Chapter 58
A Naturalistic Approach to the Cocktail Party Problem

Ervin R. Hafter, Jing Xia, and Sridhar Kalluri

Abstract While studies of simple acoustic features have provided excellent bases for models of spatial hearing, we are seeking, here, to create a new paradigm for examination of shared attention and scene analysis in natural environments, where the listener is confronted with semantic information from multiple sources. In this new simulation of the cocktail party problem, a subject (S) is questioned, on-line, about information heard in multiple simultaneous stories spoken by different talkers. Questions based on brief passages in the stories are presented visually for manual response. To ensure that responses are based on semantic information rather than just keywords, the latter are replaced in the questions with synonyms. Pay is for performance, and S knows that while a majority of the questions come from a "primary talker," there is potential value in obtaining information from secondary sources. Results, to date, suggest that obtaining semantic information from separate stories is limited by two spatial factors, an exclusive filter that protects information from the attended talker and an inclusive filter that incorporates information from secondary talkers.

E.R. Hafter (✉)
Department of Psychology, University of California,
Berkeley, CA, USA
e-mail: hafter@berkeley.edu

J. Xia
Department of Psychology, University of California,
Berkeley, CA, USA

Starkey Hearing Research Center,
Berkeley, CA, USA

S. Kalluri
Starkey Hearing Research Center,
Berkeley, CA, USA

B.C.J. Moore et al. (eds.), *Basic Aspects of Hearing*,
Advances in Experimental Medicine and Biology,
DOI 10.1007/978-1-4614-1590-9_58, © Springer Science+Business Media New York 2013

1 Introduction

It is impressive to note how long the field of spatial hearing has been fascinated with our ability to single out speech from one talker in the presence of others, an issue beautifully elaborated in Colin Cherry's (1953) prescient discussion of the "cocktail party problem." What began as a way of talking about selective attention has grown into an important tool in auditory scene analysis (for reviews see Treisman 1969; Bronkhorst 2000).

Clearly, monaural cues such as level, pitch, inflection, and language support source segregation in the cocktail party, but much of the research has focused on binaural cues and sound localization. A common technique offered in Cherry (1953) utilizes the so-called Dichotic Listening (DL), where different steams of words are presented to the two ears via headphones. The instruction is to "shadow" the stimulus in the attended ear, that is, to repeat it as it goes along. Questions about features of the talkers or semantic content in both attended and unattended ears are generally saved to the end of a session. Typically, shadowing is accurate, but little is recalled about the unattended stimulus other than such acoustical features as fundamental frequency or prosody. However, some of the unattended words are held in short-term memory, as shown by querying the subject, S, immediately after a sudden cessation (Norman 1969). Further evidence of the processing of speech in the unattended ear comes from the finding that ~30 % of subjects noted the presence of their own name in the unattended ear (Moray 1969; Conway et al. 2001) and that there can be stammers in shadowing when a word in the unattended ear relates to a word in the attended ear (Lewis 1970). A major issue in DL is concerned with the act of shadowing, itself. The problem is that during a study intended to quantify the effects of shared attention between auditory tasks, S is doing another task, preparing for and executing speech production. Could attention to this motor task account for some of the difference in information derived from the shadowed and unshadowed ears? Nevertheless, DL has been important because it examines attention to streams of natural language and because it has pointed out the importance of acoustical segregation in shared attention.

A different issue with the DL paradigm is its relatively slow information rate. This is well addressed by a newer approach, the Coordinate Response Measure or CRM (Brungart 2001; Kidd et al. 2008). There, S listens to multiple, simultaneous talkers saying sentences that are identical in syntax and timing. A typical sentence is, "Ready, (call sign), go to (keyword-1), (keyword-2) now." In a common variation, the keyword lists contain a color and a number. The instruction is to identify the talker who says a given call sign and to repeat the two keywords spoken by that talker. An important advantage of the technique is that it allows study of informational masking through classification of errors based on intrusions of incorrect keywords spoken by unattended talkers (Ihlefeld and Shinn-Cunningham 2008). Another advantage is that by removing variance in syntax and shrinking the lists of possible responses, it creates a very high information rate, providing literally thousands of trials for testing of hypotheses. Like DL, CRM has highlighted the importance of acoustic differences such as the fundamental frequency of competing talkers and their locations in space. However, while the reduced variance afforded by closely matched stimuli increases statistical power, it may also be detrimental to understanding how listeners act in a real-world situation such as the cocktail party, where each talker tells a different story and the listener's task is to extract semantic meaning from the stories. In sharp contrast, the essential distinction in CRM is based on phonetic cues, that is, the sound of the keyword, rather than the gist it conveys. This is clearly illustrated by Woods and Kalluri (2011) who, using speech with CRM-like syntax, albeit with individual talkers staggered in time, report success accounting for identification of nonsense syllable keywords on the basis of audibility of the syllables.

