Received Date: 15-Jun-2015

Revised Date: 18-Oct-2015

Accepted Date: 06-Nov-2015

Article Type: Article

Modeling Two-Channel Speech Processing with the EPIC Cognitive Architecture

David E. Kieras (kieras@umich.edu)

Electrical Engineering & Computer Science Department, University of Michigan

2260 Hayward Street, Ann Arbor MI 48109-2121, USA

(corresponding author)

Gregory H. Wakefield

Electrical Engineering & Computer Science Department, University of Michigan

Eric R. Thompson

Battlespace Acoustics Branch, Air Force Research Laboratory, Wright-Patterson AFB

Nandini Iyer

Battlespace Acoustics Branch, Air Force Research Laboratory, Wright-Patterson AFB

Brian D. Simpson

Battlespace Acoustics Branch, Air Force Research Laboratory, Wright-Patterson AFB

Keywords: Cognitive architecture; two-channel speech; auditory perception; auditory streams

Abstract

An important application of cognitive architectures is to provide human performance models that capture psychological mechanisms in a form that can be "programmed" to predict task

performance of human-machine system designs. While many aspects of human performance have been successfully modeled in this approach, accounting for multi-talker speech task performance is a novel problem. This paper presents a model for performance in a two-talker task that incorporates concepts from psychoacoustics, in particular, masking effects and stream formation.

1. Introduction

A classic problem in cognitive psychology is the "cocktail party effect" in which a person is surrounded by several people speaking simultaneously, and is nonetheless able to follow a single speaker well enough to maintain a conversation, although some information about what the other speakers are saying appears to be available under some conditions. The early study of these phenomena (e.g. Cherry, 1953; Moray, 1959) defined the current concept of selective attention; the human listener was said to be able to selectively attend to one of the signal sources and "filter out" the others. In the decades since, a large number of additional studies and theoretical work has clarified what properties of the acoustic and perceptual situation contribute to the effect. Providing a comprehensive review is not possible in the limited space available for this paper, but relevant surveys are provided by Darwin (1997), Yost (1997), Bronkhorst (2000), Haykin and Chen (2005), Schneider, Li, and Daneman (2007), and Moore and Gockel (2012), and more recent individual studies are cited in what follows.

The most common experimental paradigm is that the subject listens to speech from two or more talkers who are speaking simultaneously, respond to information provided by only one of them, called the *target*, and ignore the information provided by the other talkers, called the *masker(s)*. The research has focussed on characterizing what aspects of the messages contribute to an interference between the target and the maskers, both in terms of the perception of sound in general, and of speech in particular. A general psychoacoustic effect is *masking*, in which a sound may become less perceptible if another sound is simultaneously present. In the case of simple sounds or signals, masking effects are generally considered to be a result of interactions in the cochlea itself, for example, the excitation pattern on the basilar membrane produced by the target sound is disrupted by the masking sound, making it less detectible. In the context of

simultaneous speech messages, a distinction is made between *energetic masking*, which refers to the interference produced at the acoustic or sensory level, as in the masking of simple sounds, and *informational masking*, which is interference produced at the higher perceptual and cognitive levels and makes it difficult for the human listener to follow the target message in the presence of the masker message, above and beyond the effects of energetic masking.

Another key concept is *auditory streams* (Bregman, 1990), the notion that the acoustic field sensed by the ears is decomposed by the auditory system into one or more temporally coherent sound sources (for reviews, see Darwin, 1997; Moore & Gockel, 2012). While, in general, the mapping between sources in the acoustic field and those perceived by the auditory system is not one-to-one, in a two-talker task, it is believed that each talker is perceived as a distinct stream, and the listener's task is to determine which sounds go with which stream and choose the appropriate response. Performance in the two-talker task thus reflects a combination of the energetic masking effects and informational masking effects on stream formation and segregation (Schneider, Li, & Daneman, 2007).

The mainstream psychoacoustic work on the cocktail-party effect has focussed on "front end" processes of signal detection and estimation, using mathematical models, but these accounts do not have a well-defined way to incorporate "back end" cognitive-strategy processes that represent how a listener will choose a response that meets the task requirements, which can be surprisingly subtle even in simple tasks (Meyer & Kieras, 1997b, 1999). In contrast, even though mainstream cognitive architecture approaches are committed to an "end-to-end" goal of representing perception through cognition to action, they have focussed primarily on cognitive processes, and have tended to ignore the difficult aspects of perceptual processes (see Kieras, in press, for an overview).

