Spatial unmasking of birdsong in human listeners: Energetic and informational factors [a)]

Virginia Best, Erol Ozmeral, Frederick J. Gallun,
Kamal Sen, and Barbara G. Shinn-Cunningham

Hearing Research Center
Boston University
677 Beacon St.
Boston, Massachusetts 02215

Submitted to the Journal of the Acoustical Society of America: 21 April 2005
Revised and resubmitted: 29 July, 2005

Running Head: Spatial unmasking of birdsong

PACS: 43.66.Dc, 43.66.Pn

Corresponding author's address:
Barbara Shinn-Cunningham
Department of Cognitive and Neural Systems
Boston University
677 Beacon St., Room 311
Boston, MA 02215

Tel: 617-353-5764
Fax: 617-353-7755
email: shinn@cns.bu.edu

[a)]Portions of this work were presented at the 2005 Mid-Winter meeting of the Association for Research in Otolaryngology.

ABSTRACT

Spatial unmasking describes the improvement in the detection or identification of a target sound afforded by separating it spatially from simultaneous masking sounds. This effect has been studied extensively for speech intelligibility in the presence of interfering sounds. In the current study, listeners identified zebra finch song, which shares many acoustic properties with speech but lacks semantic and linguistic content. Three maskers with the same long-term spectral content but different short-term statistics were used: (1) chorus (combinations of unfamiliar zebra finch songs); (2) song-shaped noise (broadband noise with the average spectrum of chorus); and (3) chorus-modulated noise (song-shaped noise multiplied by the broadband envelope from a chorus masker). The amount of masking and spatial unmasking depended on the masker and there was evidence of release from both energetic and informational masking. Spatial unmasking was greatest for the statistically-similar chorus masker. For the two noise maskers, there was less spatial unmasking and it was wholly accounted for by the relative target and masker levels at the acoustically better ear. The results share many features with analogous results using speech targets, suggesting that spatial separation aids in the segregation of complex natural sounds through mechanisms that are not specific to speech.

I. INTRODUCTION

In natural environments, sound sources of interest often must be extracted from a background of noise and other distracting sounds. There is a rich history of studies addressing this problem in the context of speech intelligibility, where a listener must extract the content of one source (a 'target') in the presence of competing sources ('maskers'; see Bronkhorst, 2000 for a recent review). Masking is thought to have two main forms. The first is 'energetic masking,' in which the masker reduces the audibility of components of the target due to interference in peripheral frequency channels. The classic illustration of energetic masking is the disruption of speech intelligibility caused by the presence of broadband noise. However, a different kind of masking can occur in addition to energetic masking, or even in the absence of frequency overlap between target and masker. If a competing signal has similar spectro-temporal characteristics it can interfere with the perception of a target at a more central perceptual level (so-called 'informational masking'; Pollack, 1975; Watson, 1987; Durlach *et al.*, 2003). For example, this kind of masking is thought to be a factor in the masking of speech by other talkers with similar voices (Carhart *et al.*, 1969; Brungart *et al.*, 2001).

In most masking situations, spatial separation of the target from the masker(s) improves performance. For primarily energetic maskers, this 'spatial unmasking' has two components. First, the relative energy of the target and masker reaching the ears changes with target and masker location. Usually, spatial separation of target and masker increases the audibility of the target in each frequency band at one of the ears. Second, binaural processing increases the audibility of a target in a particular band if the target and masker contain different interaural time and/or level differences (Zurek, 1993; Bronkhorst, 2000). For primarily informational maskers, the benefit of spatial separation can be much greater than for energetic maskers. In these

conditions it is thought that the differences in perceived location strengthen the formation of distinct objects and reduce confusion between the two sources (Freyman *et al.*, 1999; Arbogast *et al.*, 2002; Kidd *et al.*, 2005).

Many studies have attempted to unravel the contribution of these various factors to speech-on-speech masking and spatial unmasking. Energetic effects are examined by using a noise masker that is matched in its magnitude spectrum to the long-term average spectrum of speech ('speech-shaped noise'). As speech maskers contain large fluctuations in energy which may allow subjects to 'listen in the gaps,' a more appropriate energetic masking control for actual speech is a noise masker modulated by the envelope of a speech signal. Although informational masking is somewhat more difficult to isolate using natural stimuli, it is often assumed to include any additional masking seen with a speech-on-speech masker that cannot be explained by energetic effects in the peripheral auditory representation. Several recent studies adopted a powerful paradigm to minimize energetic masking and emphasize informational masking by processing competing speech signals to have very little masking due to spectral overlap (e.g., see Arbogast *et al.*, 2002). These studies suggest that fundamentally different mechanisms underlie spatial release from energetic and informational masking. For instance, one important feature of spatial release from informational masking is that it appears to be robust to reverberation, unlike spatial release from energetic masking (Kidd *et al.*, 2005).

In natural environments, both energetic and informational masking undoubtedly influence the perception of sound sources (e.g., see Oh and Lutfi, 1999). Interestingly, few studies have examined spatial unmasking with complex natural sounds other than speech. In the current study, zebra finch songs were used to replace human speech in some of the classic masking conditions described above. One reason for using zebra finch song as a stimulus is that it has a

structure that is similar to speech: both are spectro-temporally complex but relatively sparse, both have clear harmonic structure, and both possess dynamic features such as frequency modulation and co-modulation across frequency (Doupe and Kuhl, 1999). By using these songs as stimuli in human experiments we can uncouple the influences of complex spectro-temporal sound structure from top-down linguistic and semantic effects that may affect masked speech perception. In previous studies this goal has been met using reversed speech, which is strongly speech-like but is not intelligible (e.g., see Freyman *et al.*, 2001). However, if spatial unmasking follows similar patterns for zebra finch song, it strongly suggests that the brain has general mechanisms for dealing with complex structured stimuli that are not specific to speech.

