# 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

**Gregory H. Wakefield (ghw@umich.edu)**
Electrical Engineering & Computer Science Department, University of Michigan
2260 Hayward Street, Ann Arbor MI 48109-2121, USA

**Eric R. Thompson (eric.thompson.22.ctr@us.af.mil)**
Ball Aerospace & Technologies Corp.
2610 Seventh St. B441, Wright-Patterson AFB, Ohio 45433

**Nandini Iyer (nandini.iyer.2@us.af.mil)**
Battlespace Acoustics Branch, Air Force Research Laboratory
2610 Seventh St. B441, Wright-Patterson AFB, Ohio 45433

**Brian D. Simpson (brian.simpson.4@us.af.mil)**
Battlespace Acoustics Branch, Air Force Research Laboratory
2610 Seventh St. B441, Wright-Patterson AFB, Ohio 45433

**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 the psychoacoustic study of speech perception, in particular, masking effects and stream formation.

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

## 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) led to a body of additional studies and theoretical work that 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. The most common experimental paradigm is that the subject must listen to simultaneous speech inputs from two or more talkers (human speakers), but respond to the information provided by only one of them. Some more recent research over the last decade has used more precise procedures to help characterize the determinants of performance; in particular many experiments have been done using the *coordinate response measure* (CRM) speech corpus which represents a highly simplified form of the command and control messages used in military settings (Bolia, Nelson, Ericson, & Simpson, 2000).

The mainstream psychoacoustic work on this problem applied the mathematical tools of signal analysis that have been successful in characterizing human ability to detect and discriminate sounds. A less formal but influential concept was *auditory streams* (Bregman, 1990), the notion that we perceive separate sound sources based on the detailed properties of the incoming sounds. In a two-talker task, each talker would be perceived as a stream, and the listener's task is to determine which sounds go with which stream and choose the appropriate response. This process must involve a combination of perceptual mechanisms and cognitive strategies. However, psychoacoustic accounts of the task have focussed on "front end" processes of signal detection and processing and did not have a well-defined way to take into account the possibly complex "back end" processes of cognitive strategies involved in the task. In contrast, cognitive architecture research developed powerful theoretical mechanisms for the "back end" processing, especially using production systems, but tended to ignore difficult details of perceptual processes.

The present paper combines mathematical models of speech perception 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 psychoacoustic speech perception model was incorporated into the EPIC cognitive architecture to provide an integrated account of performance in a well-studied two-talker speech perception task. We devised a relatively simple speech perception model and a strategy which together account for important factors that determine performance.

An earlier form of this model appears in Kieras, Wakefield, 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 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 review of the two-talker CRM listening task, an overview of EPIC will be presented and key extensions of the auditory processing module will be introduced. Within the framework imposed by these extensions, a model for the two-talker CRM listening task will be proposed and fit to the human data.

## Replication of a Two-Talker Dataset

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 selecting the appropriate Color/Digit pair from a 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 whose words are simultaneous, and must choose the color-digit pair associated with the Target callsign, and was instructed to ignore the Masker message. The responses are scored as matching the Target message, the Masker message, or Neither.

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

Rather than show his results in this paper, however, we present the results for a methodologically improved replication which is very similar in design and results to Brungart (2001). The replication followed the conditions and procedures of Brungart (2001) in all respects except two: (1) The SNR, which ranged from -12 to +15 dB in the original study, was shifted to a lower range (-18 to +9 dB) in the interest of studying performance at SNRs closer to