The present paper combines basic psychoacoustic mechanisms with a cognitive architecture to model human performance in a two-talker listening task. EPIC (Executive/Process-Interactive Control) is one among several architectures whose goal is to provide an integrated account of human abilities and limitations in perception, cognition, and action. A relatively simple psychoacoustic model was incorporated into the EPIC cognitive architecture, and a task strategy was expressed as production rules, to provide an "end-to-end" account of performance in a well-studied two-talker speech perception task. Both the perception model and the task strategy are required to account for important effects in the task performance.

Earlier forms of this model appear in Kieras, Wakefield, Thompson, Iyer, & Simpson (2014) and Wakefield, Kieras, Thompson, Iyer, & Simpson (2014); the model presented here has the same strategy component, but the perceptual models are considerably improved, taking into account how pitch differences affect both content detection and stream segregation. The result is a model with far fewer parameters that must be estimated from the data. A detailed comparison of the improved perceptual model with the previous one is not possible in the available space here; the reader can compare this model with the one in Kieras, et al (2014).

Following a summary of the experimental task and its results, an overview of EPIC will be presented and key extensions of the auditory processing module will be introduced. Within the framework imposed by these extensions, a model for the two-talker listening task will be proposed and fit to the human data.

2. The experimental task

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

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

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

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

2000).

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

3. An experiment and its data

This paper provides a model for the data from Experiment 2 in Thompson, Iyer, Simpson, Wakefield, Kieras, & Brungart (2015). This experiment, based on Brungart (2001), manipulated the acoustic similarity of the two talkers, varying from Different Sex (DS), to Same Sex (SS), to Same Talker (ST), and also manipulated the relative loudness of the two messages, the Signal-to-Noise ratio (i.e. the Target-to-Masker ratio) from -18 to +9 dB, and provided performance incentives to help stabilize the subject strategies. Finally, in addition to the proportion of *Both-Correct* responses (both Color and Digit are Target), they also reported the proportions of responses that matched Target, Masker, or Neither separately for Color and Digit.

3.1. The Thompson et al. results

Because of the multiple factors and measures involved, the effects are somewhat complex. The six panels of Fig. 1 show the results as the *observed* points (solid points and lines; the *predicted* points will be explained later). Each panel plots the proportion of responses that matched the Target, the Masker, and Neither, as a function of the signal-to-noise (SNR) ratio in dB. The upper panels display the proportions for Color responses; the lower panels display the proportions for Digit responses. In addition, the panels show the proportion of Both-Correct

responses in which both Color and Digit are from the Target message. These black curves are the same in the upper and lower panels. The left-to-right panels display the results for the similarity of the Target and Masker talkers. From left to right, the stimulus conditions are Different Sex, Same Sex but different talkers, and Same Talker.

——————————Insert Figure 1 about here——————————

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

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

3.2. Accounting for the phenomena

To date, a satisfactory theoretical account of two-talker CRM effects is lacking in the speech-perception field. Discussions have focused on the relative importance of informational masking over energetic masking, the roles of selected and divided attention, and the formation and maintenance of auditory streams. However, none of these concepts have been operationalized to the point of providing a quantitative theoretical account of experimental outcomes. What follows is an attempt to help bridge this gap.

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

specific effects in this data.

4. The architecture and model

An EPIC architecture model comprises a simulated human which interacts with a simulated task environment; the architecture describes the fixed components of the simulated human, controlled by a task-specific strategy represented as production rules. Due to space limitations, the usual description of the architecture is not provided here; see Meyer and Kieras (1997a, 1999) or Kieras (in press) for more discussion. The focus of this presentation is on the mechanisms of the auditory processor that have been added to the architecture, and the production-rule strategy for the task.

4.1. Model summary

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

4.1.1. Speech auditory input

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

Auditory perception constructs *auditory objects* based on properties of the physical input. There are two kinds of auditory object: *word objects* represent individual perceived words that

have a temporal duration; *stream objects* represent perceived sound sources for these word objects.

