The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources

John F. Culling[a)]
School of Psychology, Cardiff University, P.O. Box 901, Cardiff, CF10 3YG, United Kingdom

Monica L. Hawley[b)] and Ruth Y. Litovsky[c)]
Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston. Massachusetts 02215

(Received 11 July 2003; revised 13 May 2004; accepted 18 May 2004)

Three experiments investigated the roles of interaural time differences (ITDs) and level differences (ILDs) in spatial unmasking in multi-source environments. In experiment 1, speech reception thresholds (SRTs) were measured in virtual-acoustic simulations of an anechoic environment with three interfering sound sources of either speech or noise. The target source lay directly ahead, while three interfering sources were (1) all at the target's location (0°,0°,0°), (2) at locations distributed across both hemifields (−30°,60°,90°), (3) at locations in the same hemifield (30°,60°,90°), or (4) co-located in one hemifield (90°,90°,90°). Sounds were convolved with head-related impulse responses (HRIRs) that were manipulated to remove individual binaural cues. Three conditions used HRIRs with (1) both ILDs and ITDs, (2) only ILDs, and (3) only ITDs. The ITD-only condition produced the same pattern of results across spatial configurations as the combined cues, but with smaller differences between spatial configurations. The ILD-only condition yielded similar SRTs for the (−30°,60°,90°) and (0°,0°,0°) configurations, as expected for best-ear listening. In experiment 2, pure-tone BMLDs were measured at third-octave frequencies against the ITD-only, speech-shaped noise interferers of experiment 1. These BMLDs were 4–8 dB at low frequencies for all spatial configurations. In experiment 3, SRTs were measured for speech in diotic, speech-shaped noise. Noises were filtered to reduce the spectrum level at each frequency according to the BMLDs measured in experiment 2. SRTs were as low or lower than those of the corresponding ITD-only conditions from experiment 1. Thus, an explanation of speech understanding in complex listening environments based on the combination of best-ear listening and binaural unmasking (without involving sound-localization) cannot be excluded.
[DOI: 10.1121/1.1772396]

PACS numbers: 43.66.Ba, 43.66.Dc, 43.66.Pn [PFA] Pages: 1057–1065

I. INTRODUCTION

Listeners are often exposed to simultaneous sounds from many sources. The problem of extracting a single target voice from a competing milieu, so that it can be individually understood, has been termed ''the cocktail-party problem'' (Cherry, 1953). Most research on the cocktail-party problem has concentrated on listeners' ability to understand one voice in the presence of one other, or against an undifferentiated babble. In a recent study, Hawley *et al.* (2004) investigated listeners' ability to understand speech in a complex listening environment. They measured speech reception thresholds (SRTs) for up to three interfering sounds of different types and in different binaural configurations. Many, but not all,[1] aspects of their results were consistent with a model of binaural processing in complex listening environments that includes separate mechanisms to exploit interaural time delays (ITDs) and interaural level differences (ILDs). In such a model, the ITDs are exploited via the mechanism of binaural unmasking, while ILDs are exploited purely by best-ear listening. Hawley *et al.* measured monaural thresholds as a means of assessing best-ear listening, and subtracted these from the thresholds for binaural listening to derive a binaural interaction effect. The present study addresses two aspects of Hawley *et al.*'s results for which the individual roles of ILDs and ITDs were not completely clear.

First, Hawley *et al.* included two spatial configurations, (30°,60°,90°) and (90°,90°,90°), in which three interfering sound sources occupied the same hemifield. In one case, they occupied different locations in that hemifield (30°,60°,90°), while in the other, they occupied the same location (90°,90°,90°). No difference was observed between these configurations. This aspect of the data could be interpreted as contrary to expectation based on equalization-cancellation (E-C) theory (Durlach, 1963, 1972), undermining the notion that binaural unmasking effects are sufficient to explain the data. E-C theory suggests that a binaural processor first attempts to equalize (through various transformations) the sound input at the two ears and then subtracts one ear's input

[a)]Electronic mail: cullingj@cf.ac.uk
[b)]Current address: Department 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 Avenue, Madison, WI 53705.

from the other. This binaural processing only improves performance when the interfering source is more intense than the target; if the target is more intense, it is processed monaurally. When the interfering source has a different ITD from the target, the optimal equalization will compensate for the interferer's ITD (since it is more intense) and the cancellation stage will, in consequence, preferentially cancel the interferer. For three interfering sources in different locations, the three interferers will have different ITDs. Since the equalization stage can only equalize a single ITD,[2] one might expect the E-C mechanism to be rather ineffective compared to the case where the interferers share the same location and ITD. Although the data appear not to fit this expectation, Hawley *et al.*'s experiment used a combination of ILDs and ITDs. In both the (30°,60°,90°) and (90°,90°,90°) configurations, there was an advantageous signal-to-noise ratio at one ear produced by the ILDs and the resulting effects of ''best-ear'' listening may have obscured a difference between these configurations.

Second, Hawley *et al.* measured higher SRTs in the (−30°,60°,90°) configuration, in which interferers were present in both hemifields, than in the (30°,60°,90°) configuration, in which all sound sources were in the same hemifield. They interpreted this result as coming from a loss of best-ear listening in the (−30°,60°,90°) configuration, but the difference might in some way have been related to the differences in ITDs between these configurations.

Experiment 1 in the present series of experiments therefore set out to clarify the contributions of these cues using head-related impulse responses (HRIRs) that were manipulated to exclude one or other binaural cue. Experiments 2 and 3 were conducted in order to further clarify the interpretation, by analyzing the role of ITDs in each frequency band.