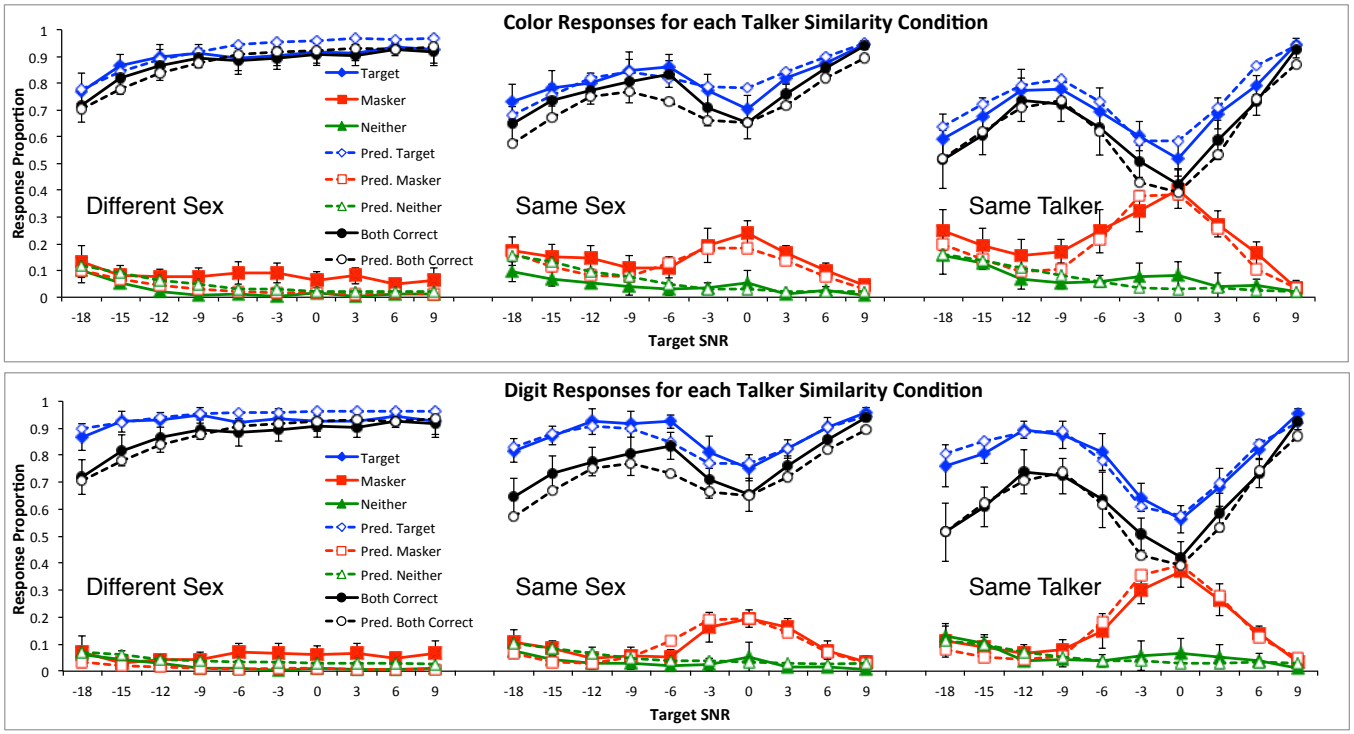


*Figure 1. Observed (solid points and lines) and Predicted (open points and dotted lines) proportion of responses as a function of SNR and talker similarity. Top panel shows Color responses, bottom panel shows Digit responses. In order from the top down, the curves are as follows: Blue curves with diamond points are for Target responses, black curves with circle points are for completely 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.*

masked detection thresholds; (2) the replication clarified the task instructions with a point reward system for correct performance, and provided performance feedback at the end of each trial and during the experiment.

### Results

The six panels of Figure 1 show these somewhat complex experiment results as the *observed* points (solid points and lines; the *predicted* points will be explained later). Each panel plots the proportion of Target, Masker, and Neither responses as a function of the signal-to-noise (SNR) ratio in dB. The upper and lower panels display the proportion for Color and Digit responses separately. 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 based on the similarity of the Target and Masker talkers. From left to right, the stimulus conditions are Different Sex, Same Sex but different talkers, and Same Talker.

The basic effects are as follows: overall, with increasing positive SNR, the completely 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 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 very low (negative) SNRs, but then improves, and then declines again in the vicinity of 0 dB SNR, and then improves again.

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

### Accounting for the Phenomena

To date, a theoretical account of the two-talker CRM results remains incomplete. Discussions have focused on the relative importance of informational masking over energetic masking, the roles of selected and divided attention, and the formation and maintenance of auditory streams. However, none of these concepts have been operationalized to the point of providing strong predictions 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.

## The Architecture and Model

An EPIC architecture model comprises a simulated task environment which interacts with a simulated human; 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 (1997, 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.

### Model Summary

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

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

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

*Word objects*. Word objects have a variety of properties, but for the purposes of this model, they may or may not have *content*, which is the recognized semantic item (e.g. *red*); this allows for a word to be "heard" but not recognized. Words also have *stream attributes*, which in this model are average loudness level (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. This model assumes that the stream attributes are *always* perceived.[1]

Whether the content of a word object is recognized in the presence of the other word objects is assumed to be a basic masking phenomenon. The probability of content detection depends on the SNR, that is, the loudness level of the word

[1] For simplicity, we are assuming that perceived pitch and loudness correspond to physical semitones and dB.

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.

*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.

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

Because the loudness and pitch of each word in the utterances varies within the same talker, it is possible for individual words from two different talkers to be mis-assigned to the streams, so that each stream is associated with a mixture of words from the two talkers. Figure 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.

The cognitive process for selecting a response makes use of the recognized content of the word objects together with the stream associated with each word object. For example, as in Figure 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.

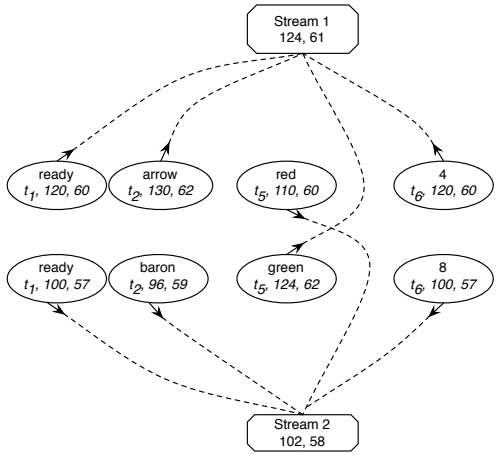


*Figure 2. Example showing contents of working memory after erroneous stream tracking. The polygonal boxes top and bottom are the two stream objects, showing mean pitch (Hz) and loudness level (dB) values. The ovals are the word objects in each message in left-to-right time order (*goto *and* now *omitted 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.*

*Theoretical summary.* In terms of conventional attention theory, this is a "*very* late selection" model - *all* of the information produced by perception is available to cognition for choosing the response.

