The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer[a)]

Monica L. Hawley[b] and Ruth Y. Litovsky[c]
Hearing Research Center and Department of Biomedical Engineering, Boston University, Boston, MA 02215

John F. Culling
School of Psychology, Cardiff University, P.O. Box 901, Cardiff, CF10 3YG, U.K.

(Received 18 January 2003; revised 14 November 2003; accepted 17 November 2003)

The ''cocktail party problem'' was studied using virtual stimuli whose spatial locations were generated using anechoic head-related impulse responses from the AUDIS database [Blauert *et al.*, J. Acoust. Soc. Am. **103**, 3082 (1998)]. Speech reception thresholds (SRTs) were measured for Harvard IEEE sentences presented from the front in the presence of one, two, or three interfering sources. Four types of interferer were used: (1) other sentences spoken by the same talker, (2) time-reversed sentences of the same talker, (3) speech-spectrum shaped noise, and (4) speech-spectrum shaped noise, modulated by the temporal envelope of the sentences. Each interferer was matched to the spectrum of the target talker. Interferers were placed in several spatial configurations, either coincident with or separated from the target. Binaural advantage was derived by subtracting SRTs from listening with the ''better monaural ear'' from those for binaural listening. For a single interferer, there was a binaural advantage of 2–4 dB for all interferer types. For two or three interferers, the advantage was 2–4 dB for noise and speech-modulated noise, and 6–7 dB for speech and time-reversed speech. These data suggest that the benefit of binaural hearing for speech intelligibility is especially pronounced when there are multiple voiced interferers at different locations from the target, regardless of spatial configuration; measurements with fewer or with other types of interferers can underestimate this benefit.
[DOI: 10.1121/1.1639908]

PACS numbers: 43.66.Pn, 43.71.Gv [PFA] Pages: 833–843

I. INTRODUCTION

In many social situations, listeners receive simultaneous sounds from many sources. Perceptually segregating a single target voice from a competing milieu, so that it can be individually understood, has been termed ''the cocktail-party problem'' (Cherry, 1953). A number of cues and processes that contribute to the solution of the cocktail-party problem have been identified. There are four that are of particular relevance to the current study.

First, spatially separating the target and interferers improves understanding of the target speech. In the free field or ''virtual free-field,'' the effect is known as ''spatial release from masking'' (Plomp and Mimpen, 1981; Bronkhorst and Plomp, 1992; Nilsson *et al.*, 1994; Koehnke and Besing, 1996; Peissig and Kollmeier, 1997; Hawley *et al.*, 1999; Shinn-Cunningham *et al.*, 2001; Litovsky *et al.*, 2002). Spatial release from masking can be regarded as having two components (Durlach, 1963; vom Hövel, 1984; Zurek, 1992; Bronkhorst, 2000): monaural advantage arises directly from improvements in signal-to-noise ratio at the ''best'' ear (BE), which are caused by headshadow; binaural advantage arises from binaural unmasking (BU) of the low-frequency parts of the speech signal, which are largely facilitated by differences in interaural time delay (ITD) between competing sources (Zurek, 1992; Bronkhorst and Plomp, 1988; Durlach, 1963; Culling and Summerfield, 1995; Breebaart *et al.*, 2001a,b,c). This BE+BU account is distinct from that provided by auditory scene analysis (Bregman, 1990), which suggests that spatial release from masking involves the grouping of sound elements from one direction and segregation of that group from elements of interfering sound in different directions. The BE+BU interpretation separates the roles of ITDs and headshadow, while, in the auditory scene analysis, both contribute to the initial determination of sound direction. The present study attempts to differentiate between these accounts by comparing monaural and binaural performance in a variety of listening situations.

Second, understanding of the target speech depends upon the temporal properties of the interfering sound. A speech interferer has a fluctuating frequency spectrum and amplitude envelope. In contrast, speech-shaped noise has a long-term spectrum which matches that of speech, but lacks such modulation (e.g., MacKeith and Coles, 1971; Plomp and Mimpen, 1979; Festen, 1993; Koehnke and Besing, 1996). The effect of the temporal envelope can be investigated using speech-modulated noise, whose temporal envelope is also derived from speech. Dips in the temporal envelope of the interferer are beneficial to understanding of the target voice, presumably due to the transitory improvement

[a)]Portions of this paper were presented at the 137th Meeting of the Acoustical Society of America, March, 1999, British Society of Audiology 2000, and the Meeting of the Association for Research in Otolaryngology, February, 2000.
[b]Current address: Dept. of Otolaryngology, University of Maryland Medical School, 16 S. Eutaw St., Suite 500, Baltimore, MD 21201.
[c]Current address: University of Wisconsin Waisman Center, 1500 Highland Ave., Madison WI 53705.

of signal-to-noise ratio (Festen and Plomp, 1990).

Third, differences in fundamental frequency (F0) between concurrent voices enable listeners to better understand those voices (Brokx and Nooteboom, 1982). Experiments with simultaneous vowels have shown this improvement in understanding to be dependent upon the harmonic structure of the interfering sound, rather than that of the target sound. Lea (1992) found that if one vowel in a pair was noise-excited, detection of the noise-excited and not the harmonic vowel improved. Similarly, Summerfield and Culling (1992) and de Cheveigné *et al.* (1995) found that if one vowel in a pair was inharmonic, identification of this vowel improved (compared to the same-F0 or both-inharmonic cases) and not that of the harmonic one. These data are consistent with the idea that the interfering source is perceptually cancelled (de Cheveigné, 1997). Therefore, when a speech interferer is replaced by speech-modulated noise there can be no advantage from F0 differences. A similar effect may be expected when multiple interferers are presented, although this would depend upon whether the putative canceling mechanism can recursively cancel multiple F0's. We are not aware of any direct perceptual evidence on the effect of multiple F0's among the interferers. These experiments have usually involved stimuli with static fundamental frequencies, but some studies (e.g., Darwin and Culling, 1990; Summerfield, 1992; Culling *et al.*, 1994) have employed modulated F0's, and their results suggest that listeners can exploit instantaneous differences in F0 as proficiently as sustained ones. Thus, when an interfering voice has the same mean F0 (as, for instance, when it is a recording of the same individual), natural modulation of the voice will introduce instantaneous differences in F0 that listeners can exploit. As an illustration of this point, we used Praat to measure the F0 in semitones of each of the voices used in the present study for all of the available recordings of their voices and for every analysis frame. We then calculated the variance of each. The mean instantaneous difference in F0 between randomly selected frames of the same voice can be predicted from the variance sum law; it is $\sqrt{2\sigma_{F0}^2}$, where σ_{F0}^2 is the variance of the voice F0. The values we derived in this way were 5.6 semitones for one voice (known as "DA") and 4.5 semitones for the other ("CW").