II. METHODS

A. Stimuli

1. Target songs

Songs from five male zebra finches (*Taeniopygia guttata*) were used as target stimuli. Between 5 and 30 songs were recorded from each of the five birds. Recordings were conducted in a single-walled sound-treated booth (Industrial Acoustics Company, NY) using a single microphone (Audio-Technica AT3031) placed 7 inches above the caged bird. Often, to entice singing, a female bird was placed temporarily in a neighboring cage in the booth. Four of the birds were recorded with a sampling rate of 32 kHz, and one with a rate of 41.1 kHz.

Recorded songs consisted of many smaller elements (syllables) arranged into repeated patterns (motifs). Five similar motifs were selected from each bird's repertoire. Each motif was

highly stereotypical for a particular bird but quite distinct from those of the other birds. For example, each bird's motif generally consists of a particular pattern of syllables repeated in a fixed order with nearly identical rhythm. Motifs vary across birds in the exact syllables making up the motif as well as the number and rhythm of the syllables. Overall duration of the motifs varied from 750 to 1000 ms across the five birds. For uniformity, all were low-pass filtered at 8 kHz before use in this experiment. The 25 motifs were used both for the identification training and for the masking experiment (see section B). A spectrogram representation of a sample target motif is shown in Figure 1a.

2. Masker stimuli

Three types of masker were used, all with the same long-term spectral characteristics but different short-term statistics. All maskers were generated with duration 1 second to ensure that all target motifs could be fully masked in time. Figure 1 (panels b-d) shows spectrogram representations of examples of each of the three maskers.

Chorus maskers: To make maskers that could easily be mistaken for targets, chorus maskers were generated by adding three song motifs from unfamiliar birds together. Six such mixtures were generated by using all possible combinations of three unfamiliar motifs drawn randomly from a set of five. These unfamiliar motifs were obtained in a previous experiment from five unfamiliar birds. Before adding the unfamiliar motifs to create the chorus, each was looped as necessary to create a 1-second long signal. An example of a chorus masker is shown in Figure 1b.

Song-shaped noise maskers: Song-shaped noise maskers were created by generating broadband noise that had a spectral profile matching that of the average of the set of chorus maskers. Twelve independent maskers were generated, and an example is shown in Figure 1c.

Chorus-modulated noise maskers: Chorus-modulated noise maskers were generated by modulating a song-shaped noise with the envelope from a random chorus masker. Six such maskers were created, using the six chorus envelopes and six different song-shaped noises. These maskers are more similar to the song targets than the song-shaped noise maskers as they contain broad fluctuations in energy over time. An example is shown in Figure 1d.

On each trial, one target and one masker were presented simultaneously at one of seven randomly-selected target-to-masker ratios (TMRs). TMR was calculated using the broadband RMS levels of the two signals. The target level was varied to produce TMRs evenly spaced between -40 dB and 8 dB. These TMRs were chosen on the basis of preliminary testing to span the sloping portions of psychometric functions relating identification performance to TMR. The overall presentation level of the stimulus was set by individual subjects such that the masker level (which remains fixed on every trial) was at a comfortable listening level and the highest TMR was not uncomfortable.

3. Spatialization

For the masking experiment, stimuli were first processed to create binaural signals containing realistic spatial cues, then presented over headphones. The stimuli were processed with pseudo-anechoic head-related transfer functions (HRTFs) measured on a KEMAR manikin at a distance of 1 meter (Shinn-Cunningham *et al.*, 2005) in the horizontal plane at the level of

the ears (0° elevation). In all trials the target was processed by the HRTFs from straight ahead (0° azimuth). The masker was processed with either the same HRTFs at 0° azimuth (‘co-located’) or with HRTFs at 90° azimuth (‘spatially separated’), as depicted in Figure 2.

The resulting spatialized target and masker signals were then added to simulate two sources with the desired spatial cues. To create a realistic, externalized percept, the left and right ear binaural signals for target and masker were summed (‘spatial’ presentation). In control trials, the energetically 'better ear' (the one with the highest target-to-masker ratio) was presented to both ears simultaneously ('diotic' control). This condition exactly reproduced the TMR at the better ear caused by the spatial configuration of target and masker, but removed any differences in perceived location of target and masker.

Casual listening confirmed that the recorded zebra finch songs signals were very ‘dry’ (i.e., not strongly affected by reverberation). Furthermore, because each signal was processed through anechoic HRTFs to generate the spatialized stimuli, any reverberant energy present in the recorded songs could not have caused any interaural decorrelation which might reduce spatial unmasking.

B. Experimental procedures

1. Subjects

Five listeners (1 male, 4 female, aged 22 – 27) were paid for their participation in the experiment, which included both training and testing (see below). Listeners were screened to ensure that they had normal hearing (within 10 dB) for frequencies between 250 Hz and 8 kHz.

2. Environment

Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. Digital stimuli were resampled to 50 kHz and sent to Tucker-Davis Technologies hardware for D/A conversion and attenuation before presentation over headphones (Sennheiser HD-580). Subjects were seated in a sound-treated booth in front of the PC terminal displaying a graphical user interface (GUI). Following each presentation, subjects identified which target bird they heard by clicking on the GUI with a mouse, allowing the PC to store their responses. MATLAB software (Mathworks Inc.) was used to generate the stimuli (offline), to control stimulus presentation, and to collect responses for later analysis.

