appeared, the strategy tries to associate them with streams regardless of whether their content is available, based only on the observed or inferred status of the Target and Masker StreamIDs. If the Target/Masker StreamID is tagged, and a color word has the same stream, then the strategy tags that color word as Target or Masker.

Digit word processing

When the digit words have appeared, the strategy performs the same tagging and inference steps on them as it does for color words, resulting in digit words with the corresponding Target or Masker tags.

Tag content

The strategy first applies the stream tagging for word objects to tag the content of the color and digit words with their source streams: If a word is tagged as Target/Masker color/digit, and has available content, it tags that content with the same Target/Masker color/digit information. These tagged content items will be used to choose the actual response. Because content might be masked, it is possible that a word will be tagged with the Target stream but have no content available to use in the response. For example, a color word object might be tagged as Target color word, but if its content is not available, there will not be any color content tagged as Target color content.

Choose response information

At this point all necessary information has been extracted from the input messages, so the strategy makes use of it to pick the best response. The strategy chooses which color or digit to use in the response, and if no selections are made, the color or digit will be guessed. In this model, maskers are not avoided;this model is a mixture model in which a "fallback" rule in the strategy uses word loudness to choose a response.

First, the strategy decides whether loudness will be used for color choice: If there are two call sign words with different genders, flip a 0.5 coin to decide whether to use loudness in the color choice. Then the following rules are applied:

Use observed target for color if not using loudness: If not using loudness in the color choice, and a stream is tagged as Target stream observed, and a color is tagged as Target color content, tag it as to-be-used in the response.

Use loudest color content if masker not observed: If using loudness in the color choice, and there is a color word that is not tagged as a Masker, and its content is available, and it is louder than some other color word, then tag that color content as to-be-used in the response.

Use observed target digit content: If a stream has been tagged as Target observed, and a digit content has been tagged as Target, then tag that digit content as to-be-used in the response.

Use color singleton: If loudness is being used in the color choice, and there is a color content present, but no color content has been tagged as to-be-used, tag the color content as to-be-used in the response, randomly choosing if there is more than one color present that meets these conditions.

Use digit singleton: If no stream has been tagged as Target observed, and there is a digit content present, but no digit content has been tagged as to-be-used, tag the digit content as to-be-used in the response, randomly choosing if there is more than one digit present that meets these conditions.

Pick response object

At this point the content to be used and avoided has been tagged, so the last task in the strategy is to pick an object on the display to click on. As a simplification of the model, EPIC's visual parameters are set to eliminate a need to move the eyes around to determine the colors and digits of the display objects, since this visual search would play no role in the phenomena of interest.

The following rules are used to pick the object to click on:

- If there is an object whose color and digit are both tagged as use, pick it.
- If not, pick an object at random whose digit is tagged as use and whose color is not tagged as avoid.
- If no digit tagged as use, then pick an object at random whose color is tagged as use, and whose digit is not tagged as avoid.
- If neither color nor digit is tagged as use, then pick an object at random whose color and digit are not tagged as avoid.

Finish trial

After the object has been clicked on, the strategy waits for the objects to disappear, cleans up working memory by discarding all of its tag information, and then waits for a new trial to start.