II. EXPERIMENT 1

A. Stimuli

The stimuli were similar to those of Hawley *et al.* except that the head-related impulse responses (HRIRs) were manipulated in the frequency domain to remove ITDs or ILDs, and only speech and speech-shaped-noise interferers were used. The target sentences were spoken by two male voices (''DA'' and ''CW'') from the MIT recordings of the Harvard Sentence lists (Rothauser *et al.*, 1969). The Harvard sentence lists are grammatically and semantically correct sentences with otherwise relatively low predictability; an example used in the present study (with keywords in capitals) was ''The SMALL PUP GNAWED a HOLE in the SOCK.'' These sentences were presented against either (1) a compound of three other sentences from the database and spoken by the same voice, but selected for greater length, or (2) a compound of three speech-shaped noises, each with the same mean long-term spectrum as the target voice.

In order to place the sounds in different virtual locations, they were convolved with anechoic HRIRs from the HMS III acoustic manikin, as published in the AUDIS catalog (Blauert *et al.*, 1998). The HRIRs were transformed into the frequency domain and processed in two different ways. First, in order to create HRIRs with no ITDs, the phase spectra of a HRIR pair were each replaced with identical phase spectra that linearly increased in phase with frequency (i.e., so that they had a zero ITD). HRIRs were then recreated by inverse Fourier transform. Second, in order to create HRIRs with no ILDs, the amplitude spectra of a HRIR pair were replaced with identical flat spectra. In conducting the latter alteration the scaling of the impulse responses was changed. These scale factors were therefore calculated and compensated for during the convolution process in order to reproduce the original rms.

Speech and speech-shaped-noise interferers were each convolved with the original HRIR and also with each of the manipulated HRIRs in order to produce stereo stimuli with both ILDs and ITDs, ILDs only, or ITDs only. These stimuli were then mixed to give interferers whose component sounds were distributed in virtual space[3] using the desired cues. The locations in virtual space were, as in Hawley *et al.*'s three-interferer conditions, all in front (0°,0°,0°), distributed in both hemifields (−30°,60°,90°), distributed in one hemifield (30°,60°,90°), or concentrated on one lateral location (90°,90°,90°).

With 3 sets of binaural cues ×4 spatial configurations there were 12 different interferer conditions. For the speech-interferer conditions, four versions of each interferer condition were created, two for each voice (''DA'' and ''CW'') using different sentence sets. There were, therefore, 2 voices ×2 sentence sets ×12 interferer conditions =48 interfering speech stimuli. An additional four interfering speech stimuli, one in each spatial configuration, were created for use in practice runs using both ILDs ands ITDs and voice ''DA.'' For the speech-shaped-noise conditions, there were 2 voice-spectra ×12 interferer conditions =24 interfering noise stimuli. In this case, the interferers with both ILDs and ITDs were also used in the practice.

The target sentences were also convolved with manipulated HRIRs, so that they possessed the same type or combination of binaural cues as the interferers against which they were to be presented (ITD, ILD or ITD+ILD). However, the targets were always convolved with HRIRs for directly in front (0°), so the binaural cues would be minimal. Ten target sentences were required for each SRT measurement. The target stimuli were created from 120 sources sentences in order to cover the 12 conditions, 60 from each voice. Once convolved with the 3 different HRIRs, there were 360. An additional 40 target stimuli were generated using voice ''DA'' for use in the practice stimuli.

B. Subjects

Thirty-six listeners with no reported hearing problems and English as a first language were recruited from among Cardiff University students in return for course credit.

C. Procedure

Each listener attended a single 2-h experimental session. For each listener, 16 SRTs were measured in all. The first 4 SRTs were a practice. The remaining 12 experimental SRTs covered each of the four spatial configurations using each of the three combinations of binaural cues. Twenty-four listen-

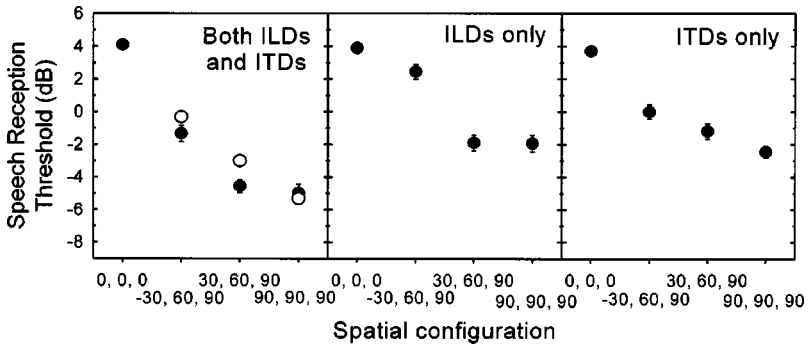

FIG. 1. Results of experiment 1 using speech interferers. Each panel shows speech reception thresholds for the four virtual listening situations for one combination of binaural cues (filled symbols). Error bars are one standard error of the mean. The leftmost panel also shows predicted thresholds based on Bronkhorst's (2000) formula (open symbols).

ers participated in the speech-interferer condition. Twelve listeners participated in the speech-shaped-noise condition.

For the speech interferer conditions, the interfering voice was the same as the target voice. Whether the interferer was the same voice, or a speech-shaped noise derived from that voice, each participant received 6 of the 12 conditions with a given target/interfering voice and 6 with the other. Each participant was paired with another who received the reciprocal allocation of voices to conditions, and experimental materials were rotated such that a given set of target sentences was heard by each listener in a different binaural condition.