In our new simulated cocktail party, S hears multiple talkers, each presenting a different stream of natural language (story). Reponses are answers to visually presented questions that are based on intrinsic meaning in selected passages, information requiring semantic rather than phonetic processing. Unlike DL, S does not speak (verbal shadowing), instead using a tactile device to answer questions. Attention is directed to the story of the "primary" talker by telling S that a majority of the questions are related to that story. S is paid for performance, so listening to the primary offers the highest payoff, but S is reminded that there is potential value in attending to secondary talkers, that is, eavesdropping. Here, we examine the utility of this paradigm in two experiments that look at spatial bandwidth for exclusion as well as inclusion of information from secondary talkers.

2 Methods

Stimuli are short stories taken from the internet, each lasting about 10 min. *N* simultaneous, but otherwise independent, stories are spoken by *N* different talkers. These are presented through separate loudspeakers placed at head height along a circle surrounding the S. At irregular moments during a session, phrases in stories are used to generate questions shown visually on a screen along with two possible answers. Information on which a question is based in local, in the sense that Ss cannot know the correct answer without attending to the story at the relevant moment. Most important is that the answers reflect semantic information in the relevant phrase rather than the phonetics of keywords. For example, the phrase in one story, "… though I was grateful to be there, I was more grateful that sleep was near…," is tested with the question "What was the narrator looking forward to?" and the two potential answers were "(1) going to bed" and "(2) exploring her new house." In the experiment, questions appear with a mean interval of 20 s.

Stories are read by professional radio announcers in anechoic space prior to sound editing. An inverse filter is used to smooth each story's time-varying sound pressure calculated through a running window. Next, the overall level is set to 70 dB SPL, before final adjustment in level in accord with equal-loudness judgments, by Ss comparing it to a "standard" story. Finally, 100 % performance in undistracted attention is ensured by rejecting questions that elicit an error when tested in the quiet.

A visual screen shows a cartoon of the spatial locations of *N* talkers in the experiment. One of these is colored to indicate the location of the "primary" talker. Questions appear below the cartoon ~1 s after the appropriate information appears in a story; S has 8 s to respond with a two-button box. Pay is for correct answers, regardless of source, but we assume that attention is focused on the "primary" story talker because S knows that it is the source for a majority of the questions. S must answer all questions and we point out that attending to a secondary source, if possible, could increase pay. All subjects go through a training procedure.

3 Experiment 1: Two Talkers

Stories from two female talkers were presented at ±7.5° or ±60° relative to the midline; one was labeled the primary. Ss were 28 young, native English-speaking subjects with normal hearing. In a control condition (A), 12 Ss knew that 100 % of the questions would come from the primary talker. In a shared-attention condition (B), the other 16 Ss knew that a majority of the questions (in actuality, 70 %) would come from the primary talker and the rest from the secondary. On half of the ~10-min sessions, the primary was on the right; on the other half, the primary was on the left.

3.1 Results and Discussion

Figure 58.1 shows performance plotted as a function of the angular distance between primary and secondary talkers. Solid and dashed lines represent answers to questions from primary and secondary talkers, respectively. High performance in the control condition (1A) indicates little interference by the mere presence of a secondary talker. The small effect of spatial separation is not significant ($t(11) = -1.219$, $p = 0.248$). With 30 % of the questions from the secondary (1B), performance on the primary fell, but this was accompanied by better than chance performance on questions from the secondary. A trade-off of this kind suggests division of attention between stories in accord with their perceived importance. Quite interesting is that performance on secondary stories was better when the talkers were closer together

