Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences

Andrew Schwartz
Harvard/MIT Speech and Hearing Bioscience and Technology Program, Cambridge, Massachusetts 02139

Josh H. McDermott
Center for Neural Science, New York University, New York, New York 10003

Barbara Shinn-Cunningham[a)]
Center for Computational Neuroscience and Neural Technology and Biomedical Engineering, Boston University, Boston, Massachusetts 02215

(Received 20 April 2012; accepted 27 April 2012)

To clarify the role of spatial cues in sound segregation, this study explored whether interaural time differences (ITDs) are sufficient to allow listeners to identify a novel sound source from a mixture of sources. Listeners heard mixtures of two synthetic sounds, a target and distractor, each of which possessed naturalistic spectrotemporal correlations but otherwise lacked strong grouping cues, and which contained either the same or different ITDs. When the task was to judge whether a probe sound matched a source in the preceding mixture, performance improved greatly when the same target was presented repeatedly across distinct distractors, consistent with previous results. In contrast, performance improved only slightly with ITD separation of target and distractor, even when spectrotemporal overlap between target and distractor was reduced. However, when subjects localized, rather than identified, the sources in the mixture, sources with different ITDs were reported as two sources at distinct and accurately identified locations. ITDs alone thus enable listeners to perceptually segregate mixtures of sources, but the perceived content of these sources is inaccurate when other segregation cues, such as harmonicity and common onsets and offsets, do not also promote proper source separation.
 [http://dx.doi.org/10.1121/1.4718637]

PACS number(s): 43.66.Pn, 43.66.Mk, 43.66.Dc [EB] Pages: 357–368

I. INTRODUCTION

Natural environments often contain multiple concurrent acoustic sources whose waveforms superimpose to create ambiguous mixtures of sound. Because the same acoustic mixture can be produced by a vast number of combinations of sources, segregating a source from a mixture presents a computational challenge (Cherry, 1953; Bregman, 1990; Darwin, 1997; Bronkhorst, 2000; Carlyon, 2004; McDermott, 2009; Shamma and Micheyl, 2010). Listeners solve this under-constrained problem by estimating which mixture elements are likely to have come from the same source, partly relying on grouping cues derived from the spectrotemporal structure of natural sound sources (Bregman, 1990). The perceived sources are perceptual estimates of the true content of physical sound sources in the external world (e.g., see Shinn-Cunningham, 2008).

Intuitively, spatial cues such as interaural time and level differences (ITDs and ILDs, respectively) seem as if they should be useful for segregating sound sources: sound elements that originate from the same location are likely to come from a common source. In realistic listening situations, such as when attending to a particular speaker in a "cocktail party" environment, spatial cues indeed appear to aid segregation, improving the comprehension of target sentences in a background of other speakers (Bronkhorst, 2000; Drullman and Bronkhorst, 2000; Freyman *et al.*, 2001; Hawley *et al.*, 2004; Ihlefeld and Shinn-Cunningham, 2007). There are a few ways in which spatial cues could aid a listener in this situation. When sound signals overlap in time and frequency so that the target is "energetically masked" by competing signals, interaural differences can be exploited to enhance the signal-to-noise ratio, lowering the relative level of the target required for it to be detected (Durlach, 1963; Akeroyd, 2004; Culling, 2007; Wan *et al.*, 2010). However, spatial cues can also help listeners identify a target sound whose elements are clearly audible, as when ambiguities of grouping and/or selection interfere with performance, a situation known as "informational masking" (Kidd *et al.*, 1998; Arbogast *et al.*, 2002; Gallun *et al.*, 2005; Shinn-Cunningham *et al.*, 2005; Ihlefeld and Shinn-Cunningham, 2007). In such circumstances, the effect of spatial cues is most evident at longer timescales, such as when selecting between already-segregated objects, or in streaming sound elements such as syllables together (Darwin and Hukin, 1997; Darwin and Hukin, 1999; Kidd *et al.*, 2005; Kidd *et al.*, 1994). Over short timescales, as when the grouping of simultaneous elements across frequency is assessed, the influence of spatial cues is usually weaker than that of other segregation cues (Buell and Hafter, 1991; Culling and Summerfield, 1995; Darwin and Hukin, 1997; Darwin and Hukin, 1999; Dye, 1990; Stellmack and Dye, 1993), including harmonicity (Moore *et al.*, 1986; de Cheveigne *et al.*, 1995; Roberts and

a)Author to whom correspondence should be addressed. Electronic mail: shinn@cns.bu.edu

Brunstrom, 1998), spectrotemporal continuity (Bregman, 1990), and common modulation (Cutting, 1975; Darwin, 1981; Hall *et al*., 1984; Cohen and Schubert, 1987; Schooneveldt and Moore, 1987).