Sounds were attenuated and mixed using a Tucker-Davis Technologies AP2 array processor and then presented to listeners via TDT System II hardware (DD1, FT6, PA4, HB6) and Sennheiser HD414 headphones in a single-walled IAC sound-attenuating chamber located in a sound-treated room. The listener made responses via a computer terminal, whose keyboard was placed within the booth and whose screen was visible through the booth window.

SRTs were measured using the method originally described by Culling and Colburn (2000) and based upon that of Plomp (1986). The listeners were instructed that the target sentence would initially be quieter than the interferers and that it would be heard from in front. The same interfering complex of three sounds was presented throughout a given SRT measurement at approximately 53 dB(A). In each run, the listener was informed using the computer terminal's screen of the transcripts of the interfering sentences. Initially, the first target sentence was presented against this interferer, both sentences beginning simultaneously, at a very adverse SNR, and the listener pressed the ''return'' key on the keyboard. The stimulus was repeated, each time at a 4-dB more favorable SNR until the listener judged that half the words of the targets sentence were audible. The listener then entered a transcript. When the listener's transcript was complete, the actual transcript was also displayed on the screen with five keywords in capitals. The listener self-marked his or her transcript and progressed to the next target sentence. The remaining nine sentences were presented at different SNRs according to a 1-up/1-down adaptive threshold algorithm, which increased SNR by 2 dB if fewer then three keywords were correctly transcribed and otherwise decreased SNR by 2 dB. The last eight SNRs derived in this way were averaged to yield a threshold value. The entire transaction was logged and displayed on the experimenter's computer monitor to ensure compliance with the instructions.

D. Results

The results of experiment 1 are shown in Figs. 1 and 2. Figure 1 shows the data for speech interferers, and Fig. 2 the data for speech-shaped-noise interferers. A three-way, mixed analysis of variance was conducted with the two types of interferer (speech and speech-shaped-noise) as a between-subject factor and the three sets of binaural cues (ILD+ITD, ILD-only, and ITD-only) and the four spatial configurations, (0°,0°,0°), (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°) as within-subjects factors. The patterns of data on these two figures are similar across both the available binaural cues and the spatial configurations. SRTs were significantly lower for speech than for speech shaped noise [$F(1,34)=57$, $p<0.0001$]. As in Hawley *et al.*'s results, the effects of different spatial configurations were smaller in magnitude for speech-shaped-noise interferers than for speech interferers. This effect was reflected by a significant interaction between interferer type and spatial configuration [$F(1,102)=13.8$, $p<0.0001$].

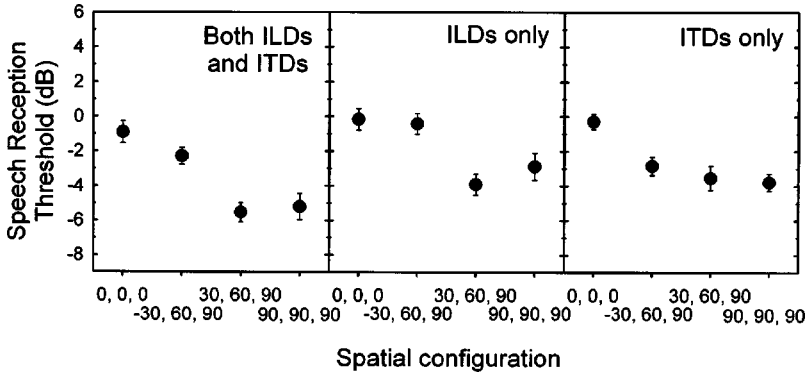

FIG. 2. As in Fig. 1, but for speech-shaped-noise interferers.

There was a significant effect of the available binaural cues [$F(1,68)=22.3$, $p<0.0001$], reflecting the fact that listeners derived more advantage from spatial separations between the target and interfering sources when both binaural cues were available than when either was available in isolation. Tukey pairwise comparisons showed that both binaural cues in combination (ILD+ITD) produced significantly lower SRTs than the ILD-only condition ($q=10.4$, $p<0.001$) and the ITD-only condition ($q=7.13$, $p<0.001$), but that the ILD-only and ITD-only conditions did not differ significantly.

There was a significant effect of spatial configuration [$F(1,102)=156$, $p<0.0001$] and the effect of the available binaural cues depended upon the spatial configuration [$F(1,6)=7.6$, $p<0.0001$]. There were significant differences between the different cue combinations in the (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°) configurations [in each case, $F(1,2)>20$, $p<0.0001$], but not in the (0°,0°,0°) configuration. The interaction was interrogated further using Tukey pairwise comparisons, which confirmed that there were no significant differences between effects of different binaural cues in the (0°,0°,0°) configuration, but showed further that the binaural advantage produced by the other spatial configurations displayed a different pattern across configuration. In the ITD+ILD and ITD-only configurations, all spatial configurations differed significantly from each other except the (30°,60°,90°) and (90°,90°,90°) configurations ($p<0.01$), whereas in the ILD-only condition the (0°,0°,0°) and (−30°,60°,90°) configurations did not differ either ($p>0.05$). In other words, in the ILD+ITD and ITD-only conditions, all forms of spatial separation between target and interferer produced a spatial advantage, but, in the ILD-only condition, this advantage did not occur when the interferers were distributed to both hemifields (−30°, 60°, 90°, configuration). The interaction between interferer type and binaural cues and the three-way interaction were both nonsignificant.