3. Identification training

Subjects were trained to identify the five target birds on the basis of their unique song motifs. Each target bird was given a name (‘Uno,’ ‘Junior,’ ‘Moe,’ ‘Toro’ and ‘Nibbles’) that subjects were trained to associate with the specific motifs. Training began with a familiarization session in which subjects could press one of five labeled buttons on the GUI and hear the song of the corresponding bird. This session continued for as long as the subject desired. Subjects became familiar with the birds relatively quickly, and reported anecdotally that the different birds were distinguishable on the basis of (a) particular syllables having a unique pitch or structure as well as (b) the temporal arrangement of syllables.

After familiarization, subjects initiated an identification test of 100 trials. In this test, a target was presented in quiet (with no masker) and the subject was required to identify the bird by clicking on the appropriate button. Correct response feedback was provided in written form on the screen. Subjects were required to achieve a score of at least 90% on this test before moving on to the masking experiment, and all subjects met this criterion on their first attempt.

4. Masking Experiment

The format of the masking tests was similar to that of the identification test, but a masking stimulus was present and no feedback was provided. In a single test, the masker type (chorus, modulated-noise, or noise), spatial configuration (co-located or separated), and presentation mode (spatial or diotic) were fixed. Each test consisted of 35 trials (five repetitions at each of the seven TMRs, randomly interleaved). Subjects were instructed to listen for the target stimulus, which was always simulated as coming from directly in front, and to identify it by clicking on the GUI.

All combinations of masker type, spatial configuration, and presentation mode were tested in a single session, for a total of 12 tests per session. The tests were presented in a different random order for each subject. In order to ensure that subjects maintained their ability to identify the target birds in quiet during the experiment, short identification tests were interleaved with the masking tests. At the beginning of a new session, subjects were required to make 24 correct identifications in a 25-trial test in order to commence the masking test. Within a session, subjects were required to make 10 correct identifications on a 10-trial test before every masking test.

Each subject completed six sessions in total, corresponding to 30 trials at every TMR in every condition. No subject completed more than one session on any given day.

5. *Generation of psychometric functions*

Data were sorted by subject, masker type, and presentation condition, and psychometric functions were plotted for each case. Raw psychometric functions were generated by plotting performance (in percent correct) as a function of TMR (see Figure 3). To enable the estimation of slope and threshold parameters, logistic functions were fit to each raw psychometric function.

III. RESULTS

A. Performance as a function of target-to-masker ratio

Figure 3 shows the mean raw psychometric functions across subjects for both co-located and spatially separated configurations (error bars indicate the across-subject standard deviation). The top, center, and bottom panels show data for the song-shaped noise masker, chorus-modulated noise masker, and chorus masker, respectively. For all conditions, performance improved with increasing TMR, from chance levels (20% correct) to near perfect identification. Furthermore, for all maskers, there was a large advantage to having target and masker spatially separated (compare squares to circles). For song-shaped and chorus-modulated noise maskers, spatial and diotic presentations gave similar results (compare filled and open symbols in top and center panels). However, for the chorus masker, spatial performance was superior to diotic performance, but only when the sources were spatially separated (compare filled and open symbols in bottom panel).

An additional observation that can be made about the psychometric functions is that their slopes vary with masker type. To quantify this effect, slopes of the logistic functions fit to the raw data were examined for each subject. Figure 4 shows the mean and standard deviation of these slope values for each masker type (pooled across subjects and psychometric functions). On average, slopes were steepest for the song-shaped noise masker, but were similar for the other two maskers. An ANOVA revealed a significant effect of masker condition on slope [$F(2,57) = 6.3$, $p = 0.0034$], and post-hoc analysis (Tukey HSD, $p = 0.05$) confirmed that the song-shaped noise masker produced steeper slopes than the other two maskers.

B. Individual masked thresholds

In order to compare performance across the various conditions, thresholds were extracted from the individual logistic functions. Threshold was defined as the TMR giving 60% accuracy, which represents a performance level half-way between chance (20%) and perfect performance (100%). Thresholds can be seen for the three maskers in the three panels of Figure 5. In each panel, the five columns represent the five subjects with thresholds plotted in dB (note that a lower value indicates less masking).

This figure demonstrates that there were large individual differences, but also highlights several main effects. First, the chorus masker was a more effective masker than the two noise maskers. In general, thresholds are higher for the chorus masker than the noise maskers, i.e., the target had to be presented at a higher intensity to reach threshold performance. An ANOVA showed a significant effect of masker type [$F(2,57) = 8.74$, $p = 0.0005$] and post-hoc analysis (Tukey HSD, $p = 0.05$) confirmed that thresholds for the chorus masker were significantly larger

than for the two noise maskers. Second, spatial separation resulted in a reduction in masking for all subjects in all conditions (compare squares and circles). A third important result is seen in the difference between the spatial (filled symbols) and the diotic (open symbols) presentation conditions. For the two noise maskers, thresholds are essentially the same for spatial and diotic presentations. For the chorus masker, the spatial and diotic conditions produced similar thresholds in the co-located configuration. However, for the spatially separated configuration the spatial condition consistently produced less masking than the diotic condition. These latter effects are quantified and examined more closely in the following section.

C. Spatial release from masking

Spatial unmasking was calculated by taking the difference in threshold between the co-located and spatially separated configurations for each subject and each condition. Mean spatial unmasking values, averaged across the five subjects, are plotted in Figure 6 (error bars show across-subject standard deviations).