The relatively weak role of spatial cues for grouping over local timescales may reflect the acoustics of realistic listening situations, in which reverberant energy and background noises randomly distort the spatial cues reaching the ears (Shinn-Cunningham *et al*., 2005; Roman and Wang, 2008; Mandel *et al*., 2010). Although the reliability of spatial information available instantaneously in a particular frequency channel is often low because of these acoustic factors, reliability can be improved by combining information across channels and time. Thus, once sound elements are grouped together locally in time and frequency, spatial auditory perception can be accurate, and can contribute more strongly to the grouping of sound over longer time scales. Nonetheless, spatial cues do impact grouping and segregation of concurrent sound elements when other spectrotemporal grouping cues are ambiguous or uninformative (e.g., see Shinn-Cunningham, 2005; Darwin and Hukin, 1999; Drennan *et al*., 2003; Best *et al*., 2007; Shinn-Cunningham *et al*., 2007), consistent with the notion that they provide some weak segregation information on their own even at local time scales.

The goal of this paper was to further explore the role of spatial cues for grouping spectrotemporal sound elements over short timescales (on the order of 100 ms). We focus exclusively on the role of ITDs, assessing their influence on both the perception of segregation (i.e., whether listeners hear multiple auditory objects when presented with a mixture of sound sources) and the perceived acoustic content of the resulting perceptual tokens. Although natural listening conditions often produce ILDs as well as ITDs, ITDs alone are sufficient to support the localization of complex sounds, and on their own can enhance speech intelligibility in complex mixtures (Culling *et al*., 2004; Edmonds and Culling, 2005; Gallun *et al*., 2005; Kidd *et al*., 2010; Best *et al*., 2006; Shinn-Cunningham *et al*., 2005). We restricted our investigations to ITDs in part because they could be cleanly manipulated without altering other factors (e.g., the relative levels of the sounds from different sources) that might affect performance on tasks assessing grouping even if they do not directly impact grouping itself.

ITDs are known to play a role in the segregation of concurrent sound elements, but the extent to which they suffice for deriving accurate representations of the acoustic content of a source remains unclear. Prior work on sound segregation has primarily utilized artificial sounds such as tones, or natural sounds such as speech. Both approaches have potential limitations. Experiments with simple artificial stimuli could overestimate the efficacy of grouping cues for deriving a sound source's content because the structure of such stimuli is far simpler than that of real-world signals. On the other hand, natural signals such as speech are sufficiently structured and familiar that many factors influence segregation, making it difficult to isolate the role of any single cue, such as ITDs.

To investigate the utility of ITD for recovering the acoustic content of sound sources, we used a class of novel synthetic stimuli introduced in a recent study of sound segregation (McDermott *et al*., 2011; Fig. 1). These synthetic stimuli have second-order spectrogram correlations matching those of natural sounds, but their spectrotemporal structure is otherwise unconstrained. As a result, they lack many of the grouping cues, such as harmonicity and sharp onsets and offsets, that promote segregation in natural signals like speech. Consistent with the notion that mixtures of these synthetic sources contain only weak grouping cues, listeners generally perceive mixtures of two such sources as a single source, and have difficulty identifying the acoustic content of the true sources even though individual synthetic sources are easily discriminable in isolation (McDermott *et al*., 2011). However, when presented with a sequence of mixtures in which each mixture contains a fixed "target," but a different "distractor" source, listeners hear the repeating target as segregated from the randomly varying background stream, and are relatively good at estimating the content of the target (McDermott *et al*., 2011). These findings suggest that the auditory system detects repeating structure embedded in mixture sequences and interprets this structure as a repeating source, which is then segregated from the other sounds in the mixture (see also Kidd *et al*., 1994).

Several features of these synthetic stimuli make them attractive for testing the role of spatial cues in sound segregation. First, because other grouping cues in these sounds are weak, even a modest role of spatial cues on source segregation may be revealed. Second, we can generate as many perceptually distinct synthetic sources as needed, all drawn from the same generative distribution. Third, the complex structure of the synthetic sources allows us to probe listeners' ability to estimate the acoustic content of the source. Fourth, the effect of source repetition on the same task provides a way to compare any ITD benefit to that of another known grouping cue without altering the local structure of the source.

We thus adapted the methods of McDermott and colleagues to measure the influence of ITD separation on segregation. Subjects were presented with mixtures of target and distractor sounds that were given either the same or different ITD, and judged whether or not a subsequent probe sound was present in the preceding mixture(s). To compare the effect of ITD to that of target repetition, mixtures were presented back to back in a sequence consisting of a repeating target sound with distractors that either differed or repeated from presentation to presentation. Across a variety of conditions, we found that the effect of ITD separation on performance was modest at best and was much smaller than the benefit of target repetition.