4.1.2. Word objects

Word objects have a variety of properties, but for the purposes of this model, they may or may not have *content*, which is the recognized semantic item (e.g. *red*); this allows for a word to be "heard" but not recognized. Words also have *stream attributes*, which in this model are average loudness level (specified in dB) and average pitch (in semitones, where the number of semitones is defined as $12 \cdot \log_2(\text{pitch in Hz})$), both averaged over the duration of the word. Semitones provide a logarithmic scale for pitch, analogous to decibels for loudness. (For simplicity, we are assuming that perceived pitch and loudness correspond to the physical measures of semitones and decibels, respectively.) This model assumes that the stream attributes are *always* perceived.

Whether the content of a word object is recognized in the presence of the other word objects is assumed to be a basic energetic masking phenomenon. The probability of content detection depends on the SNR, that is, the loudness level of the word relative to the other word objects that are simultaneously present, and the pitch difference between the two word objects. With respect to the latter, studies show that discrimination of simultaneous vowel sounds improves with pitch difference, though increasing the difference beyond about 4 semitones produces no further improvement (Assmann & Summerfield, 1990). This effect was incorporated in the model by computing an *Effective SNR* that is the weighted sum of the loudness difference in dB (the SNR) and the pitch difference in semitones capped at 4.

4.1.3. Stream objects and stream tracking

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

The auditory perceptual processor assumes that there are as many stream objects as input sources, each with a unique but arbitrary *StreamID* attribute, and attempts to assign each incoming word object to one of the streams, using the stream-related attributes of loudness and

pitch to do so. Once the assignment is done, the stream percepts are updated to reflect the loudness and pitch properties of the words assigned to them, and the next pair of word objects will be assigned to the updated streams. Thus the auditory processor *tracks* the streams.

4.1.4. Cognitive strategy and response choice

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

Because the loudness and pitch of each word in the utterances varies within the same talker, it is possible for individual words from two different talkers to be mis-assigned to the streams, so that each stream is associated with a mixture of words from the two talkers. Fig. 2 shows an example in which the Color words have been assigned to the wrong stream, while the Digit words were assigned to the correct stream. This will lead to a response with the Masker Color and the Target Digit.

————————Insert Figure 2 about here————————

The cognitive process for selecting a response makes use of the recognized content of the word objects together with the stream associated with each word object. For example, as in Fig. 2, if the word object whose content is the Target Callsign *Baron* is associated with Stream2 and there are two word objects associated with the same stream whose content has been recognized as the Color *Red* and the Digit *8*, then *Red 8* will be used to specify the response to be made.

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

4.1.5. Theoretical summary

In terms of conventional attention theory, this is a "*very* late selection" model — *all* of the information produced by perception is available to cognition for choosing the response. The problem of trying to handle two simultaneous messages is not represented as a failure to select

the correct stream prior to cognition, but rather that masking effects and errors in stream assignment will result in a collection of information about the perceived messages that may be incomplete or incorrect (e.g. as in Fig. 2), and the task strategy must make use of this information to choose a response that meets the task requirements.

4.2. Model details and parameters

4.2.1. Corpus statistics drive the model

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

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

4.2.2. Content detection parameters

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

——————————Insert Table 1 about here——————————

The content detection process is modeled along the lines suggested by Wichman & Hill (2001). With a low probability (the lapse rate α), subjects will fail to recognize content (even at very high SNRs); otherwise, the probability of content detection follows a gaussian detection function of Effective SNR, with parameters of mean μ and standard deviation σ. The parameters w, α and σ are assumed to be constant across the type of content word (Callsign, Color, Digit), while μ is assumed to have a different value for each type of content word (Callsign, Color, Digit). For

completeness, the content detection functions for the filler words *ready*, *goto*, and *now*, were specified, but for simplicity were made the same as the Callsign detection function because the *content* of the filler words plays no role in stream tracking or response strategy.

4.2.3. Stream tracking details and parameters

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

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

4.2.4. Cognitive processor strategy exploration

The auditory perception components in the EPIC architecture take the input utterance segments and perform content detection and stream tracking and provide the resulting content and StreamID attributes of the individual word segments, like that shown in Fig. 2, to the cognitive processor, which is running a strategy implemented in production rules.

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

incorporates symmetrical inferences.