E. Discussion

1. The combination of ILDs and ITDs

The results obtained for the combination of ILDs and ITDs seem to be in agreement with previous studies. Bronkhorst (2000) derived a descriptive expression from several sets of published data (Bronkhorst and Plomp, 1992; Plomp and Mimpen, 1981; Peissig and Kollmeier, 1997), which allows us to predict binaural intelligibility level differences (BILDs) for SRTs measured against multiple speech interferers. In his expression, masking release, R, is predicted for a frontal target source for N interferers with azimuths, θ_i, as follows:

$$R=\left[\alpha\left(1-\frac{1}{N}\sum_{i=0}^{N}\cos\theta_i\right)+\beta\frac{1}{N}\left|\sum_{i=0}^{N}\sin\theta_i\right|\right]. \qquad (1)$$

The parameters α and β are constants, derived by Bronkhorst in a regression analysis. Their values (1.38 and 8.02) have not been altered to accommodate the present data set. Values of R produced by the formula were used to derive predicted differences between the measured (0°,0°,0°) data and the other three configurations. Figure 1 shows predictions based on his formula and parameters with open symbols. The fit appears to be quite good. The formula predicts that there is a substantial masking release in the (−30°,60°,90°) and (30°,60°,90°) configurations, although not quite as large as that observed in the experiment. In the equation, the cosine term evaluates the average angular disparity between the target and each of the interferers, while the sine term makes a symmetry-dependent contribution, which is lower when the arrangement is more symmetrical. It is the latter term, therefore, that introduces a difference between the predictions for (−30°,60°,90°) and (30°,60°,90°), while the former term introduces a difference between (30°,60°,90°) and (90°,90°,90°). It is also worth noting that there is nothing in this formula that would reduce predicted thresholds directly as a result of interferers having different locations and therefore different ITDs.

2. Effect of ITDs alone

Eliminating ILDs from the stimuli produced an effect of spatial configuration that was reduced in magnitude (Figs. 1 and 2, rightmost panels), but similar in form to that of the combination of ILDs and ITDs (leftmost panels). These results show that the binaural system is able to exploit ITDs not only when a single interferer is present (Bronkhorst and Plomp, 1992), but also when multiple interferers have multiple sources. At first sight, the result seems inconsistent with Durlach's (1963, 1972) E-C model, since the cancellation mechanism can only apply a single delay and cancel operation; in the current experiment one would expect this mechanism to eliminate only one of the three spatially distributed interferers and produce a rather small binaural advantage. In order to assess whether this interpretation is justified we evaluated this listening situation using a conceptual approach developed by Levitt and Rabiner (1967).

Levitt and Rabiner showed that it was possible to predict the effect of interaural temporal disparities in the BILD by assuming that the binaural advantage produced in each frequency band by the temporal differences is equivalent to an increase in SNR of the same magnitude. They divided the frequency spectrum into third-octave bands and used an expression for the size of the pure-tone BMLD at each center frequency to give the effective improvement in SNR for that band. They then used AI theory (Fletcher and Galt, 1950; Kryter, 1962) to predict improvement in speech recognition.

It is not straightforward to use Levitt and Rabiner's (1967) method to predict the BILD produced by ITDs for multiple interfering sources of the present experiment directly from the stimulus configuration. Experiments 2 and 3 therefore assessed empirically whether it is reasonable to suppose that the observed pattern of data can be predicted by improvements in this "effective" SNR in each frequency band. Further discussion of the effects of ITDs is deferred until after these experiments are described.

3. Effect of ILDs alone

The results of the ILDs-only condition are quite striking, in that they indicate quite clearly that listeners' use of ILDs

is overwhelmingly dominated by best-ear listening. The conclusion is mainly based upon the fact that SRTs were the same in the (−30°,60°,90°) configuration, where the interfering sources were spatially separated, as in the (0°,0°,0°) configuration, where they were coincident with the target source. This result is in marked contrast to what one would expect if listeners were using ILD as a sound-localization cue and attending to sound coming from directly in front. If that were so, one would expect that the (−30°,60°,90°) configuration would show some advantage over the (0°,0°,0°) configuration. This result is entirely consistent with best-ear listening, because in the (−30°,60°,90°) configuration interfering sources are located on both sides of the head, so that neither ear is shadowed from all the interferers by the head.

In addition, if listeners were attending to a particular location, one would also expect there to be little difference between the (−30°,60°,90°) and (30°,60°,90°) configurations, since in each case there are interfering sounds 30°, 60°, and 90° from the target source. In fact, there is a difference of about 4 dB between these configurations. Again, this result is entirely consistent with best-ear listening, because in the (30°,60°,90°) configuration, all three interfering sources are in the same hemifield leaving one ear in an acoustic shadow, while in the (−30°,60°,90°) configuration, neither ear is shadowed from all the interferers.

4. Effects of interferer type

Several recent studies have observed that spatial unmasking is greater for multiple speech interferers than for noise interferers (Peissig and Kollmeier, 1997; Noble and Perret, 2002; Hawley *et al.*, 2004). This effect was also observed for reversed-speech interferers (Hawley *et al.*, 2004). Comparing across Figs. 1 and 2, the present results replicate this effect. It is not obvious how these results can be interpreted in terms of simple binaural processing strategies. One possibility is that, for speech interferers, there is an additional effect of informational masking which makes the threshold particularly high in the (0°,0°,0°) configuration. A more detailed evaluation of the possible role of informational masking is presented in Hawley *et al.* (2004).

III. EXPERIMENT 2

Levitt and Rabiner (1967) showed that the effects of ITDs in the BILD can be predicted from pure-tone masking release data by (1) assuming that the effect of a given binaural configuration is, effectively, to reduce the spectrum level of the noise in accordance with magnitude of the pure-tone BMLD at each frequency, and (2) predicting intelligibility in the *effective* noise level using the articulation index (Kryter, 1962). In order to apply this model to the current data set, we first measured pure-tone BMLDs for the ITD-only, speech-shaped-noise maskers from experiment 1.