Fourth, the interfering speech carries linguistic content, which can be confused with the content of the target voice. This confusion can be regarded as a form of "informational masking." Such masking is a disruption of performance that cannot be accounted for by a simple model of energetic masking (i.e., overlap in the frequencies of the target and interferer). Rather, the masker carries some other information regarding the stimuli and listening conditions, which interferes with perception of the target content (Pollack and Pickett, 1958; Lutfi, 1990; Kidd *et al.*, 1998). Most of what is known about informational masking has been investigated using nonspeech stimuli; however, recent studies using speech as both target and interferer suggest that informational masking might play an important role in the cocktail-party problem (Brungart *et al.*, 2001). When a real-speech interferer is replaced by speech-modulated noise, one may expect some advantage to accrue from the removal of this interference. Thus, this effect tends to oppose the effect of losing F0 differences. In order to differentiate these two effects, one can employ a time-reversed speech interferer, which possesses an F0, but lacks recognizable linguistic content above the phoneme level. A time-reversed speech interferer may show some release of informational masking due to the removal of these components of the interferer information.

In summary, there are four main effects that have been studied with respect to the "cocktail party effect," but their relative importance, especially in multi-talker environments, is poorly understood. While many studies have investigated these four effects individually, few have addressed interactions between them. In addition, few paradigms have been extended towards more complex, ecologically relevant situations in which multiple competing sources occur from various directions. The purpose of the present study was to explore the interaction between the number of interfering sounds, the role of BE and BU when the spatial distribution of interfering sounds are manipulated, and the role of spectral, temporal, and linguistic content. The study thus addresses the problem of understanding the more complex listening situations that are routinely encountered in real life.

The most comprehensive study conducted to date on the effects of multiple sources is that of Peissig and Kollmeier (1997). Peissig and Kollmeier used a virtual sound field presentation of a target source directly ahead and one, two, or three interfering sources, consisting of either speech or speech-shaped noise. In each case, they measured speech reception thresholds (SRTs) using a subjective method with one of the interfering sources in each of 17 different directions. Other interfering sources were in fixed positions. They found that (a) speech produced less interference than noise, and (b) spatial release from masking was smaller with speech than with noise for a single interfering source, but was more robust as additional interfering sources were introduced, such that it showed greater spatial release from masking than noise for three interferers. The results raised some interesting questions.

First, a potential problem with the BE+BU view of spatial unmasking is that models of binaural unmasking appear capable of suppressing only a single interfering source direction, whereas cocktail parties are usually populated by multiple, spatially separated, interfering voices (Peissig and Kollmeier, 1997). The reduction in spatial unmasking that occurred when a single noise interferer was replaced by several suggests support for the BE+BU view. On the other hand, the robustness of spatial unmasking for multiple speech interferers suggests that speech may be an exception to this rule. Peissig and Kollmeier (p. 1668) explain the robustness of spatial unmasking for speech interferers in terms of BE+BU by suggesting that modulation in the interfering sources allows the binaural system to switch between different interferers, cancelling whichever is most energetic at a given point in time. This explanation can account for the robustness of performance with multiple speech interferers, which display independent modulations in their temporal envelopes, compared to performance with multiple continuous-noise interferers, which have no modulation. However,

modulation is one of many physical differences between speech and continuous noise. In order to test Peissig and Kollmeiers explanation, therefore, the present study also used multiple speech-modulated noises. These interferers are identical to the speech-shaped noise interferers except for the critical factor of modulation, which is based upon that of the speech interferers. If independent modulation of interfering sources is the critical factor in producing robust spatial unmasking for multiple interferers, these speech-modulated maskers should produce similarly robust unmasking. In addition, while Peissig and Kollmeier's three-interferer spatial configurations always had at least two interferers in different locations, the present study directly contrasts situations in which three interfering sources are spatially separated with situations in which they are spatially coincident.

Second, the exact role of best-ear listening is unclear in Peissig and Kollmeier's study. Ambiguity occurs for two reasons. One is that they did not contrast best-ear performance with binaural performance. The other is that the condition with three interferers always had fixed-position interferers on both the right and left. In the present study, best-ear performance was measured for all conditions and subtracted from binaural performance to yield a measure of binaural advantage. In addition, conditions were included that contrast three interfering sources in the same hemifield, with a condition in which the interferers are distributed in both hemifields.

A final point of difference between our approach and that of Peissig and Kollmeier is that, in their study, speech intelligibility was measured using a subjective method, whereby subjects adjusted the level of the test sentence to that which corresponded to a subjective judgment of 50% intelligibility. This method was justified on the basis that it enabled data to be collected more rapidly and that a close correlation had been observed in previous studies between objective and subjective SRTs. We preferred to measure speech intelligibility under various interfering conditions using a performance measure.

II. METHODS

A. Listeners

A total of 32 paid participants, 18–36 years old, were recruited from the Boston University community (9 males and 23 females); all were native speakers of English with audiometric thresholds at or below 15 dB HL between 250 and 8000 Hz. None of the listeners were familiar with the sentences used in this study.

B. Conditions

Each listener completed testing in three to six sessions of 1.5 to 2 h each. During these sessions they contributed a single SRT in each of 48 conditions (3 numbers of interferers$\times$4 spatial configurations$\times$4 interferer types). Sixteen listeners provided these SRTs with binaural presentation and 16 with monaural presentation, so the monaural and binaural data sets were collected in exactly the same way but from different sets of listeners. Each listener from the monaural condition could be paired with one from the binaural condition, who completed the different conditions of spatial configuration and interferer type in the same order using the same materials.

Different sets of 16 target-sentence lists were used for data collection using different numbers of interferers. To decrease the effect on the thresholds from using different target-sentence lists and to minimize any order effect, a Latin square design was utilized in which each list was paired with each condition only once and each list occupied a particular place in the order only once. Thus, each listener performed one SRT measurement for each condition and using each list. Each number of interferers had a separate Latin square order using a different set of lists.

C. Simulated anechoic space

Anechoic head-related impulse responses (HRIRs) from the HMSIII acoustic manikin and distributed in the AUDIS collection (Blauert *et al.*, 1998) were used to simulate the spatial locations. The stimulus intended for each position was convolved with the set of HRIRs for the left and right ear. All stimuli for each ear were digitally added and presented to the listener through Sennheiser HD433 headphones while they were seated in a double-walled IAC sound-attenuated booth. For the monaural conditions, only the left headphone was stimulated since this was usually the ''better monaural ear'' defined as the ear with the better signal-to-noise ratio; in the majority of simulated configurations the interfering virtual sound sources were situated to the listeners' right, and were therefore less intense at the left than the right ear.