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

4.2.5. Strategy summary

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

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

4.3. Model fitting and results

The parameter values shown in Table 1 were determined by Monte-Carlo runs of the EPIC model using a grid search on high-performance computer clusters provided by AFRL through mindmodeling.org. The search goal was to maximize r^2 between predicted and observed values for the Target and Masker Color and Digit probabilities (blue and red curves in Fig. 1). Each Monte-Carlo run used 3000 trials per talker/SNR condition. There are a total of 240 empirical data points with at least 120 degrees of freedom; eight parameter values were varied in the search. The best-fit values are shown in Table 1.

Fig. 1 shows the predictions from the EPIC model as open points and dotted lines. All three conditions are well handled with a small set of parameters that describe how the auditory perceptual process is affected by the acoustic properties of the input as provided by the corpus statistics based on the segmentation. It is especially noteworthy that unlike the model presented in Kieras et al. (2014), there are no parameters that are specific to talker similarity conditions — the pitch difference used in content detection and stream tracking accounts for these effects.

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

However, there is a clear tendency for the Both-Correct points to be generally under-predicted, probably because our simple model of the stream tracking is not efficient enough. To show this, in Fig. 3 are the conditional probabilities of selecting the Target, Masker, or Neither Digit given that the listener has correctly identified the Target Color. On the whole, the likelihood of choosing the Target Digit is far higher than the others, meaning that if the listener has tracked the Target stream correctly as far as the Color, then he or she is very likely to get the Digit correct as well because the stream tracking is very likely to be correct. In contrast, choosing a Masker Digit would be a case of the stream tracking "switching" to the Masker stream. Notice that the model is consistently less likely to choose the Target Digit than the subjects, and is more likely to switch to the Masker or Neither Digit, especially when the tracking is more difficult in the

vicinity of 0 dB SNR. The result is a tendency to under-predict the Both-Correct responses, even though the individual Target and Masker responses are well predicted.

——————————Insert Figure 3 about here——————————

5. Conclusions

We now have a successful model of the two-talker task that explains performance in terms of the familiar concepts of masking and stream perception, but does so in the form of a computational cognitive architecture that distinguishes between the perceptual mechanisms that analyze the auditory input, and the cognitive mechanisms that apply a strategy for how to use the perceptual information to perform the task. The model is specified enough to provide a detailed quantitative account of the data using a small number of parameters estimated to fit the data.

The most important theoretical conclusion suggested by this model is that human performance in this task is limited by low-level perceptual difficulties, not central limitations such as an "attentional filter" or "bottleneck." This is an example of the "very late selection" concept encouraged by the EPIC architecture, in which perception delivers as much information as it can to cognition where the strategy uses it to produce the response (c.f. Kieras, in press). Thus, rather than trying to explain how auditory perception has some kind of "selective attention" mechanism that in some ill-defined way "filters out" the masker message, we proposed rather simple, general, and quantitatively-specified low-level perceptual mechanisms for masking and stream tracking, and a fairly simple cognitive strategy, so that the model produces the overall effect of selective attention, but without any concept of a "filter" or "bottleneck" (c.f. Meyer and Kieras, 1997a, b; 1999).

Of course, the model needs further development. Improving the efficiency of the stream tracking mechanism might make the model's tracking as "sticky" as human tracking. But the single most important phenomenon challenging the model is the substantial degradation of performance when more than one masking talker is speaking. Compared to a single masker, if there are two maskers, performance is disproportionately poor, and is only a little worse if a third masker is added. The model as described here functions in the two- and three-masker cases, and approximates the poorer performance if the content detection functions are made substantially

less sensitive and steeper, representing stronger masking effects. However, we need a more fundamental explanation for the effect of multiple maskers than ad-hoc parameter adjustments. We are currently exploring a finer temporal grain size in the model which can support a "glimpsing" approach to content detection (e.g. Brungart & Iyer, 2012) to explain these effects in a more general way.

Acknowledgements

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, (D.E.K and G.H.W) and grants from the Air Force Office of Sponsored Research (AFOSR) (N.I. and B.D.S.).

References

Assmann, P.F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America.* 88, 680–697.

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

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

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

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

Brungart, D.S., & Iyer, N. (2012). Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *Journal of the Acoustical Society of America.* 132(4), 2545-2556. [http://dx.doi.org/10.1121/1.4747005]

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

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

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

Kieras, D.E. (in press). A summary of the EPIC Cognitive Architecture. In S. Chipman (Ed.), *The Oxford Handbook of Cognitive Science*.