A. Method

Masked detection thresholds were measured for pure tones at 15 frequencies in $\frac{1}{3}$-oct intervals between 200 and 5080 Hz. The maskers were the ITD-only, speech-shaped-noise maskers employed in experiment 1. Four listeners each produced 120 thresholds (4 spatial configurations $\times$15 frequencies $\times$2 interfering voices). Recall that speech-shaped noises were modeled on two different voices. Thresholds were measured in 2I-FC, 2-down/1-up adaptive-threshold procedure with trial-by-trial feedback. The last 12 of 20 reversals contributed to each mean threshold. BMLDs for the (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°) configurations was determined by subtracting the equivalent thresholds in the (0°,0°,0°) configuration.

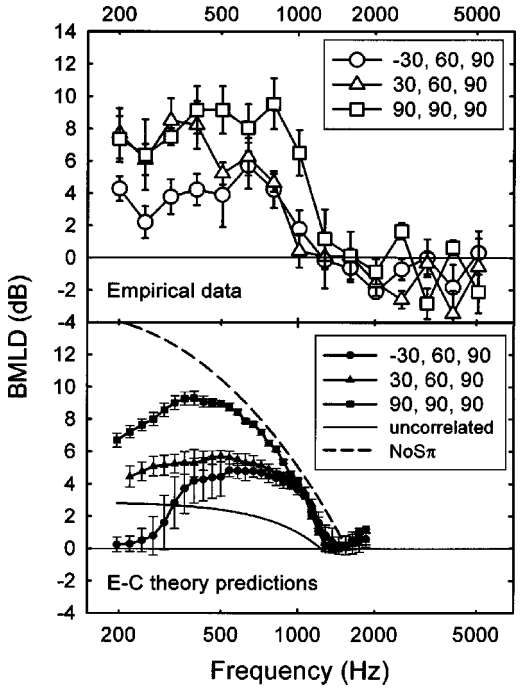

FIG. 3. Upper panel: pure-tone BMLDs for So tones measured at third-octave frequencies between 200 and 5008 Hz against three speech-shaped-noise interferer complexes with ITDs only. Error bars are one standard error of the mean. Lower panel: mean predictions for the same three conditions from E-C theory, implemented using Eq. (2) as well as for predicted thresholds in uncorrelated noise and for NoSπ. Error bars are one standard deviation.

B. Results

Eight of the 480 thresholds were rejected because they differed by more than 10 dB from the mean for frequency and spatial configuration. The remaining results, averaged across the four listeners, are plotted in the upper panel of Fig. 3. It is evident that all three spatial configurations produce a BMLD relative to the (0°,0°,0°) configuration at low frequencies; the differences in ITD between the three maskers in the (−30°,60°,90°) and (30°,60°,90°) configurations do not abolish masking release, although they do reduce it at some frequencies compared to the (90°,90°,90°) configuration. Masking release is smaller in magnitude in the (−30°,60°,90°) configuration than in the other two up to 400 Hz. From 504 Hz upwards, the release is larger in magnitude in the (90°,90°,90°) configuration than in the other two. From 1600 Hz upwards there was no masking release.

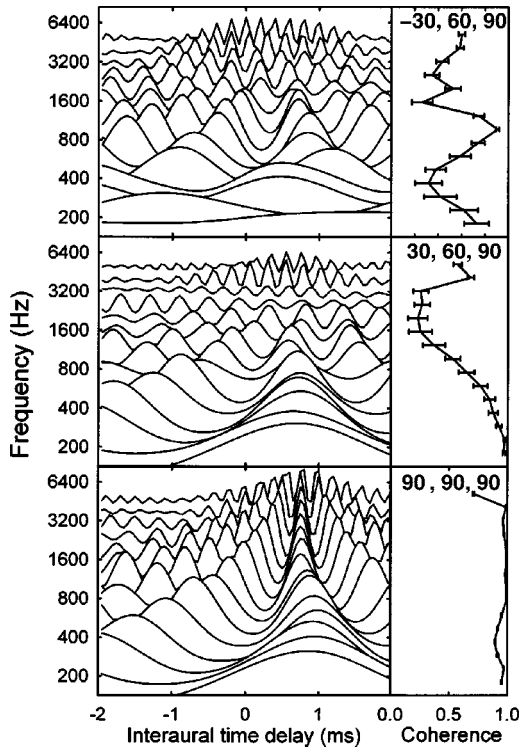

FIG. 4. The left-hand panels show a series of cross-correlations of corresponding left- and right-hand frequency channels from a gamma-tone filterbank (Patterson *et al.*, 1987, 1988) at third-octave frequencies between 200 and 5008 Hz within a 100-ms exponentially tapering temporal window. Separate panels show such cross-correlograms for the (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°) interferers (ITD-only, speech-shaped noise). The right-hand panels show the corresponding coherence (the maximum value of the cross-correlation function) averaged over a series of approximately 100-ms analysis windows. The duration and shape of these windows was measured by Culling and Summerfield (1998). Error bars are one standard deviation of this mean across the series of windows.

C. Discussion

Experiment 2 shows that the BMLD is a surprisingly robust effect in the (−30°,60°,90°) and (30°,60°,90°) configurations, where the interferers come from different directions. It is noteworthy that this was also the case in the (−30°,60°,90°), where the range of interferer ITDs encompasses that of the target. These results can be better understood with reference to the interaural statistics of the combined interference stimuli. Figure 4 shows cross-correlograms and channel-by-channel coherence measurements for the (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°) interference stimuli (ITD-only, speech-shaped noise). The cross-correlograms on the left-hand panels show the modulation of Pearson's *r* with interaural time delay at third-octave frequencies between 200 and 5000 Hz. The right-hand panels show the coherence (the maximum of the cross-correlation function) at the same frequencies. The error bars on the right-hand panels show the standard deviation of the coherence across a series of 100-ms temporal windows. The coherence values give some indication of the potential for binaural unmasking at each frequency.