D. Sound sources

The speech tokens were from the Harvard IEEE corpus (Rothauser, 1969). The recordings[1] used were from two male speakers, each contributing half of the sentences. Six of the longest sentences for each talker were reserved for use as interferers to ensure that all targets were shorter than the interferers. The remaining sentences were made into 64 lists of ten sentences each maintaining a single talker for each list. The interferers paired with the target list were from the same talker.

An interferer of each type (speech; reversed speech; speech-shaped noise; speech-modulated, speech-shaped noise) was made based on each of the six interferer sentences. The noise interferers were filtered to match the long-term spectrum of the speech interferers, calculated for each talker separately. The noise samples were cut to the same length as the matching speech interferer and scaled to the same root-mean-square value. For the speech-modulated, speech-shaped noise, the envelope was extracted from the speech interferer and was used to modulate the noise tokens, giving the same coarse temporal structure as the speech. The envelope of running speech was extracted using a method similar to that described by Festen and Plomp (1990), in which a rectified version of the waveform is low-pass filtered. A first-order Butterworth low-pass filter was used with a 3-dB cutoff at 40 Hz. The time-reversed interferer was

TABLE I. Location of interferers.

Interferer location	Front	Left or distributed on both sides	Distributed on right	Together on right
One interferer	0°	−30°	+60°	+90°
Two interferers	0°, 0°	−30°, +90°	+60°, +90°	+90°, +90°
Three interferers	0°, 0°, 0°	−30°, +60°, +90°	+30°, +60°, +90°	+90°, +90°, +90°

speech reversed in time, end to end. Reversed-speech interferers had the same coarse and fine temporal-spectral structure as speech, but no intelligibility.

E. Sound-source locations

The target location was always at the front (0°). There were conditions with one, two, or three interferers, which were all of the same type in a given condition. Up to three interferers were placed either in the front (0°,0°,0°) distributed on both sides (−30°,60°,90°), distributed on the right side (30°,60°,90°), or from the same location on the right side (90°,90°,90°). See Table I for the full specification of these conditions. The level of each interferer was fixed and so the overall level of the interferers was increased as more interferers were added.

F. SRTs

SRTs were measured using a method similar to that developed by Plomp (1986). Listeners were seated in the sound-attenuated booth in front of a terminal screen. A practice SRT with three interferers for each of the interferer types was given at the start of each session to familiarize the subject with the interferer types and the task.

At the start of each SRT measurement, the level of the target was initially very low. The listener heard the same target sentence and interferer combination repeatedly. Each time the listener pressed the return key the same target sentence and interferer combination was replayed, but with the signal-to-interferer ratio increased by 4 dB. When the listener judged they could hear ''more than half'' of the sentence, they typed in their first transcript. From that point on, an SRT was measured using a one-down/one-up adaptive SRT technique targeting 50% correct speech reception (Levitt, 1971).

Correct speech reception was self-assessed by the listener. After listening to each sentence, the listener typed in their transcript. On pressing the return key, the correct target text was also printed on the screen. Each IEEE sentence had five designated key words and these words were in capital letters in the transcript (e.g., The BIRCH CANOE SLID on the SMOOTH PLANKS.). The listener compared the two transcripts and typed in how many key words were correct. The level of the each trial was raised by 2 dB if two or fewer key words were correct and the level was lowered by 2 dB if three or more key words were correct. The entire transaction was logged in a data file and displayed on the experimenter's computer monitor for verification of scoring reliability. The SRT was determined by averaging the level presented on the last eight trials.[2]

In the speech condition, listeners needed to know the text of the interfering sentences because the interferers were from the same voice as the target sentences and in some conditions all sentences were presented from the front location. The texts of any speech interferers were therefore printed on the screen prior to the start of an SRT measurement. The content, number, and locations of the interferers were fixed throughout the run. In conditions that contained a nonspeech interferer, ''unintelligible'' was printed on the screen.

III. RESULTS AND DISCUSSION

The data were analyzed using the assumption that the observed differences between SRTs for different spatial configurations are the result of two independent processes (best-ear listening and binaural advantage) which are additive in decibels. Using these assumptions, the raw SRTs for monaural and binaural listening were used to calculate three additional statistics.

First, the total advantage of separation for each listener in each condition is determined by subtracting the SRT from a given separated condition from that for the corresponding unseparated condition. The advantage of separation for the binaural condition is called the ''total advantage of separation,'' since it contains advantages due to both head shadow (monaural factor) and binaural processing.

Second, the monaural advantage of separation for each listener in each condition (i.e., best-ear listening) is defined as the difference in SRT between each monaural spatially-separated condition and the corresponding unseparated condition.

Third, the binaural advantage is defined as the part of the total advantage that is not accounted for by the monaural advantage. It is obtained by subtracting (in decibels) the monaural advantage from the total advantage of separation. For this purpose the listeners from the monaural and binaural conditions were paired.[3] This difference measure reflects the binaural processing that occurs in different situations, since it is only present when two ears are available and reflects the benefit over listening with just the better monaural ear.

All five measures are discussed below, but statistical analysis is reserved for the derived monaural and binaural advantages of separation. This statistical choice avoided reanalyzing the same data in different ways. The decision to analyze the component advantage of separation is supported by Figs. 1–3, which show that the component effects produce a clearer, more easily interpreted, pattern than the raw data. Scheffé *post hoc* contrasts between means were performed on all significant results from each ANOVA, using

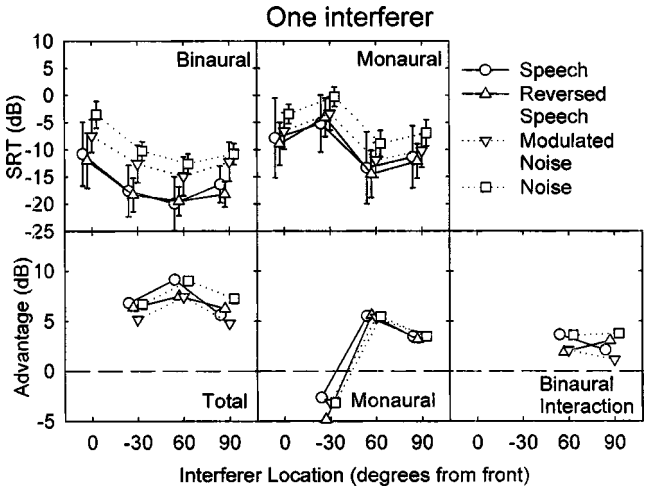

FIG. 1. SRTs and advantages of spatial separation for a single interfering source. The top two panels show means of the raw SRTs, with standard error bars, using two ears (binaural) and using only the left ear (monaural). The lower three panels show the advantage of spatial separation derived by subtracting away the SRT for the nonseparated condition using two ears (total) and using only the left ear (monaural). The binaural advantage is the difference between the total and the monaural advantage.