Spatial release from masking in the spatial listening condition was similar for the song-shaped noise and chorus-modulated noise maskers (means of 16.8 dB and 15.3 dB respectively). For these noise maskers, spatial unmasking in the diotic listening condition was also substantial (means of 16.5 dB and 14.8 dB) and not significantly different from the spatial condition (paired t-tests; $p = 0.5$, 0.4, respectively). In contrast, for the chorus masker, there was a large advantage in the spatial listening condition (mean 21.1 dB compared to 11.1 dB for the diotic control), a difference that was highly significant (paired t-test, $p = 0.001$).

IV. DISCUSSION

A. Spatial unmasking of birdsong and speech in human listeners

For the two noise maskers (song-shaped noise and chorus-modulated noise), benefits of around 17 dB and 15 dB (respectively) were observed with spatial separation of the target and masker. These values are quite large compared to spatial benefits reported for the intelligibility of speech masked by speech-shaped noise. In his comprehensive review, Bronkhorst (2000) reported between 6 and 10 dB of spatial release from masking for various types of speech material in this configuration (target in front, masker in front or to the side). One likely explanation for this difference is that zebra finch song contains more high-frequency energy (above 2 kHz) than speech.

Figure 7a compares the power spectral density of zebra finch song and speech (calculated using the MATLAB function ‘psd’). The zebra finch song curve is based on the 25 target tokens used in the current experiment, and the speech curve is based on a sample of similar size from a well-known speech corpus (Bolia *et al.*, 2000). While the speech signals contain significant low-frequency energy and have spectral levels that drop off gradually above 1 kHz, the songs have most of their energy between 2 – 5 kHz (see also Zann, 1996). Given the small wavelengths at these frequencies and the size of the human head, the head-shadow effect for zebra finch song is large and greatly improves the target-to-masker ratio in the better ear when sources are spatially separated. Indeed, analysis of the long-term broadband TMR at the better ear showed an increase of approximately 18 dB with spatial separation, which can fully account for the benefits observed for the noise maskers. The idea that advantageous energy in the better ear is driving much of the observed spatial unmasking is consistent with the observation that the unmasking

was equal in the spatial and diotic conditions for these maskers. In other words, for the noise maskers, the benefit of moving the masker to the side can be entirely explained by the change in energy at the better ear.

The fact that no additional spatial unmasking was observed with spatial presentation relative to diotic presentation for the noise maskers implies that ITD processing did not provide any release from masking for these stimuli. In contrast, for speech, additional advantages of binaural processing of up to 7 dB are typically observed (Carhart *et al.*, 1967; Dirks and Wilson, 1969). Models of binaural unmasking show that ITD effects are dominant for frequencies up to 500 Hz, and essentially disappear for frequencies above 2 kHz (Zurek, 1993, see Figure 7b). Thus, this apparent discrepancy between spatial unmasking of zebra finch song and speech presumably reflects the different amounts of low-frequency energy present in the stimuli. Unlike speech, the zebra finch songs used in this study have very little energy below 2 kHz (Figure 7a). It may also be that the information below 2 kHz is less important for the identification of zebra finch song than for the identification of speech (and perhaps the importance of this energy for identification.

For the chorus masker, the better ear advantage was roughly 11 dB (i.e. diotic thresholds for the spatially separated configuration were 11 dB lower than for the co-located configuration). This benefit is smaller than that found for the noise maskers, presumably because the chorus masker is spectro-temporally sparser and therefore a less effective energetic masker than the noise maskers. However, in contrast to the noise maskers, the spatial release from masking in the spatial condition was greater than in the diotic condition for all subjects. It can be assumed that this extra spatial unmasking (approximately 10 dB) is not due to binaural processing, as it did not occur for the noise maskers (which have more energetic overlap and hence are more likely to gain an advantage from such within-channel processing). We attribute the large extra spatial

unmasking seen with the chorus masker to a reduction in informational masking due to perceived differences in target song and chorus locations.

The mean spatial release from masking of 21.1 dB in the chorus condition of this study is large when compared to the spatial release reported in past speech studies. For studies of speech intelligibility against a background of same-talker speech, reported spatial unmasking values range up to 14 dB (e.g see Freyman *et al.*, 1999). As discussed already, head-shadow contributes much more to spatial unmasking for zebra finch song than it does for speech. This large contribution of better-ear TMR benefit at least partially accounts for the large amounts of spatial unmasking observed in the current study. The 10 dB of extra unmasking that we attribute to informational unmasking (although remarkable) is within the range observed in speech tasks dominated by informational masking. In situations where informational masking dominates, release from masking can reach up to 18 dB (Arbogast *et al.*, 2002; Kidd *et al.*, 2005).

B. Evidence for different forms of masking

To summarize, the different maskers in the current experiment resulted in different patterns of masking and of spatial unmasking. A clear indication that the chorus masker produced the most informational masking is the fact that subjects made substantial identification errors even when the target was clearly audible in a chorus background (Figure 3, TMRs of 0 and 8 dB). However, perhaps the most important finding was that the benefit of spatial separation for the chorus masker was much *greater* than for the noise maskers, even though the energetic gain due to separation was *smaller.* This is consistent with the idea that spatial separation can act to reduce informational masking in situations where the target and masker are have similar short-term spectro-temporal characteristics and are easily confused with one another (Durlach *et al.*,

2003b). For this experiment, the spatial percept helped listeners group target segments together (and segregate them from masker segments), improving identification performance.