The E-C model suggests that the binaural system detects a signal through the size of the residue after cancellation, but if the masker is incoherent, the masker will not cancel properly and will also be present in this residue. Detection will, therefore, be best with a coherent masker, but a second condition must also be met for binaural unmasking to be effective. In order to avoid being cancelled with the masker, the interaural phase of the signal must differ from that of the masker at the frequency in question. Durlach and Colburn (1978) pointed out that to a first approximation the pure-tone BMLD is dependent upon the phase difference between signal and masker. In principle, therefore, one can generate an approximate E-C prediction[4] for the BMLD at any frequency, ω, from the coherence, ρ, of the masker at that frequency and the phase difference $(\phi_s - \phi_m)$, between the signal and the masker at its maximum in the cross-correlation function.

Durlach (2003) has provided an expression that allows us to predict from E-C theory the binaural advantage, BMLD (in dB), from ω, c, and $(\phi_s - \phi_m)$:

$$\text{BMLD} = 10\log_{10}\left[\frac{1+\sigma_\epsilon^2-\cos(\omega(\phi_s-\phi_m))\exp(-\omega_s^2\sigma_\delta^2)}{c(1+\sigma_\epsilon^2-\exp(-\omega_s^2\sigma_\delta^2))+(1-c)(1+\sigma_\epsilon^2)}\right]. \quad (2)$$

In this formula, c is the proportion of noise that is common at both ears. It can be related to ρ using Eq. (3). σ_δ and σ_ϵ are taken from Durlach (1972) and have the fixed values of 0.000 105 and 0.25, respectively. Phase and frequency are in radians and radians/second:

$$c = \frac{\sqrt{\rho}}{\sqrt{\rho}+\sqrt{1-\rho}}. \quad (3)$$

The lower panel of Fig. 3 shows predictions that are based on these formulae combined with coherence and phase difference values measured from each type of ITD-only, speech-shaped-noise stimulus.[5] The plotted curves take account of Durlach's (1963) assumption that listeners' thresholds are never below their monaural thresholds, so where the formula returns a negative BMLD, the value has been set to zero. The predictions from E-C theory are broadly consistent with the observed thresholds in the upper panel of Fig. 3, although there is a marked deviation at low frequencies for the (−30°,60°,90°) condition. Also plotted is the theoretical prediction for uncorrelated noise, derived by setting ρ to zero and $(\phi_s - \phi_m)$ to π in Eq. (2) and for NoSπ. Comparing this curve with the others demonstrates that the (−30°,60°,90°) and (30°,60°,90°) maskers are not equivalent to uncorrelated noise from the standpoint of E-C theory. Clearly, E-C theory can predict a robust BMLD for multiple, spatially distributed interferers and more so than one might expect on the basis of being able to cancel just one of them. We set out below an explanation of how the E-C mechanism handles these multiple interfering sources.

In the (90°,90°,90°) configuration, Fig. 4 shows that the coherence is high at all frequencies. The target speech signal is not, however, out of phase with the complex of interferers at all frequencies. The phase of the target is always zero. The phase of the complex of interferers can be seen in the left panel in Fig. 4 from its cross-correlation function. At 504 Hz

the masker cross-correlation has a trough near zero ITD, indicating a phase difference of π radius between masker and signal, so a maximal masking release will occur at this frequency [i.e., the (90°,90°,90°) threshold is close to the NoSπ threshold]. However, at all other frequencies the phase difference is smaller. Whenever target and masker differ in ITD, there will always be frequencies at which the phase difference is less than π (or even zero). It is for this reason that different ITDs are never as effective in producing large BILDs as the NoSπ condition (e.g., Schubert, 1956), for which signal and masker are out of phase at all frequencies.

In the (30°,60°,90°) configuration, the coherence of the masker is lower at high frequencies than in the (90°,90°,90°) configuration, but is similar up to about 500 Hz where the pure-tone thresholds for these two configurations diverge (Fig. 3). E-C theory predicts some difference between (30°,60°,90°) and (90°,90°,90°) below 500 Hz, but not a very large one (~2 dB).

In the (−30°,60°,90°) configuration, there are few frequencies that show high coherence, reflected by the consistently small BMLD in Fig. 3. Predictions and observations show some similar features as a function of frequency. It is noteworthy, however, that both the observed and the predicted BMLDs tend to be above those predicted for uncorrelated noise.

IV. EXPERIMENT 3

Although Levitt and Rabiner (1967) used the articulation index to predict the BILD, we decided to employ an empirical approach by adopting their assumption that the effective spectrum level of the masker is reduced by its interaural configuration with respect to the signal. Culling and Summerfield (1995) postulated that, in the binaural system, each frequency band operates independently, such that the unmasking in each individual frequency band is unaffected by across-frequency differences in interaural configuration (note that peaks in cross-correlation functions in Fig. 4 do not all occur at the same delay). Here we combine these ideas to predict that the effective improvement in SNR at each frequency can be measured from pure tone BMLDs (like those from in experiment 2) without regard to the differences in equalization parameters required in different frequency channels. The size of the pure-tone BMLD predicts the effective reduction in the masker level. Therefore, an equivalent reduction in the actual level of the masker should yield the same thresholds. Experiment 3 tests this prediction.