α=0.05. *Post hoc* one-sample *t*-tests were used to demonstrate deviation of spatial advantages from zero. Bonferroni correction was not used for these *t*-tests because they were intended to identify which spatial advantages differed from zero rather than whether any of them differed.

A. One interferer

The results for a single interferer are shown in Fig. 1.

1. Raw SRTs

For the binaural condition, the SRTs decrease as the interferer location is separated from 0°, the location of the target, regardless of interferer type. The effect of interferer type is seen as an overall shift in the SRTs. The lower SRTs for speech and reversed speech probably reflect the exploitation of differences in F0 between target and interferer (Brokx and Nooteboom, 1982), which may have enabled the interferer to be cancelled (de Cheveigné, 1997).

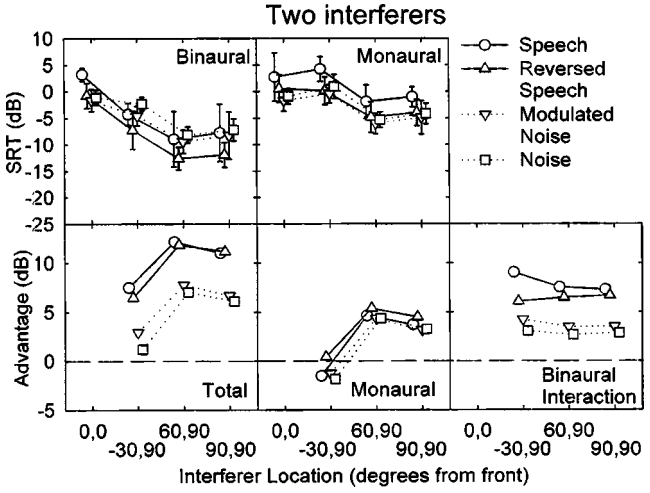

FIG. 2. As in Fig. 1, but for two interfering sources.

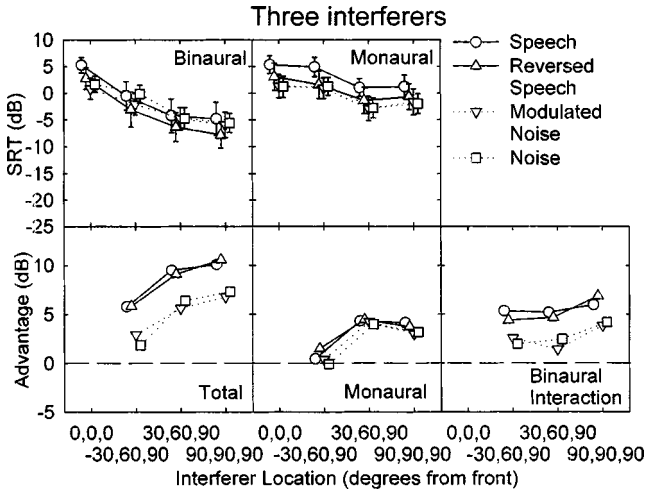

FIG. 3. As in Figs. 1 and 2, but for three interfering sources.

For the monaural condition, the SRTs increase for an interferer at −30° and then fall for interferers at 60° and 90°. The increase at −30° is expected, since, for this interferer location, the left ear is on the same side as the interferer, and so the SNR is not favorable. The ordering of the interferer types is the same as was seen for the binaural condition. However, the difference between the modulated noise interferer and the speech (1.4 dB) and reversed speech (1.9 dB) interferers is not as marked as it was for the binaural condition

2. Advantages of separation

There is a large total advantage (about 6 dB) when moving the interferer 30°, 60°, or 90° from the target location. A similar effect of location is observed using each interferer.

The monaural advantage of separation was negative for an interferer at −30° due to the unfavorable SNR, but +6 and +3 dB for interferers at 60° and 90°, respectively. A two-factor ANOVA (4 interferer types×3 interferer locations) revealed a significant effect of interferer location [$F(2,30)=80$, $p<0.0001$], but not interferer type and no interaction. *Post hoc* analysis of interferer location revealed that all levels of interferer location differed from each other [$F(2,30)=145,87,7.5$]. The monaural advantages generally differed significantly from zero [$t(15)>2.9$], except for speech interferers in the −30° and 90° locations. Figure 1 shows that these means were similar to those for the other interferer types and the lack of significance can be attributed to greater variance. Advantage of separation was negative for the interferer at −30°, and positive for 60° and 90°.

The binaural advantage for the interferer at −30° was not calculated since the monaural measurement was not made from the ear with the best signal-to-noise ratio and thus the difference between binaural and monaural measurement includes more than binaural processing in this case. A two-factor ANOVA for the remaining data (4 interferer types×2 interferer locations) revealed no significant effects. The majority of binaural advantages were significantly greater than zero [$t(15)>2.2$] at 60° and 90° and are in the range of 2–4

dB, consistent with previous reports. The exceptions were 90° for speech and modulated noise and 60° for reversed speech.

B. Two interferers

The results for two interferers are shown in Fig. 2.

1. Raw SRTs

The binaural SRTs for two interferers also decrease as the interferers are separated from the target location. However, the ordering of the interferer types is different from that seen in the one-interferer case; the speech interferer now gives among the highest SRTs, while the reversed speech remains the lowest. The relative increase in SRTs against the speech interferers compared to the one-interferer case may reflect an increase in linguistic interference, while the reversed speech retains an advantage due to exploitation of F0 differences. The SRT for the speech interferers is higher than that for the reversed speech interferers by an average of 3.7 dB across locations.

The monaural SRTs were lower than the unseparated condition for the (60°,90°) and (90°,90°) conditions. SRTs in the (−30°,90°) did not differ from the unseparated condition, presumably because the beneficial effect of headshadow is removed when interfering sources are placed on both sides. SRTs for the speech interferer were higher than for the other interferer types. This result contrasts with the single-interferer case, in which speech and reversed speech gave the lowest SRTs.

2. Advantages of separation

The total advantage of separation is up to 12 dB for speech and reversed-speech interferers. Speech and reversed speech had a larger total advantage of separation than modulated noise and noise interferers. This advantage of separation was greater than observed with only a single interferer. The (60°,90°) and (90°,90°) conditions gave a large advantage and the (−30°,90°) a smaller one.