Secondary support for these different kinds of masking comes from the psychometric functions described in section IIIA. It has been noted previously that more 'informational' maskers tend to give rise to shallower psychometric functions than more 'energetic' maskers (Festen and Plomp, 1990; Kidd *et al.*, 1998; Lutfi *et al.*, 2003). One reason that has been put forward for this is that informational maskers are generally less homogeneous than energetic maskers. If the different maskers in the inhomogeneous set are differentially effective (and give rise to psychometric functions with different thresholds), then averaging across these maskers will give rise to a shallower slope even if each masker-specific psychometric function is equally steep (see Durlach *et al.*, 2005). A second explanation for shallower slopes in informational masking is that the masking is due to confusion and thus depends less directly on target-to-masker ratio than energetic masking.

In the current study (Figure 3, 4) the psychometric functions were steepest for the song-shaped noise masker, the masker that was most homogeneous from trial-to-trial and which caused little informational masking. The chorus masker gave rise to shallower psychometric functions, consistent with both factors: the individual maskers were drawn from a highly inhomogeneous set, and the chorus masker may interfere with target song identification in a way that is only weakly dependent on TMR. The fact that the chorus-modulated noise masker (which had a similar amount of inhomogeneity) produced equally shallow psychometric functions suggests that masker inhomogeneity was a dominant factor affecting slope values in the current study. Interestingly, however, an analysis of the average performance level for the different masker tokens in the current study revealed no greater variability across chorus maskers and

chorus-modulated noise maskers than across song-shaped noise maskers. Furthermore, there appeared to be no specific interaction between particular targets and particular chorus or chorus-modulated noise masker tokens. It remains to be seen whether this is a result of the heavy data reduction required for this analysis; perhaps a larger data set might reveal such effects.

It was interesting to find (as discussed in the previous section) that, in general, the contributions of energetic and informational masking and the relative benefits of spatial separation follow very similar patterns for zebra finch song and speech. Although there are differences in the extent to which spatial separation improves identification of these two natural stimuli, most of these differences can be attributed to differences in the acoustics of the stimuli (e.g. differences in which frequency range contains information about the speech or song content). In particular, it seems that identifying a bird in a chorus poses a problem similar to understanding a talker in the presence of other talkers. Such fundamental similarities suggest that there are general mechanisms for segregating complex sounds that are not unique to speech.

Traditional models of spatial unmasking (such as those estimating speech intelligibility in the presence of interference) cannot predict the effects of informational unmasking or the benefits of spatially separating target and masker when informational masking is dominant. Traditional models consider energy effects at the ears as well as binaural processing (see Colburn, 1996 for a review), operating on each frequency channel independently. Thus, such models only explain within-channel masking effects. In the current study, as well as in the speech studies discussed earlier, benefits of spatial separation have been observed that (a) do not depend on frequency overlap and (b) are much larger than traditional ‘energetic’ unmasking effects. Extensions of existing models are required to explain these effects and produce a complete picture of spatial unmasking with complex stimuli. Some modeling efforts have had success in predicting the

effect of masker uncertainty on tone detection (e.g. Lutfi, 1993; Oh and Lutfi, 1999). However, some aspects of informational masking, such as confusion between target and masker components, have not yet been modeled. Furthermore, no models have been applied to explain performance in more complex tasks such as understanding speech in a mixture of similar maskers.

C. Final comments

This study demonstrated that for the human listener, spatial separation enhances the identification of a familiar zebra finch song in the presence of different kinds of interference. The results give insight into what kinds of masking can occur with these signals, and what factors can provide release from masking in humans. However, it is not known how relevant this situation is to the birds that use these signals for communication in real environments. While there is evidence that songbirds are capable of segregating mixtures of signals (see Hulse, 2002 for a review), spatial factors have not yet been examined. Furthermore, some evidence suggests that spatial cues are not as salient for small birds as they are for humans (Park and Dooling, 1991; Dent and Dooling, 2004; however see Nelson and Suthers, 2004). It would be interesting to test zebra finches on the same stimuli used in this experiment, to address whether spatial unmasking enhances song identification in these birds.

V. CONCLUSIONS

Spatial separation enhanced the ability of human listeners to identify familiar zebra finch songs in the presence of interfering sounds with identical long-term spectra, but the nature of the benefit varied with the short-term statistics of the interference. All maskers showed a large

benefit due to energetic advantages at the acoustically 'better ear.' For noise maskers this advantage could fully explain the observed spatial unmasking. However, for maskers made up of unfamiliar songs, there was an additional advantage of spatial separation that could not be solely explained by energetic effects. It appears that when the target and masker had similar short-term spectro-temporal characteristics, differences in perceived location helped listeners segregate the sources, leading to large reductions in informational masking. The data are consistent with previous studies examining speech recognition in the presence of noise and competing speech sources and provide further evidence that both energetic and informational masking influence behavior in natural acoustic settings.

ACKNOWLEDGMENTS

This work was supported by a grant from the Air Force Office of Scientific Research to BGSC (Grant FA9550-04-1-0260).

REFERENCES

Arbogast, T. L., Mason, C. R. and Kidd, G. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**(5), 2086-2098.

Bolia, R. S., Nelson, W. T., Ericson, M. A. and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**(2), 1065-1066.

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

Brungart, D. S., Simpson, B. D., Ericson, M. A. and Scott, K. R. (**2001**). “Informational and energetic masking effects on the perception of multiple simultaneous talkers,” J. Acoust. Soc. Am. **110**(5), 2527-2538.

Carhart, R., Tillman, T. W. and Greetis, E. S. (**1969**). “Perceptual masking in multiple sound backgrounds.,” J. Acoust. Soc. Am. **45**(3), 694-703.

Carhart, R., Tillman, T. W. and Johnson, K. R. (**1967**). “Release of masking for speech through interaural time delay,” J. Acoust. Soc. Am. **42**, 124-138.