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

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.

Moore, B.C.J, & Gockel, H.E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society*, 367, 919-931. doi:10.1098/rstb.2011.0355

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

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

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

Thompson, E.R., Iyer, N., Simpson, B.D., Wakefield, G.H., Kieras, D.E., & Brungart, D.S. (2015). Enhancing listener strategies using a payoff matrix in speech-on-speech masking experiments. *Journal of the Acoustical Society of America*, 138(3), 1297-1304. [http://dx.doi.org/10.1121/1.4928395]

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.

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

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

Tables

Table 1. Best-fit parameter values

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

Figure File Names & Captions

KierasFig1.eps

Fig. 1. Data from Thompson et al.(2015) Experiment 2. Observed (solid points and lines) and Predicted (open points and dotted lines) proportion of responses as a function of SNR and talker similarity. Top panel shows Color responses, bottom panel shows Digit responses. In order from the top down, the curves are as follows: Blue curves with diamond points are for Target responses, black curves with circle points are for Both-Correct responses (both color and digit

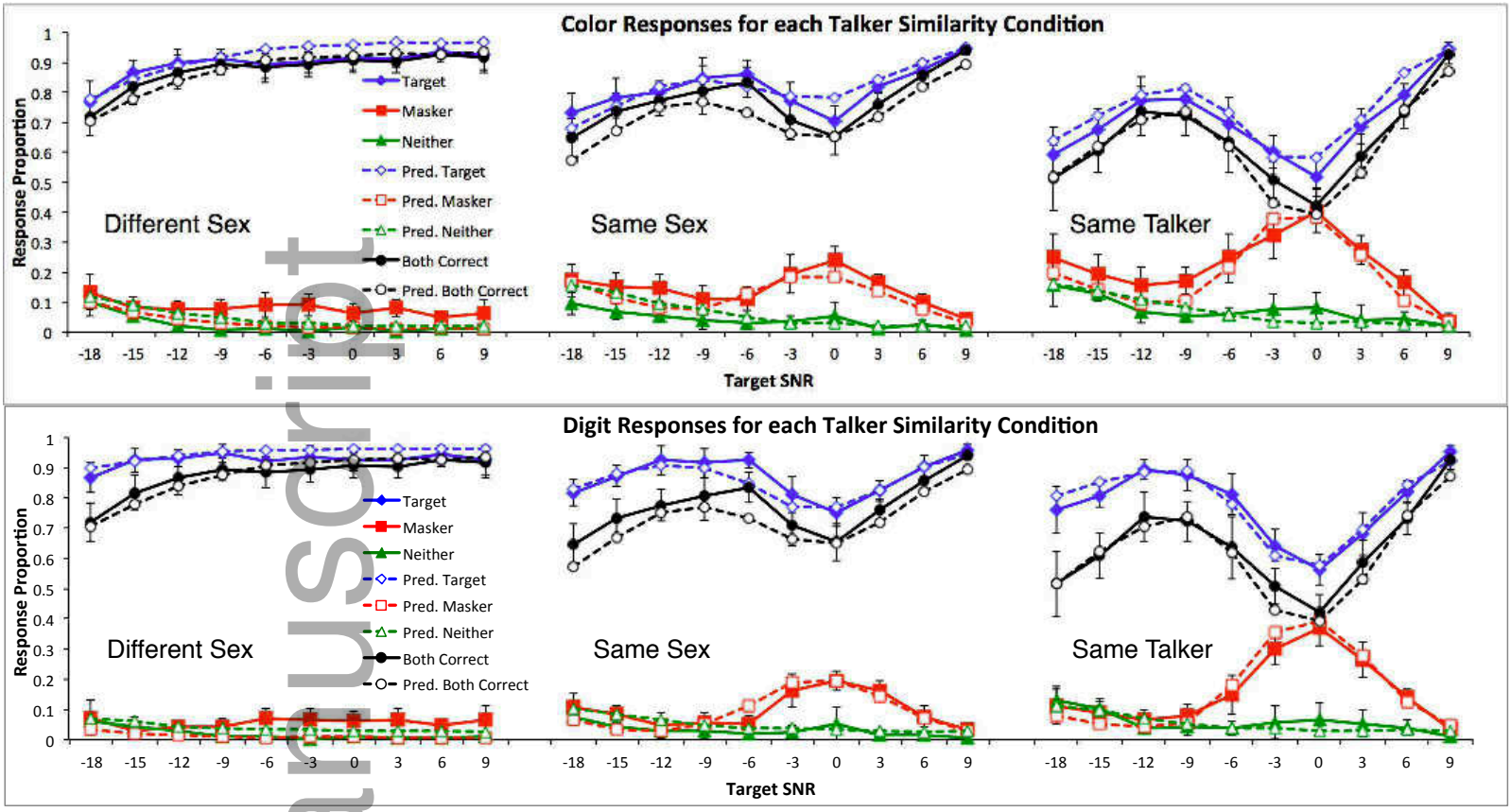
from the Target), and are the same in the top and bottom panels; red curves with square points are for Masker responses, and green curves with triangles for neither Target nor Masker. Error bars show 95% confidence intervals for the means averaged over individual subject proportions.