The monaural advantage of separation was subjected to a two-factor ANOVA (4 interferer types×3 interferer locations), which revealed a significant effect of interferer location [$F(2,30)=136$, $p<0.0001$], but not interferer type and no interaction. *Post hoc* comparisons revealed that all levels of interferer location differed [$F(2,30)=238,162,7$]. The monaural advantage for the (−30°,90°) location was not significantly different from zero for the speech and reversed speech interferers [$t(15)<1.6$ in each case], but was significantly *below* zero for the two noise-based interferers [$t(15)>2.3$ in each case]. For all other conditions the monaural advantages were significantly above zero [$t(15)>3$ in each case].

The binaural advantage of separation was subjected to a two-factor ANOVA (4 interferer types×3 interferer locations) of binaural advantages for the two-interferer conditions. The ANOVA revealed a significant effect of interferer type [$F(3,45)=7.1$, $p<0.001$], but no effect of location or interaction. *Post hoc* comparisons of interferer type revealed that speech gave greater binaural advantage than noise and modulated noise [$F(3,45)=17,11$]. All conditions gave mean binaural advantages that were significantly above zero [$t(15)>3.6$ in each case].

The significant effect of interferer type confirms that the origin of the changes in the ordering of the interferer types when a second interferer is introduced result from changes in the effectiveness of binaural processing. With more than one interferer the binaural system is more effective at alleviating interference from a speech or reversed speech source than noise or modulated noise. This effect was replicated in the three-interferer conditions.

C. Three interferers

The results for three interferers are shown in Fig. 3.

1. Raw SRTs

The binaural SRTs decrease as the interferers are separated from the target location. The ordering of the interferer types is similar to that seen for the two-interferer conditions.

The monaural SRTs were lower than the unseparated condition for the (30°,60°,90°) and (90°,90°,90°) interferers, but, as in the two-interferer case monaural SRTs at (−30°,60°,90°), with interferers on both sides, were similar to the unseparated case. SRTs for the speech interferer were higher than for the other interferer types.

2. Advantages of separation

The total advantage of separation is up to 10 dB for speech and reversed speech interferers. As in the two-interferer case, the speech and reversed speech interferers gave a larger total advantage of separation than the two noise-based interferers. Conditions (30°,60°,90°) and (90°,90°,90°) gave a large and similar advantage, while (−30°,60°,90°) gave a smaller advantage.

The monaural advantage was subjected to a two factor ANOVA (4 interferer types×3 interferer locations). The ANOVA revealed a significant effect of interferer location [$F(2,30)=243$, $p<0.0001$] on monaural advantage, but not of interferer type and no interaction. *Post hoc* comparisons of different locations revealed only that the (−30°,60°,90°) condition differed significantly from the (30°,60°,90°) and (90°,90°,90°) conditions [$F(2,30)=153,97$]. Monaural advantage for the (−30°,60°,90°) location was not significantly different from zero for any interferer type, whereas the monaural advantage in all other conditions differed significantly from zero [$t(15)>4.2$ in each case].

A two-factor ANOVA (4 interferer types×3 interferer locations) for binaural advantage revealed a significant effect of interferer type [$F(3,45)=7.7$, $p<0.0005$] and interferer location [$F(2,30)=11.4$, $p<0.0005$], but no interaction. *Post hoc* comparisons of interferer type showed that speech and reversed speech gave consistently larger binaural advantages than did modulated noise or noise interferers [$F(3,45)=13.6,11.3,11.8,9.7$]. Comparisons between interferer locations revealed that binaural advantage in the (90°,90°,90°) condition was significantly different from the other two [$F(2,30)=15.9,18.3$]. Interferer configurations (30°,60°,90°) and (−30°,60°,90°) were not different. How-

ever, binaural advantages in every condition except modulated noise interferers at (30,60,90) were significantly greater than zero [$t(15) > 3.7$].

IV. GENERAL DISCUSSION

The experiment was intended to bridge the gap in complexity from the relatively simple situations that have been extensively researched in previous studies to more complex and realistic listening situations. This was achieved by measuring SRT's both monaurally and binaurally against one, two, or three interferers in four different spatial configurations. In each of these conditions, the interferer was either speech, reversed speech, speech-shaped noise, or speech-modulated noise. The data analysis involved a separation between monaural and binaural effects, making use of the assumption that overall performance is the sum of the effects of best-ear advantage and binaural advantage. The fact that the resulting ''advantage'' measures produce a much simpler and clearer projection of the data than the raw SRTs suggests that this analysis is appropriate. However, the advantages observed for multiple voice-based interferers were larger than can be accounted for by models of binaural unmasking (Zurek, 1992). The patterns of SRTs and spatial advantages revealed a number of effects.

A. Monaural advantage

Monaural listening through the left ear was sufficient to produce an advantage of spatial separation when the interferer(s) all occurred on the right, due to the effect of head-shadow. If one assumes that this advantage arises purely from best-ear listening, the size of this effect is predictable from the acoustics associated with sound waves reaching the head and the importance of the frequencies involved for speech understanding (Zurek, 1992).

The monaural spatial advantage disappeared once multiple interfering sources were spatially distributed on the right and left, since the signal-to-noise ratio for the target presented from front was now reduced by the interferer on the left. Although unsurprising, this effect has important practical implications, since it implies that head-shadow plays a minor role in commonly encountered listening situations when competing sources are distributed in both hemifields. The result also clarifies those of Peissig and Kollmeier (1997). In their study, the fixed sources were always on either side of the head when three interfering sources were used, so their results with three interferers should probably be interpreted as including only effects of binaural advantage.

B. Binaural advantage

When both ears were available to the listener and the target sound was spatially separated from the interferers, a binaural advantage occurred. This advantage has been modeled on the basis of the strength of binaural unmasking at different frequencies and the importance of those frequencies to speech understanding (Levitt and Rabiner, 1967; vom Hövel, 1984; Zurek, 1992). For a single noise interferer, the binaural advantage is predicted to be 3 dB when the spatial separation is 90°. From the present data set, the prediction appears to hold for all interferer types in the one-interferer case. However, for multiple interferers it seems sufficient to explain the data only for noise-based interferers (Fig. 4).

In contrast to the monaural advantage discussed above, the binaural advantage was robust in all spatial configurations, whether competing sources were spatially coincident, distributed across locations, in the same hemifield, or on both the right and left. The role of binaural advantage in complex listening situations is probably greater, therefore, than monaural head-shadow. The fact that binaural advantage was robust against spatially distributed interferers is surprising in the context of models of binaural unmasking that depend upon a highly coherent masker. Multiple interferers with different delays will have reduced coherence and so might be expected to have markedly reduced binaural unmasking. For instance, Durlach's (1963) equalization-cancellation model can cancel an interferer with a specified interaural time delay, but if multiple interferers have multiple delays, one would expect it to be able to cancel only one of them. A follow-up study, Culling *et al.* (2003) analyzes this effect in greater detail and shows that models of binaural unmasking are more robust to reduced coherence than one might expect. On the other hand, it seems unlikely that binaural unmasking can account for all the spatial advantages observed with speech interferers (see Sec. IV E).