Colburn, H. S. (**1996**). "Computational models of binaural processing," in: *Auditory Computation*, ed. H. L. Hawkins, T. A. McMullen, A. N. Popper and R. R. Fay (Springer-Verlag, New York).

Dent, M. L. and Dooling, R. J. (**2004**). “The precedence effect in three species of birds (*Melopsittacus undulatus, Serinus canaria, and Taeniopygia guttata*),” J. Comp. Psychol. **118**(3), 325-331.

Dirks, D. D. and Wilson, R. H. (**1969**). “The effect of spatially separated sound sources on speech intelligibility,” J. Speech Hear. Res. **12**, 5-38.

Doupe, A. and Kuhl, P. K. (**1999**). “Birdsong and speech: Common themes and mechanisms.,” Ann. Rev. Neurosci. **22**, 567-631.

Durlach, N. I., Mason, C. R., Gallun, F. J., Shinn-Cunningham, B. G., Colburn, H. S. and G. Kidd, J. (**2005**). “Psychometric functions for fixed and randomly mixed maskers,” J. Acoust. Soc. Am. **in preparation**.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S. and Shinn-Cunningham, B. G. (**2003**). “Note on informational masking,” J. Acoust. Soc. Am. **113**(6), 2984-2987.

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

Freyman, R. L., Helfer, K. S., McCall, D. D. and Clifton, R. K. (**1999**). “The role of perceived spatial separation in the unmasking of speech,” J. Acoust. Soc. Am. **106**(6), 3578-3588.

Hulse, S. H. (**2002**). "Auditory scene analysis in animal communication," in: *Advances in the Study of Behavior*, *31*,ed. P. Slater, J. Rosenblatt, C. Snowdon and T. Roper (Academic Press.

Kidd, G., Mason, C. R., Brughera, A. and Hartmann, W. M. (**2005**). "The role of reverberation in release from masking due to spatial separation of sources for speech identification," Acustica united with Acta Acustica **in press**.

Kidd, 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**(1), 422-431.

Lutfi, R. A. (**1993**). "A model of auditory pattern analysis based on component-relative entropy," J. Acoust. Soc. Am. **94**, 748-758.

Lutfi, R. A., Kistler, D. J., Callahan, M. R. and Wightman, F. L. (**2003**). "Psychometric functions for informational masking," J. Acoust. Soc. Am. **114**(6), 3273-3282.

Oh, E. L. and Lutfi, R. A. (**1999**). "Informational masking by everyday sounds," J. Acoust. Soc. Am. **106**(6), 3521-3528.

Park, T. J. and Dooling, R. J. (**1991**). "Sound localization in small birds: Absolute localization in azimuth," J. Comp. Psychol. **105**, 125-133.

Pollack, I. (**1975**). "Auditory informational masking," J. Acoust. Soc. Am. **57**, S5.

Shinn-Cunningham, B. G., Kopco, N. and Martin, T. (**2005**). “Localizing nearby sound sources in a classroom: binaural room impulse responses,” J. Acoust. Soc. Am. **in press**.

Watson, C. S. (**1987**). "Uncertainty, informational masking and the capacity of immediate auditory memory," in: *Auditory Processing of Complex Sounds,* ed. W. A. Yost and C. S. Watson (Lawrence Erlbaum, Hillsdale, NJ), 267-277.

Zann, R. A. (**1996**). *The Zebra Finch: A Synthesis of Field and Laboratory Studies* (Oxford University Press, New York).

Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in: *Acoustical Factors Affecting Hearing Aid Performance,* ed. G. A. Studebaker and I. Hochberg (Allyn and Bacon, Boston), 255-276.

FIGURE CAPTIONS

Figure 1. Spectrograms of (a) an example target song, (b) a chorus masker, (c) a song-shaped noise masker, and (d) a chorus-modulated noise masker.

Figure 2. The two spatial configurations examined. The target (T) was always located directly in front of the listener. The masker (M) was either co-located with the target (left panel) or spatially separated at 90° to the right (right panel).

Figure 3. Mean psychometric functions showing percentage correct for different TMRs. Data points represent the mean across subjects and error bars show standard deviations. Each panel shows results for one masking condition, and the four curves in a panel represent the different presentation conditions. Symbol type indicates spatial configuration (circles: co-located, squares: spatially separated). Symbol shading indicates listening condition (filled: spatial, open: diotic).

Figure 4. Mean slopes of psychometric functions for each masker condition. Slope values were extracted from logistic fits to the raw data. Bars represent the mean across subjects and psychometric functions for a particular masker, and the error bars represent standard deviations across the pooled values.

Figure 5. Thresholds measured from fits to individual psychometric functions. The three panels show thresholds for the three masking conditions. For all panels, the five subjects are represented along the abscissa. Symbol type indicates spatial configuration (circles: co-located, squares: spatially separated). Symbol shading indicates listening condition (filled: spatial, open: diotic). Note that a lower threshold indicates better performance.

Figure 6. Mean spatial unmasking (threshold for separated configuration minus threshold for co-located configuration). The three masking conditions are represented along the abscissa, and the bars represent the mean across subjects (filled: spatial, open: diotic). Note that a higher value represents a larger benefit of spatial separation. Error bars show standard deviations.

Figure 7. (a) Mean power spectral density plots of zebra finch song and speech (see text for details of the samples used). (b) Maximum binaural advantage predicted by the model of Zurek (1993) as a function of frequency. This maximum corresponds to the detection of an interaurally out-of-phase signal in diotic noise (figure adapted from Zurek, 1993).