A. Stimuli

The (0°,0°,0°) speech-shaped noise maskers were filtered in the frequency domain in order to attenuate each frequency by the magnitude of measured pure-tone BMLD at that frequency from experiment 2. BMLD was linearly interpolated between the frequencies measured in experiment 2. Attenuated stimuli of this sort were created to simulate the masking release of the (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°) configurations. SRTs were then measured in seven conditions using diotic target speech as in experiment 1. Four of these conditions were replications of the four ITD-only conditions from experiment 1. In addition, there were simulations of the (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°) configurations based on the filtered copies of the (0°,0°,0°) masker.

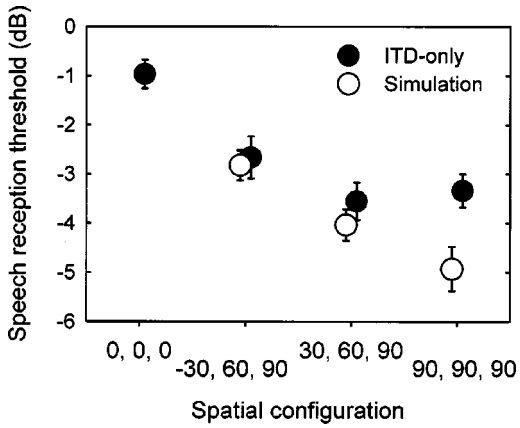

FIG. 5. Replicated SRTs for ITD-only stimuli (filled symbols) and SRTs for stimuli that simulate effects of the (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°) spatial configurations (open symbols). Simulation was achieved by filtering the (0°,0°,0°) interferers, so that each frequency is attenuated in accordance with the measured pure-tone BMLDs (see Fig. 3).

B. Procedure

Fourteen listeners each took part in a single 2-h session. During these sessions they completed a total 16 SRTs. The first two were practice runs, similar to those of experiment 1 and the following 14 were two SRTs in each of the seven conditions. As in experiment 1, the sentence materials were rotated round the different conditions from one participant to the next.

C. Results

The results are plotted in Fig. 5. Thresholds from the simulation condition were similar to, or lower than, those for the ITD-only condition. The results were analyzed using a 2×3 analysis of variance. This analysis covered the ITD-only versus simulation conditions and the three spatial configurations, (−30°,60°,90°), (30°,60°,90°), and (90°,90°,90°), that confer binaural advantage. The (0°,0°,0°) configuration could not be accommodated within this factorial analysis, since it was not replicated for the ITD-only and simulation conditions. The analysis revealed significant main effects of spatial configuration [$F(2,26)=11.9, p<0.0005$] and ITD-only versus simulation [$F(1,13)=7.8, p<0.02$]. There was also a significant interaction between the two [$F(2,26)=3.6, p<0.05$].

The interaction was interrogated using simple main effects: the simulation condition produced significantly lower thresholds than ITD-only condition in the (90°,90°,90°) configuration [$F(1)=17.7, p<0.005$], but did not differ significantly in the (−30°,60°,90°) and (30°,60°,90°) configurations.

D. Discussion

A significant difference was observed between the ITD-only and the simulation conditions only in the (90°,90°,90°) configuration. This result is therefore partially consistent

with Levitt and Rabiner's contention that the BILD results from an effective attenuation of the masker's spectrum in line with the pure-tone masking release at each frequency. If we assume that the pure-tone masking release reflects the action of an E-C mechanism, then such a mechanism can also explain the observed BILDs in the (−30°,60°,90°) and (30°,60°,90°) configurations.

The right-hand panels of Fig. 4 show a further noteworthy effect. The standard deviations of the coherence measurements are much larger in the (−30°,60°,90°) and (30°,60°,90°) configurations than in the (90°,90°,90°) configuration. In order to understand speech in noise, one would expect that the binaural system would need to extract information about the modulation of the residue from cancellation over time in each frequency channel. This modulation in this residue would mirror modulations in coherence of the signal +masker (Culling and Colburn, 2000). Figure 4 shows that in the (−30°,60°,90°) and (30°,60°,90°) configurations, where several maskers occupy different spatial positions, there is considerable modulation in coherence across time in the interferer complex itself. One might expect that this coherence-modulation noise would provide an additional source of high thresholds in the (−30°,60°,90°) and (30°,60°,90°) configurations. However, it appears on current evidence that consideration of this ''noise'' is not necessary to predict the observed performance.

The failure of experiment 3 to produce similar thresholds for the simulated effect of binaural unmasking to the ITD-only condition in the (90°,90°,90°) configuration is an outstanding puzzle. The main purpose of the experiment was to test whether the (−30°,60°,90°) and (30°,60°,90°) configurations could be simulated in this way, since it was difficult without a more detailed examination to see how an E-C mechanism would deal with these maskers. However, it is the (90°,90°,90°) configuration that was not well simulated by filtering the (0°,0°,0°) interferer. Levitt and Rabiner's assumption therefore appears to predict better performance for ITD-only stimuli than was observed. While we are not currently able to explain this result, it does, at least, help to refocus our inquiry into the lack of difference between (30°,60°,90°) and (90°,90°,90°). This lack of difference has been a consistent feature of all the experiments in this series. The current results suggest that, consistent with E-C theory, the (90°,90°,90°) configuration *should be* better than the (30°,60°,90°) configuration; the pure tone thresholds are lower and the simulation based on these thresholds did yield lower SRTs in (90°,90°,90°) than in (30°,60°,90°) (albeit nonsignificantly). It remains to find out why listeners seem to underperform (compared to the prediction) in the (90°,90°,90°) configuration.