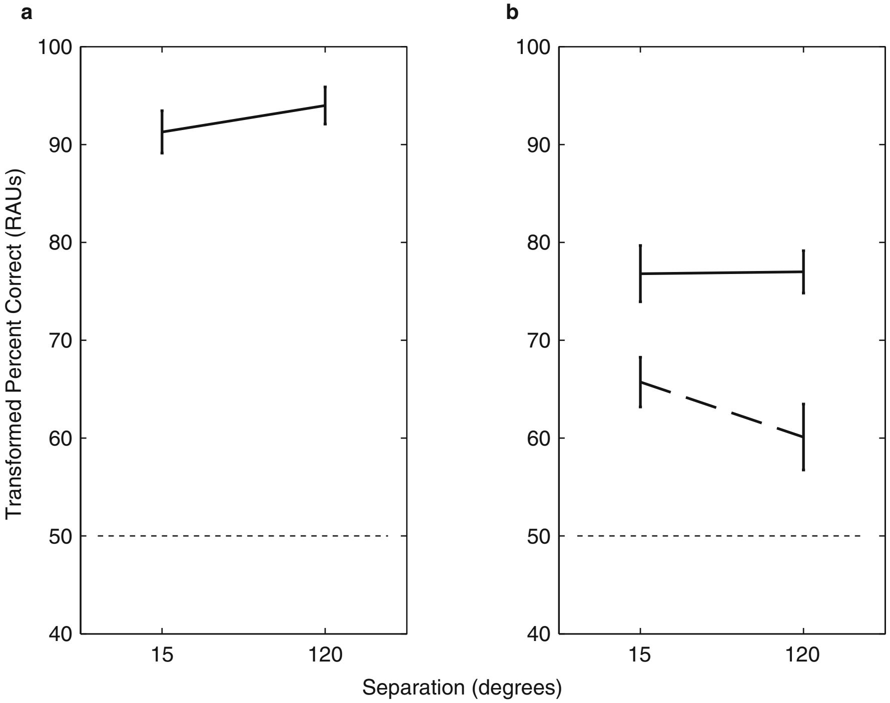

Fig. 58.1 Mean percent correct performance, transformed into rationalized arcsine units (RAU; Studebaker 1985) and plotted as a function of the angular separation between primary and secondary talkers. Chance performance is indicated by *dotted lines*. For Panel **a**, all questions were from the primary talker. For Panel **b**, 70 % of the questions were from a primary talker and 30 % from a secondary. *Solid lines* show performance based on questions from the primary talker. *Dashed lines* show performance based on questions from the secondary talker

($t(15)=2.139$, $p=0.045$), the opposite of what would be expected from a spatial release from masking.

4 Experiment 2: Three Talkers

Stories from three female talkers were presented at 0°, ±7.5°, and ±60° relative to the midline, with the one on the midline labeled “primary.” Ss were a new set of 28 young, native English-speaking subjects with normal hearing. In the control condition (A), 12 Ss knew that 100 % of the questions would come from the primary talker. In the shared-attention condition (B), a different 16 Ss knew that a majority of the questions (actually, 60 %) would come from the primary talker, and the rest would be split evenly between secondary talkers (actually 20 and 20 %).

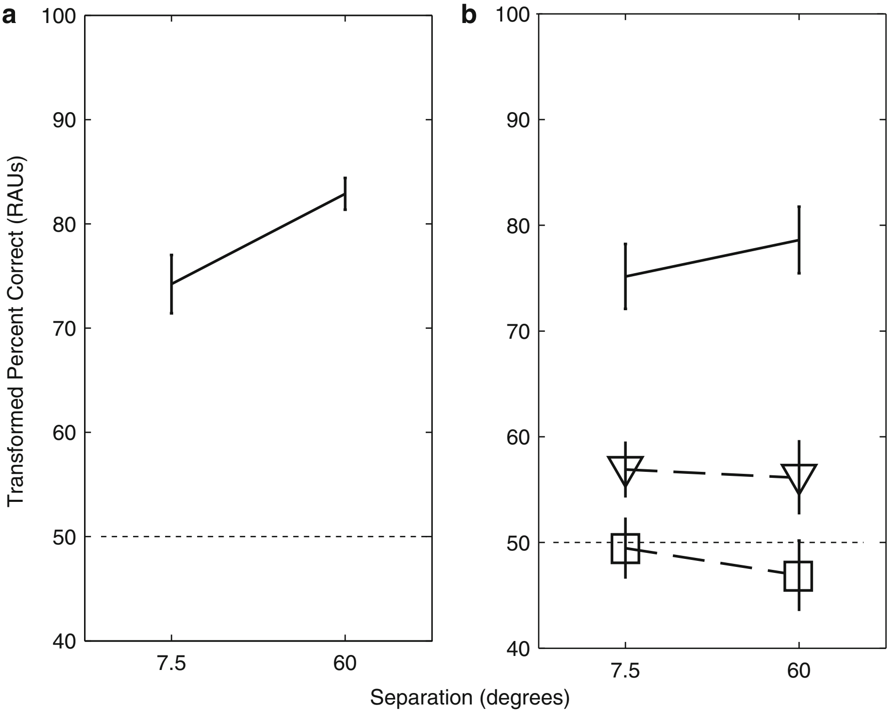

Fig. 58.2 Mean percent correct performance, transformed into rationalized arcsine units (RAU; Studebaker 1985) and plotted as a function of the angular separation between primary talker at 0°. Chance performance is indicated by *dotted lines*. For Panel **a**, all questions were from the primary talker. For Panel **b**, 60 % of the questions were from the primary talker and 20 % from each of the two secondary talkers. *Solid lines* show performance based on questions from a primary talker and *dashed lines* show performance based on the secondary talkers. Additionally, Panel **b** divides secondary performance for the two different secondary talkers; these are plotted by *triangles* and *squares*