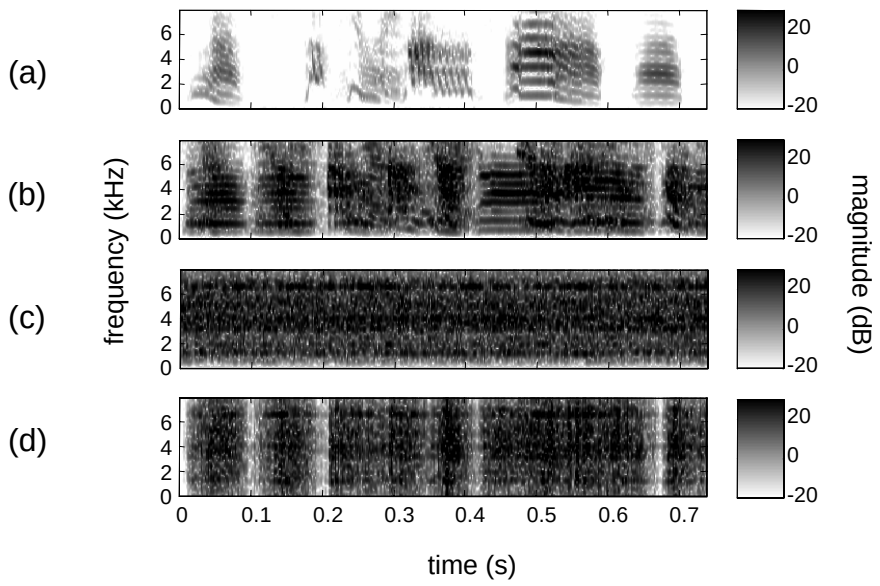
(a)
(b)
(c)
(d)
frequency (kHz)
6
4
2
0
20
0
-20
magnitude (dB)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
time (s)

Figure 1.

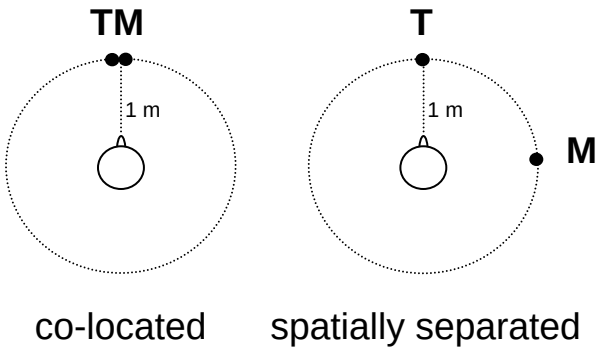

Figure 2.

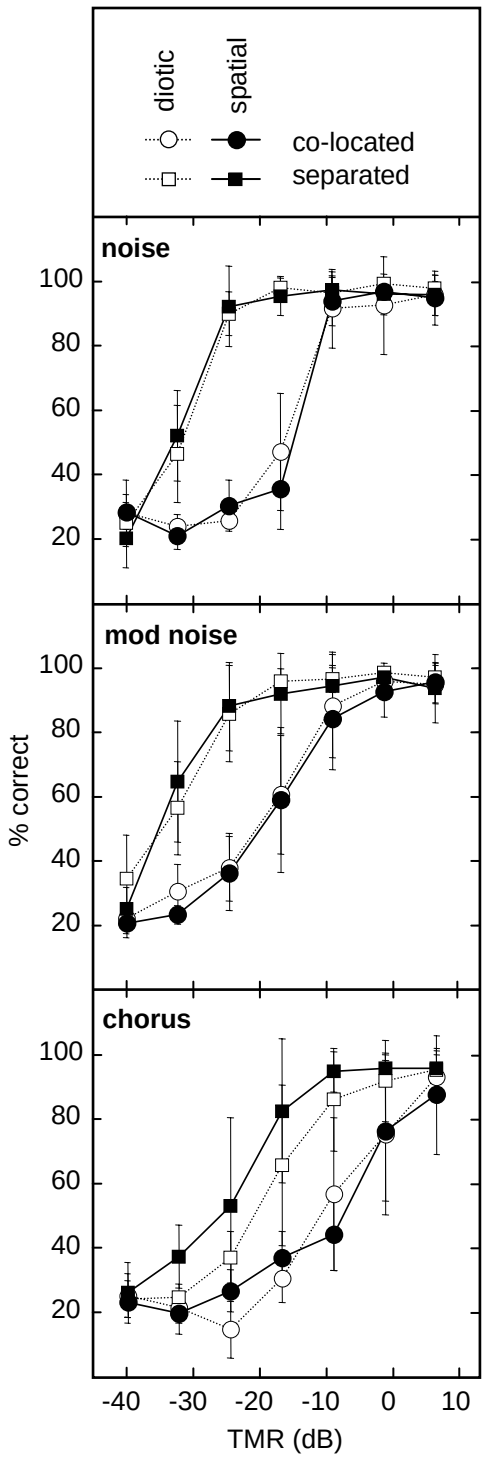
diotic
spatial
co-located
separated
noise
mod noise
chorus
% correct
100
80
60
40
20
-40
-30
-20
-10
0
10
TMR (dB)

Figure 3.

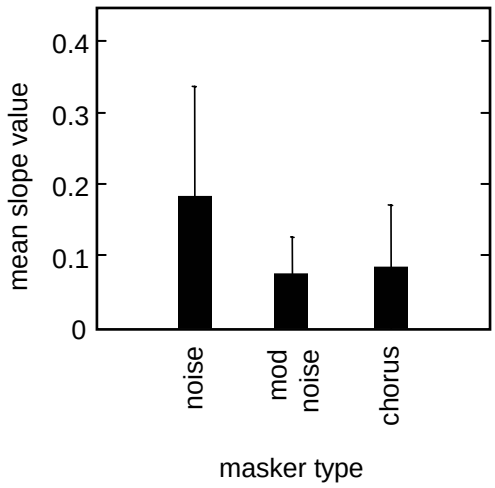
mean slope value
0.4
0.3
0.2
0.1
0
noise
mod noise
chorus
masker type

Figure 4.