V. CONCLUSIONS

Previous research has mostly examined the effect of one masking sound on speech intelligibility in noise. The present study and that of Hawley *et al.* (2004) have extended this research to cover the effects of multiple independent interfering sounds in common or distributed locations. The findings suggest that existing and well-documented mechanisms, best ear listening and binaural unmasking, are largely sufficient to explain performance in these circumstances. Two findings argue against a significant role for sound localization.

First, in experiment 1, SRTs in the ILD-only condition were lower only when the interfering sources were in one hemifield, allowing the contralateral ear an advantageous signal-to-noise ratio. If listeners used binaural cues to attend to the locations of target sources, one would expect improved intelligibility when the interfering sources were separate from the target source regardless of the effect at an individual ear. This result was observed for both speech and speech-shaped noise interferers.

Second, at least for the case of a speech-shaped noise, a combination of binaural unmasking and best-ear listening appear sufficient to explain listeners' performance with multiple, spatially separated interferers. Although one might expect both the BMLD and the BILD to be very poor for three spatially distributed interferers, we found that this intuition is neither predicted by conventional theories of binaural unmasking, nor observed experimentally. Experiment 1 found that, using ITDs alone, listeners were able to produce a spatial unmasking effect for spatially distributed interferers of both the speech and speech-shaped-noise types. Experiment 2 showed that the BMLD is quite robust to spatial distribution of speech-shaped noise interferers and E-C theory predicted the BMLD for these interferer complexes with reasonable accuracy. Experiment 3 showed that the BILD in the ITD-only condition was equal to or less than the effect of reducing the spectrum noise level at each frequency in accord with the size of the pure-tone BMLD at that frequency. Given that the pure-tone BMLDs were broadly predictable from E-C theory, such a mechanism (operating independently in each frequency channel) appears sufficient to explain the intelligibility data for these configurations.

Thus, simple binaural processing strategies, such as channel-independent equalization-cancellation, are quite robust in complex listening situations and can explain the data for speech-shaped-noise interferers quite adequately. However, larger effects of spatial unmasking are observed when multiple speech or reversed speech interferers are used (Hawley *et al.*, 2004). The pattern of thresholds is very similar, but the effects are larger. It is not obvious how these data can be explained by simple binaural processing strategies, but it is apparent from the present experiment that ILDs and ITDs make independent contributions to the spatial unmasking for speech, just as they do for speech-shaped-noise. The same arguments against a role for sound localization can therefore be applied.

ACKNOWLEDGMENTS

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

[1]The exceptional effect was an interaction between voicing of the interfering sounds and spatial distribution. If the interferers were speech or reversed speech, then the advantage of spatial separation attributable to binaural interaction was about twice as great as when the interferers were speech-shaped noise or speech-modulated, speech-shaped noise.

[2]The E-C model can also compensate for interaural differences in level. In this article, the time- and level-equalization processes will, until Sec. III C,

be assumed to operate efficiently, leaving the multiplicity of interfering sources at different ITDs as the main factor limiting performance. ILDs will mainly be considered for their effect on monaural performance at the ear with the most favorable signal-to-noise ratio.

[3]Although the stimuli contained realistic ILDs and ITDs for external virtual locations, the stimuli tended to be perceived as within the head, even when the full set of binaural cues was included.

[4]This application of E-C theory is only approximate because it assumes that the stimulus is composed of one noise that is identical at the two ears except for some interaural time delay and two that are independent and applied to different ears. The stimulus is not constructed that way, but by adding together three noises with different interaural time delays. The approximation relies on the assumption that within a given frequency channel, there is no effective difference between these two constructions, provided that the resulting coherence and time delay are identical. The advantage of making the approximation is that the same formula can be applied to practically any stimulus configuration.

[5]Interaural phase and coherence of the masker complex at each frequency were measured from the output of a gammatone filterbank using a cross-correlation with a 100-ms window. The phase of the target was always zero. Twenty samples were taken at 100-ms intervals and BMLDs calculated separately for each sample.

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.

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

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.

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 Colburn, H. S. (**2000**). ''Binaural sluggishness in the perception of tone sequences and speech in noise,'' J. Acoust. Soc. Am. **107**, 517–527.

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

Culling, J. F., and Summerfield, Q. (**1998**). ''Measurements of the binaural temporal window using a detection task,'' J. Acoust. Soc. Am. **103**, 3540–3553.

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

Durlach, N. I. (**1972**). ''Binaural signal detection: Equalization and cancellation theory,'' in *Foundations of Modern Auditory Theory Vol. II*, edited by J. V. Tobias (Academic, New York).

Durlach, N. I. (**2003**). Personal communication.

Durlach, N. I., and Colburn, H. S. (**1978**). ''Binaural Phenomena,'' in *The Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic, New York).

Fletcher, H., and Galt, R. H. (**1950**). ''The perception of speech and its relation to telephony,'' J. Acoust. Soc. Am. **22**, 89–151.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). ''The benefit of binaural hearing in a cocktail party: Effect of location and type of masker,'' J. Acoust. Soc. Am. **115**, 833–843.

Kryter, K. D. (**1962**). ''Methods for calculation and use of the articulation index,'' J. Acoust. Soc. Am. **34**, 1689–1697.

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.

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

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1987**). ''An efficient auditory filterbank based on the gammatone function'' paper presented to the IOC speech group on auditory modelling at the Royal Signal Research Establishment, 14–15 December.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). ''Spiral vos final report, Part A: The auditory filter bank,'' Cambridge Electronic Design, Contract Report (APU 2341).

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. (**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.

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.

Schubert, E. D. (**1956**). ''Some preliminary experiments on binaural time delay and intelligibility,'' J. Acoust. Soc. Am. **28**, 895–901.