4.1 Result and Discussion

Figure 58.2 shows performance plotted as a function of the angular distance from the primary and two secondary talkers. Solid and dashed lines represent answers to questions from primary and secondary talkers, respectively. With distraction from both sides (2A), performance in the control condition shows more interference than in Fig. 58.1a. However, there was a significant spatial release from masking, i.e., better performance with separations of 60° ($t(11) = -2.639$, $p = 0.027$). Perhaps the small but insignificant release seen in Fig. 58.1a reflects a ceiling effect that hid a release from masking. Comparison of Fig. 58.2a, b shows that primary performance was not further compromised in the shared-attention task, though the seeming spatial release for the primary talker in Fig. 58.2b is not significant ($t(15) = -1.157$, $p = 0.266$). Post-session comments from some Ss said that one of the secondary talkers had a particularly "high-pitched" and "animated" voice that made her seem to stand out. Results here for secondary talkers are thus parsed into two dashed lines, one (triangles) for the more distinctive talker and one (squares) for the others. For the squares, performance did not differ from chance, but for the triangles, performance was better than chance (one tailed $t(15)=2.870$, $p=0.0058$).

5 Summary and New Directions

In the SCP, speech flows rapidly and near-continuously as in natural discourse, and, unlike trial-based tasks, the speech is not interrupted by silent periods during which Ss respond. Also, cognitive demands are greater when Ss must maintain a more constant level of attention because they cannot anticipate which portion of the stories will be tested. Such factors are important if we hope to simulate the high-stress acoustic communications often encountered in real life. For future work, we are especially interested in how high stress exaggerates the detrimental effect of hearing loss or aging and whether technological interventions might help overcome the deficits.

Consistent with experiments using DL, our Ss showed strong limitations of semantic processing of simultaneous speech from multiple talkers. While subjects could derive semantic information from both of two talkers, when there were three, performance on one of the secondary talkers was no better than chance. Also consistent with studies using CRM, we found a larger spatial release from masking with two secondary talkers than with one.

What else may be concluded from this approach to the study of auditory attention in more natural environments? Although the paradigm is new and the data are somewhat preliminary, we will point to a few encouraging features. While there is usually a release from masking based on spatial separation, we found the opposite to be true for information from the secondary talker, as seen in Fig. 58.1b. In accord with Best et al. (2006), we propose another way of thinking about spatial bandwidths, one that assumes inclusion of information from a secondary talker when it falls into an attention band focused on the primary talker. Also interesting is that in our three-talker task, Ss were better able to derive information from one of the secondary sources when the (female) talker's voice was more readily distinguished from the primary source.

Results in Experiment 2 indicate that attention was shared between only two of the three talkers, but this limitation will be examined further with the introduction of more distinctiveness between voices. Further steps with this paradigm include direct comparisons between semantic and phonetic cues for attention, the speed with which attention can be switched between talkers, and ways in which hearing impairment and its treatment interact with listening in a real-world cocktail party. We are not surprised that processing of basic acoustic features such as fundamental frequency or location is present when listeners respond to the gist of a spoken message. Our hope is that the ability to examine the latter in our simulated cocktail party will offer new insights into top-down effects in auditory attention.

References

Best V, Gallun FJ, Ihlefeld A, Shinn-Cunningham BG (2006) The influence of spatial separation on divided listening. J Acoust Soc Am 120:1506–1516

Bronkhorst AW (2000) The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Acustica 86:117–128

Brungart DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am 109:1101–1109

Cherry E (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975–979

Conway R, Cowan N, Bunting F (2001) The cocktail party phenomenon revisited: the importance of working memory capacity. Psychon Bull Rev 9:331–335

Ihlefeld A, Shinn-Cunningham BG (2008) Spatial release from energetic and informational masking in a selective speech identification task. J Acoust Soc Am 123:4369–4379

Kidd G Jr, Mason CR, Richards VM, Gallun FJ, Durlach N (2008) Informational masking. In: Yost WA (ed) Auditory perception of sound sources. Springer, New York, pp 143–189

Lewis JL (1970) Semantic processing of unattended messages using dichotic listening. J Exp Psychol 85:225–228

Moray N (1969) Attention in dichotic listening: affective cues and the influence of instructions. Q J Exp Psychol 11:56–60

Norman DA (1969) Memory while shadowing. Q J Exp Psychol 21:85–93

Studebaker GA (1985) A “rationalized” arcsine transform. J Speech Hear Res 28:455–462

Treisman AM (1969) Strategies and models of selective attention. Psychol Rev 76:282–299

Woods W, Kalluri S (2011) Cognitive and energetic factors in complex-scenario listening. In: First international conference on cognitive hearing science for communication. Linkoping, Sweden, pp 19–22