KierasFig2.eps

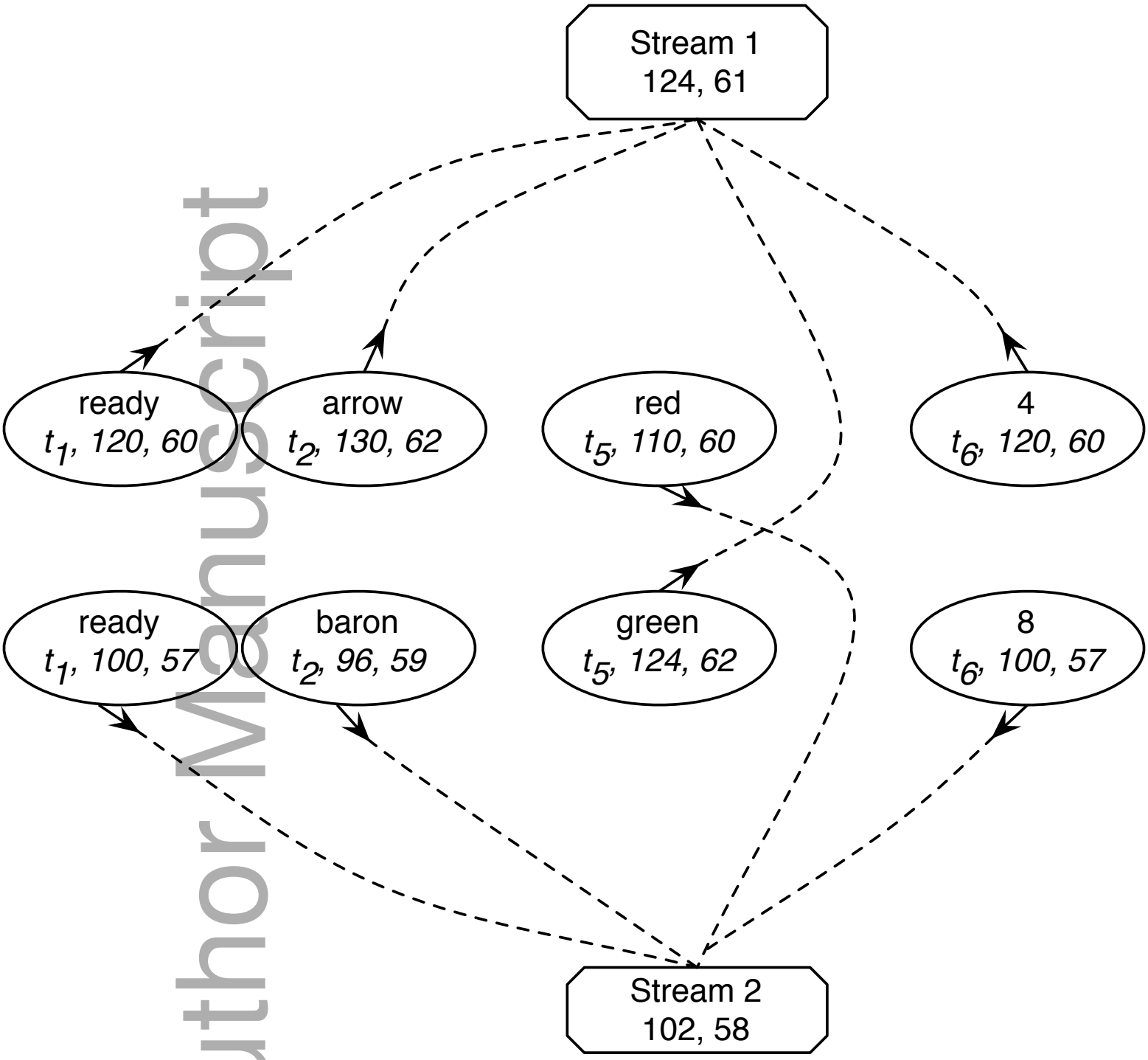
Fig. 2. Example showing contents of working memory after erroneous stream tracking. The polygonal boxes top and bottom are the two stream objects, showing mean pitch (Hz) and loudness level (dB) values. The ovals are the word objects in each message in left-to-right time order (goto and now omitted for clarity), showing the content, time stamp, pitch, and loudness. During perception, each word was associated with its closest stream, but because the Color word pitches were discrepant, they were assigned to the wrong stream.