C. Dip listening

Another well-known effect is that of ''dip-listening'' where listeners exploit transitory reductions in the power of the interferer in order to pick up information from the target (Festen and Plomp, 1990). Dip listening can be most clearly seen in the current data set through the differences between noise and modulated noise interferers; only the latter gives the listener the opportunity to listen in the dips and thereby achieve a lower SRT. There is a strong effect of dip listening in the single-interferer case of 2–3 dB. As additional interferers are added, the effect is attenuated, because the dips in one interferer become filled in by the energy of another asynchronously modulated interferer (Bronkhorst and Plomp, 1992). In the three-interferer case the SRTs are indistinguishable. Dip listening also, therefore, plays only a minor role in complex listening environments with multiple, relatively distant source like those simulated here.

D. F0 differences

SRTs were lower for single interfering sources that were voiced (speech and reversed-speech) than for ones that were noise-based (noise and modulated noise). The advantage of voiced interferers is seen in the difference between the overall SRTs for these conditions (Fig. 1). In contrast, when two and then three interferers were tested, this difference was not observed. The results may be best understood in terms of a cancellation mechanism that relies on F0 differences (e.g., de Cheveigné 1997), although an informational masking account is also possible (see Sec. IV F). The F0-difference interpretation can account for the fact that the effect is limited to the single-interferer situation, since multiple voices, with

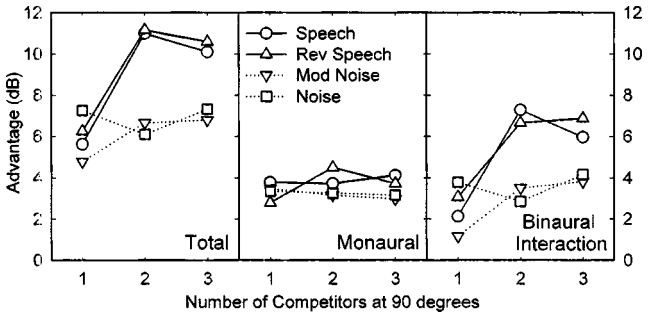

FIG. 4. Spatial advantage as a function of number of interfering sounds at 90° for each interferer type. The left-hand panel shows the total advantage, the middle panel shows the advantage when using only the best ear (monaural advantage), and the right-hand panel shows the difference between these two, attributable to binaural interaction.

multiple F0s, would require multiple rounds of cancellation. It seems likely that the system is incapable of making more than one such cancellation, but even if multiple rounds are possible, the target sound would be progressively distorted by the comb-filtering effects that accompany the cancellation.

SRTs for voiced interferers were substantially higher when there were two or three voiced interferers. The distinction between voiced and noise-based interferers is especially evident in Fig. 5 where the increase in SRT resulting from additional interferers (as large as 14 dB) is compared with the expected increase based on the increased energy in the interferers (3–6 dB). For noise-based interferers, the incremental change in SRT as the second, and then the third, interferers were added can be explained by the increased energy in the interferers (see thick horizontal bars in Fig. 5). In contrast, for the speech and reversed-speech interferers, the incremental change in SRT with added interferers is substantially larger.

E. Voicing/spatial advantage interaction

The interferer type interacted with spatial separation; the effect of spatial separation of interferers from the target was greater when either of the two voiced interferers was used (though only in the two- and three-interferer cases). A similar effect was recently reported by Noble and Perret (2002) and is consistent with the results of Peissig and Kollmeier (1997), who also found that spatial unmasking was more robust with multiple speech interferers than with multiple noise interferers. For the latter binaural advantage is limited to about 3 dB (Bronkhorst and Plomp, 1992). However, the present result is inconsistent with Peissig and Kollmeier's suggested explanation in terms of suppressing different interfering sources at different times. If this explanation were correct, then the speech modulated noise used in the present experiments would also have permitted spatial advantage to be robust against multiple interferers. We have no alternative explanation. However, it is noteworthy that it was a substantial effect (≈3 dB) and was only observed in the most complex and realistic of listening situations. It is therefore worthy of further investigation.

Other than Peissig and Kollmeier's results, the nearest precedents for the effect in the literature are the rather small

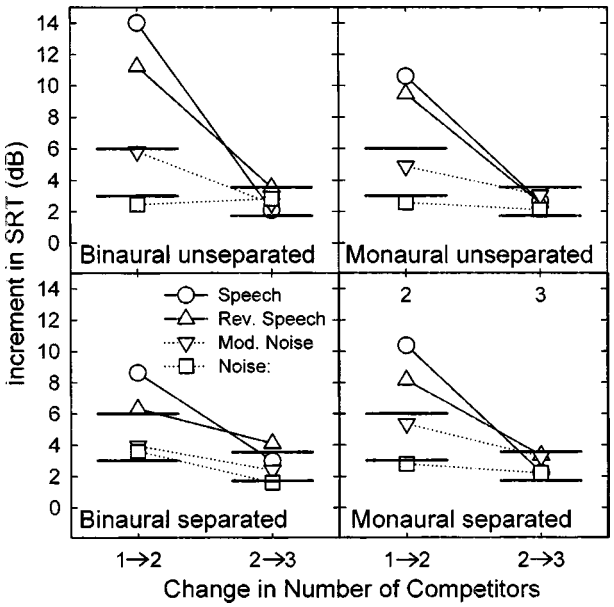

FIG. 5. Change in SRT as each additional interfering sound is added as a function of the total number of interfering sounds. The lower set of horizontal bars shows the expected average increment in threshold when there are random phase relationships between the components of the existing and the added interferers (i.e., there is a 3-dB increase in expected SRT as a result of a 3-dB increase in total masker level when a second interferer is added). The upper set of horizontal bars represent the maximum expected increase in thresholds if the components of the added interferer are perfectly in phase with those of the existing interferers (e.g., adding a second interferer causes a 6-dB increase in total masker level and in SRT).

interactions in "double-vowel" identification reported by Shackleton *et al.* (1994), and later corroborated by Culling *et al.* (1994). Shackleton *et al.* used a design in which the dependent variable was the percentage of simultaneous, synthesized vowel pairs for which listeners correctly identified both vowels. They found an interaction between the presence of a difference in F0 and the presence of a difference in ITD, such that percent correct was higher when two vowels differed in both these parameters. Culling *et al.* used a method more similar to the measurement of SRT in that the threshold for correct vowel identification was measured against a single competing vowel, which varied somewhat from trial to trial. They found a similarly small effect. The effects described in these studies seem to differ in magnitude from the one found here, but the one found here was only evident using multiple interferers. It may be that there is a small interaction for a single interferer and that that interaction grows as more interferers are introduced.