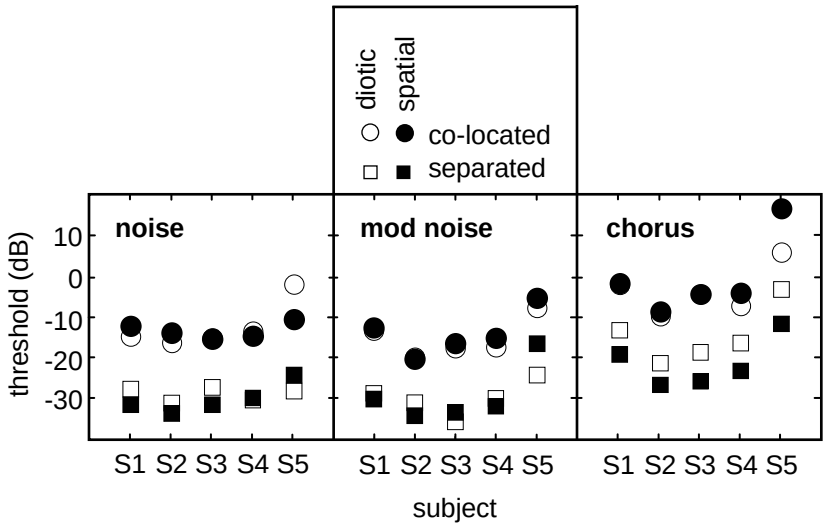
diotic
spatial
co-located
separated
noise
mod noise
chorus
threshold (dB)
10
0
-10
-20
-30
S1 S2 S3 S4 S5
S1 S2 S3 S4 S5
S1 S2 S3 S4 S5
subject

Figure 5.

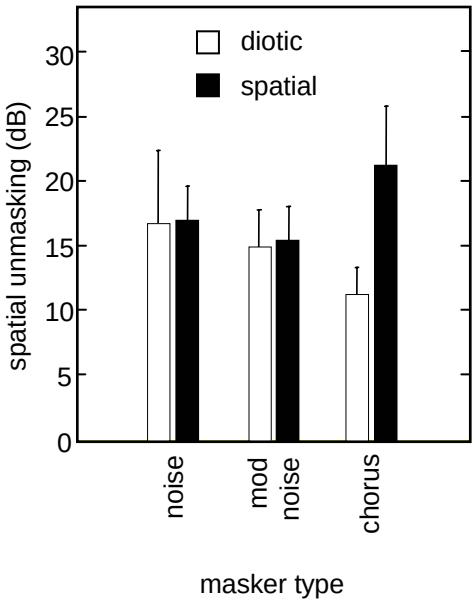
diotic
spatial
spatial unmasking (dB)
0
5
10
15
20
25
30
noise
mod noise
chorus
masker type

Figure 6.

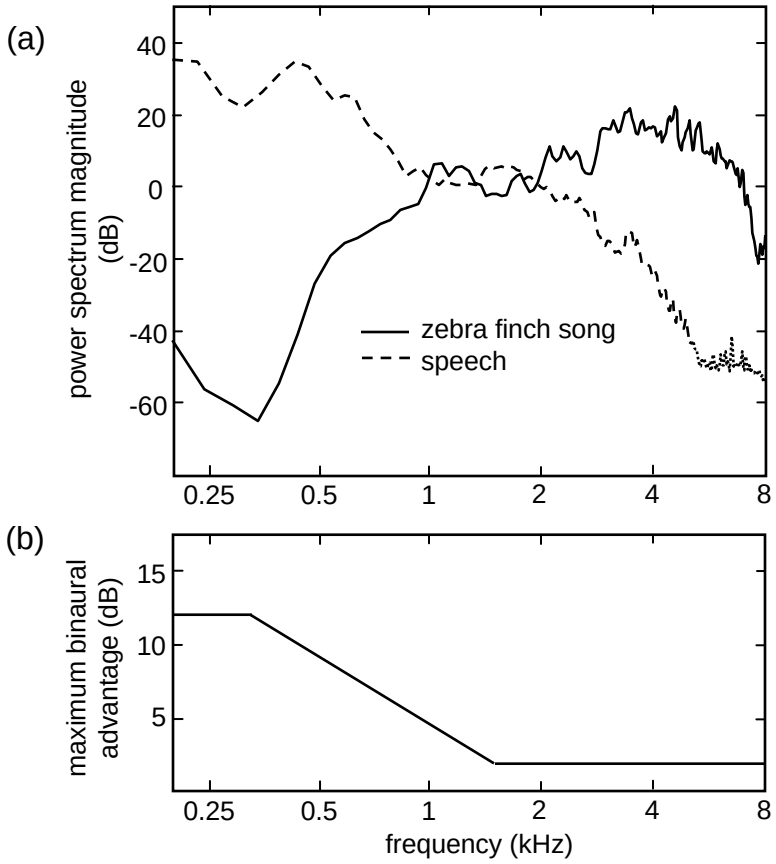
(a)
power spectrum magnitude (dB)
40
20
0
-20
-40
-60
zebra finch song
speech
0.25
0.5
1
2
4
8
(b)
maximum binaural advantage (dB)
15
10
5
0.25
0.5
1
2
4
8
frequency (kHz)

Figure 7.