KierasFig3.eps

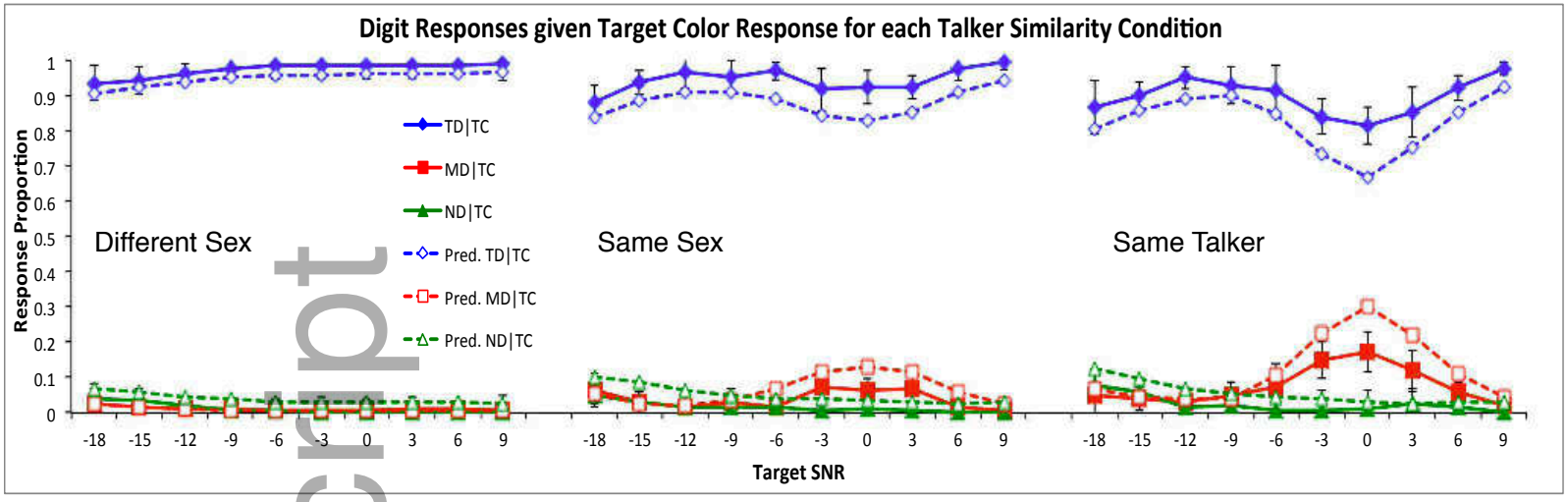
Fig. 3. Data from Thompson et al.(2015) Experiment 2. Observed (solid points and lines) and Predicted (open points and dotted lines) conditional probabilities of selecting a Target, Masker, or Neither Digit response given that the Color response was the correct target Color, as a function of SNR and talker similarity. In order from the top down, the curves are as follows: Blue curves with diamond points are for selecting the Target Digit (TD|TC), red curves with square points are for selecting the Masker Digit (MD|TC), and green curves with triangles are for selecting a Digit that is neither the Target nor a Masker Digit (ND|TC).



tops_12180_f1.eps



tops_12180_f2.eps



tops_12180_f3.eps