Curiously, the F0 effect, if one defines it as the difference between the reversed-speech and modulated noise conditions, also appears to interact with monaural versus binaural presentation. This interaction may be seen in the one-interferer case (compare the erect and inverted triangles on the top two panels of Fig. 1) where effects of F0 difference are large; they are consistently larger in the binaural than in the monaural condition, regardless of spatial configuration. The reasons for this effect remain obscure.

F. Informational masking

"Informational masking" is disruption to the processing of a target sound without energetically masking it. For in-

stance, the masker may be in different frequency channels or presented to a different ear, and so does not prevent detection of the target. If the content of the interferer is similar to that of the target, the two can become confused and tasks such as target identification can be disrupted (Pollack and Pickett, 1958; Lutfi, 1990; Kidd *et al.*, 1998; Brungart *et al.*, 2001).

One condition for informational masking is that the target content is above detection threshold. In the present study, all the interferers had the same long-term spectrum as the target speech; hence there was always some overlap in the energies of the target and interferers, and energetic masking was always present. Signs of informational masking must, therefore, manifest themselves as an excess masking in particular conditions. In addition, one should expect more informational masking where the overlap in spectro-temporal pattern is relatively incomplete. In the one-interferer cases, the modulated noise contained periods during which the energy in the interferer was significantly reduced. With additional interferers the overlap was more constant for both noise and modulated noise. In contrast, the speech and reversed speech naturally contain dynamic variations in spectrum, and are therefore unlikely to completely overlap in spectrum with the targets at a given instant in time. Thus, one would expect the two voiced interferers to be more likely to display informational masking effects. It is possible that informational masking can be seen in two aspects of the present data set.

First, when multiple interferers were present there was a consistent 2-dB difference between the speech and reversed-speech SRTs. This effect may represent informational masking at the linguistic level, and this suggestion is supported by the fact that when no binaural unmasking is possible (monaural and nonspatially separated configurations) multiple speech interferers produce the highest SRTs of all interferer types. The underlying mechanism is, at this point, largely a matter of speculation. Words from an interfering voice may be intruding into the perceived target sentence. The grammatical and semantic information in the masking stimuli may also be automatically recruiting the listener's attentional resources and reducing the depth of processing that can be applied to the target voice. Using the current paradigm, it is not possible to differentiate the effects of intrusion and attentional distraction; although the listeners transcripts were recorded, the listeners were aware of the content of the interfering sentences, and would have been unlikely to include in their transcripts words that they knew were intrusions. Furthermore, for the two-interferer case, there is evidence that a component of the binaural interaction is a release of this form of informational masking, since there is greater advantage for speech than for reversed speech

Second, the added interference produced by multiple speech and reversed-speech interferers may reflect increased informational masking. This effect was considered above with respect to the effect of F0 differences, but an increase in informational masking may provide an alternative explanation. This account relies upon the reversed speech acting as an informational masker at a lower linguistic level. It is possible, for instance, that reversed speech can recruit attentional and cognitive resources that noise-based interferers do not because they engage phonetic and lexical processing resources even if they do not yield meaningful lexical units for higher levels of processing. It is possible that reversed-speech engages many of these processes by activating an initial mechanism that searches for sources containing language-based information.

Informational masking perhaps offers a more coherent account of the interaction between voicing of the interferers and spatial separation because informational masking can be released by the spatial separation (Brungart *et al.*, 2001). However, there are other problems with this account. First, the effect of F0 differences is very well established and reduction of this effect must account for at least some of the increase in SRTs that occurs as a second voice-based interferer is introduced. Second, in the multiple-interferer cases, thresholds in the reversed speech condition are no worse than for the two noise-based interferers. Thus, there is no obvious evidence of an additional masking effect for the reversed speech interferer with respect to other interferer types, only with respect to the single-interferer case.

It should be possible to differentiate between aspects of the current data set that can be explained by informational masking and those that can be explained by F0 differences by repeating elements of the experiment using an additional masker type. Shannon *et al.* (1995) showed that very accurate speech recognition could be achieved by listening to a noise that was modulated within a discrete number of frequency channels by the speech envelope within those channels. If a sufficiently small number of frequency channels is used, such speech lacks an F0, but should still possess many of the attributes necessary to cause both types of informational masking considered above. If such interferers show a pattern of thresholds similar to the speech interferers in the current study, then this finding would strongly support the informational masking account.

Although the effects that can unambiguously be attributed to informational masking in the current data set are not very large, it should be noted that some aspects of the SRT paradigm we employed were not optimal for the observation of informational masking. In particular, the use of a fixed interfering sentence or set of sentences throughout a given SRT measurement and the presentation of the text of the interfering messages at the beginning of the measurement will have substantially reduced the uncertainty about the interferer content. Uncertainty about the interferer is supposed to be a vital aspect of informational masking, so this methodology may have served to reduce the size of the effects observed.

V. CONCLUSIONS

The results obtained in the present study suggest that listeners' ability to function in complex environments, such as a cocktail party, not only depends the type, number, and location of interfering sounds, but also on interactions between these factors. A number of the effects observed in the current study are well established, but the interactions between interferer types and spatial configuration have not been previously reported and not always easily explained. Further research is necessary to explore and account for these phenomena. However, from a practical point of view

the most significant finding is that in complex listening environments the effects of binaural advantage and fundamental frequency difference seem to be interdependent, while the role of dip listening is reduced. These findings both clarify our understanding of the cocktail party problem and its solution, and should inform our choice of appropriate stimuli for clinical testing of binaural processing.

ACKNOWLEDGMENTS

This work was supported by UK MRC and NIH-NIDCD Grant Nos. R01-DC00100 and R29-DC03083.

[1]The recordings were obtained from Patrick Zurek, Ph.D., of the Massachusetts Institute of Technology.

[2]The SRT was the mean of the levels on trials 4, through to 11, inclusive. Although no 11th trial existed, the level at which it would have been presented was determined by the result of the tenth trial.

[3]Subjects from the two conditions were paired according to the order in which they took part; subject 1 from the monaural condition was paired with subject 1 from the binaural condition. These pairs experienced the within-subjects conditions and materials in the same order but using a different number of ears. Alternative pairings would have no effect on the means of the resulting differences, but would have resulted in larger variances due to variations in list difficulty.