The problems of trying to handle two simultaneous messages is not represented as a failure to select the correct stream prior to cognition, but rather that masking effects and errors in stream assignments will result in a collection of perceptual information about the messages that may be incomplete or incorrect (e.g. as in Figure 2), and the task strategy must make use of this information to choose a response that meets the task requirements.

## Model Details and Parameters

*Corpus statistics drive the model.* We computed the average 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. Inside the auditory system module, pitch differences were always capped at 4 semitones, a constant value based on Assmann & Summerfield (1990) and not estimated to fit the data.

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

*Stream tracking details and parameters.* The stream tracking parameters are also 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 - λ). The pitch difference was capped at 4 semitones. The new word segments are then assigned to streams so as to minimize the total distance between all words and their assigned streams. The streams are then updated to include their newly assigned word segments, and the resulting means used to predict the segment that follows.

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

*Cognitive processor strategy exploration.* The auditory perception components in the EPIC architecture take the input utterance segments and perform content detection and stream tracking and provide the resulting content and StreamID attributes of the individual word segments, like that shown in Figure 2, to the cognitive processor, which is running a strategy implemented in production rules. Over the course of this work, a variety of strategies were considered, and two key options were identified. The first is that in the 2-channel task, symmetrical inferences can be made; for example, if we know that one of the Color words is from the Masker stream, we can infer that the other Color word has to be from the Target stream.

The 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 known Masker content, 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. 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 is better than a pure guess. Thus the *Use-Maskers* strategy will use content known to be from the Masker stream if Target content was not detected, but only if the identity of Target stream had been *inferred* from detection of Masker callsign content. This model used both the symmetrical inferences and the Use-Masker options.

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 |

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

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

## Model Fitting and Results

The parameter values shown in Table 1 were determined by Monte-Carlo runs of the EPIC model with a grid search of the parameter values using high-performance clusters provided by AFRL through mindmodeling.org. The search goal was to maximize $r^2$ between predicted and observed values for the Target and Masker Color and Digit

probabilities (blue and red curves in Figure 1). Each Monte-Carlo run used 3000 trials per talker/SNR condition. There are a total of 240 empirical data points with at least 120 degrees of freedom; eight parameter values were varied in the search. The best-fit values are shown in Table 1.

Figure 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 detection and 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 completely-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 completely-correct points to be generally under-predicted, probably because our simple model of the stream tracking is not "sticky" enough. That is, a detailed look shows that subjects are more likely than the model to choose the Target Digit if they have chosen the Target Color, as opposed to switching to the Masker or Neither Digit. The result is a tendency to under-predict the completely-correct responses, even though the individual Target and Masker responses are well predicted.

## Conclusions

The EPIC auditory architecture has been extended to include explicit mechanisms for auditory stream perception and tracking. These mechanisms rely on acoustic properties of the speech input itself, in this case, the statistics of the corpus.

We now have a successful model of the two-talker task in which stream tracking based on basic acoustic characteristics of speech accounts very well for data from the two-talker task. Further refinement of the model for the stream tracking process may improve the fit, and there may be ways to reduce the number of free parameters in the detection functions. Work in progress suggests that this model may also scale to three- and four-talker tasks; in fact, the model as described functions in the three- and four-talker cases; the theoretical issue is how to correctly capture the substantially poorer performance produced by having multiple maskers.

In addition, the two-talker model can account for the original Brungart (2001) data if complex suboptimal mixture model strategies are implemented to represent the apparently under-constrained strategies adopted by the subjects. This last result urges that better experimental control of subject strategies, as in our replication experiment, should be used in future experiments on this topic, and that modeling should attempt to explore alternative subject strategies systematically.

## 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, and the U. S. Air Force 711 HW Chief Scientist Seedling program.

## References

Assmann, P.F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *J. Acoust. Soc. Am.* 88, 680–697.

Bolia, R., Nelson, W., Ericson, M., and Simpson, B. (2000). A speech corpus for multitalker communications research. *J. Acoust. Soc. Am.* 107, 1065–1066.

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

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

Cherry, E. Colin (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *J. Acoust. Soc. Am.*. 25 (5): 975–79

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

Meyer, D. E., & Kieras, D. E. (1999). Precis to a practical unified theory of cognition and action: Some lessons from computational modeling of human multiple-task performance. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII*.(pp. 15-88) Cambridge, MA: M.I.T. Press.

Thompson, E.R, Iyer, N., Simpson, B.D., Wakefield, G.H., Kieras, D.E., & Brungart, D.S. (submitted). Payoff matrices and optimal listener strategies in speech-on-speech masking. Submitted to *J. Acoust. Soc. Am.*

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