Blauert, J., Brueggen, M., Bronkhorst, A. W., Drullman, R., Reynaud, G., Pellieux, L., Krebber, W., and Sottek, R. (**1998**). ''The AUDIS catalog of human HRTFs,'' J. Acoust. Soc. Am. **103**, 3082.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001a**). ''Binaural processing model based on contralateral inhibition. I. Model structure,'' J. Acoust. Soc. Am. **110**, 1074–1089.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001b**). ''Binaural processing model based on contralateral inhibition. II. Dependence on spectral parameters,'' J. Acoust. Soc. Am. **110**, 1089–1105.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001c**). ''Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters,'' J. Acoust. Soc. Am. **110**, 1105–1118.

Bregman, A. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge).

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

Bronkhorst, A. W., and Plomp, R. (**1988**). ''The effect of head-induced interaural time and level differences on speech intelligibility in noise,'' J. Acoust. Soc. Am. **83**, 1508–1516.

Bronkhorst, A. W., and Plomp, R. (**1992**). ''Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing,'' J. Acoust. Soc. Am. **92**, 3132–3139.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). ''Intonation and the perceptual separation of simultaneous voices'' J. Phonetics **10**, 23–36.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). ''Informational and energetic masking effects in the perception of two simultaneous talkers,'' J. Acoust. Soc. Am. **109**, 1101–1109.

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

Culling, J. F., and Summerfield, Q. (**1995**). ''Perceptual segregation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay,'' J. Acoust. Soc. Am. **98**, 785–797.

Culling, J. F., Summerfield, Q., and Marshall, D. H. (**1994**). ''Effects of simulated reverberation on binaural cues and fundamental frequency differences for separating concurrent vowels,'' Speech Commun. **14**, 71–96.

Culling, J. F., Hawley, M. L., and Litovsky, R. Y. (**2003**). ''The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources,'' submitted to J. Acoust. Soc. Am. .

Darwin, C. J., and Culling, J. F. (**1990**). ''Speech perception seen through the ear,'' Speech Commun. **9**, 469–476.

de Cheveigné, A. (**1997**). ''Concurrent vowel identification. III. A neural model of harmonic interference cancellation,'' J. Acoust. Soc. Am. **101**, 2857–2865.

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). ''Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation,'' J. Acoust. Soc. Am. **97**, 3736–3748.

Durlach, N. I. (**1963**). ''Equalization and cancellation model of binaural unmasking-level differences,'' J. Acoust. Soc. Am. **35**, 1206–1218.

Festen, J. M. (**1993**). ''Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice,'' J. Acoust. Soc. Am. **94**, 1295–1300.

Festen, J. M., and Plomp, R. (**1990**). ''Effects of fluctuating noise and interfering speech on the speech-reception SRT for impaired and normal hearing,'' J. Acoust. Soc. Am. **88**, 1725–1736.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**). ''Speech intelligibility and localization in complex environments,'' J. Acoust. Soc. Am. **105**, 3436–3448.

Kidd, Jr., G., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**). ''Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns,'' J. Acoust. Soc. Am. **104**, 422–431.

Koehnke, J., and Besing, J. M. (**1996**). ''A procedure note for testing speech intelligibility in a virtual listening environment,'' Ear Hear. **17**, 211–217.

Lea, A. (**1992**). ''Auditory models of vowel perception,'' Ph.D. thesis, Nottingham (unpublished).

Levitt, H. (**1971**). ''Transformed up-down methods in psychoacoustics,'' J. Acoust. Soc. Am. **49**, 467–477.

Levitt, H., and Rabiner, L. R. (**1967**). ''Predicting binaural gain in intelligibility and release from masking for speech,'' J. Acoust. Soc. Am. **42**, 620–629.

Litovsky, R. Y., Fligor, B., and Tramo, M. (**2002**). ''Functional role of the human inferior colliculus in binaural hearing,'' Hear. Res. **165**, 177–188.

Lutfi, R. A. (**1990**). ''How much masking is informational masking?'' J. Acoust. Soc. Am. **88**, 2607–2610.

MacKeith, N. W., and Coles, R. R. A. (**1971**). ''Binaural advantages in hearing speech,'' J. Laryngol. Otol. **85**, 213–232.

Noble, W., and Perret, S. (**2002**). ''Hearing speech against spatially separate competing speech versus competing noise,'' Percept. Psychophys. **64**, 1325–1336.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). ''Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise,'' J. Acoust. Soc. Am. **95**, 1085–1099.

Peissig, J., and Kollmeier, B. (**1997**). ''Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners,'' J. Acoust. Soc. Am. **101**, 1660–1670.

Plomp, R. (**1986**). ''A signal-to-noise ratio method for the speech-reception SRT of the hearing impaired,'' J. Speech Hear. Res. **29**, 146–154.

Plomp, R., and Mimpen, A. M. (**1979**). ''Speech-reception threshold for sentences as a function of age and noise level,'' J. Acoust. Soc. Am. **66**, 1333–1342.

Plomp, R., and Mimpen, A. M. (**1981**). ''Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech-reception threshold for sentences,'' Acustica **48**, 325–328.

Pollack, I., and Pickett, J. M. (**1958**). ''Stereophonic listening and speech intelligibility against voice babbles,'' J. Acoust. Soc. Am. **30**, 131–133.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). ''I.E.E.E. recommended practice for speech quality measurements,'' IEEE Trans. Audio Electroacoust. **17**, 227–246.

Shackleton, T. M., Meddis, R., and Hewitt, M. J. (**1994**). ''The role of binaural and fundamental frequency differences in the identification of concurrently presented vowels,'' Q. J. Exp. Psychol. A **47A**, 545–563.

Shannon, R. V., Zeng, F.-G., Kamath, J., Wygonski, J., and Ekelid, M. (**1995**). ''Speech Recognition with Primarily Temporal Cues,'' Science **270**, 303–304.

Shinn-Cunningham, B. G., Schickler, J., Kopco, N., and Litovsky, R. Y. (**2001**). ''Spatial unmasking of nearby speech sources in a simulated anechoic environment,'' J. Acoust. Soc. Am. **110**, 1118–1129.

Summerfield, Q. (**1992**). ''Roles of harmonicity and coherent frequency modulation in auditory grouping,'' in *The Auditory Processing of Speech*, edited by M. E. H. Schouten (Mouton de Gruyter, New York).

Summerfield, Q., and Culling, J. F. (**1992**). ''Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency,'' J. Acoust. Soc. Am. **92**, 2317(A).

vom Hövel, H. (**1984**). ''Zur Bedeutung der bertragungseigenschaften des Außenohres sowie des binauralen Hörsystems bei gestörter Sprachübertragung,'' Ing. Dissertation, RWTH Aachen 1984; as cited by Bronkhorst, A. (**2000**). Acustica **86**, 117–128.

Zurek, P. M. (**1992**). ''Binaural advantages and directional effects in speech intelligibility,'' in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, Boston), pp. 255–276.