Despite the small benefit of ITD separation on objective measures of target identification, pairs of concurrent sources with distinct ITDs were generally perceived as two sound sources at different locations. We directly assessed sound localization to verify this subjective impression. We asked listeners to localize sounds in mixtures like those described above, using a graphical interface that allowed them to respond with one or multiple locations for each trial. We found that subjects were able to use this method to accurately localize two simultaneous sources, even in conditions

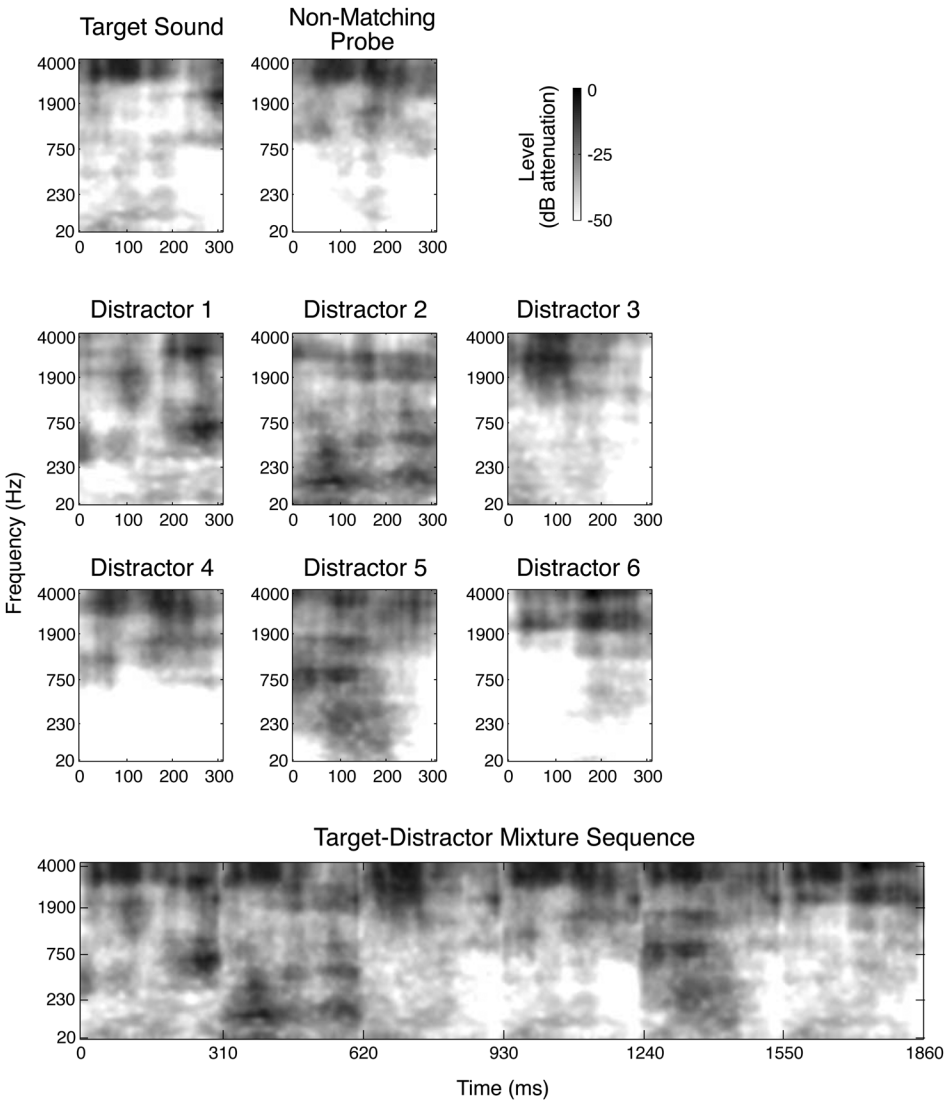

FIG. 1. Spectrograms of stimuli from an example trial of experiment 1. Top left shows the target sound, and next to it the non-matching probe. Second and third rows show six distractor stimuli, drawn from the same distribution as the target. Bottom row shows the mixture sequence that would be presented on a varying-distractor diotic trial, in which the target sound was successively presented with each of the six distractor sounds. Although the target sound is not visually apparent in the mixture sequence spectrogram, listeners typically hear the target repeating, and can distinguish it from the non-matching probe. Spectrograms were upsampled by a factor of four and blurred to lessen the visual edge artifacts of pixelation.

when they were not able to correctly identify the spectrotemporal content of these sources.

II. EXPERIMENT 1

A. Methods

1. Stimuli

We utilized the methods of an earlier paper (McDermott *et al.*, 2011), briefly reviewed here. The approach allows synthesis of novel sources that contain some of the basic spectrotemporal structure of natural sounds but that lack strong grouping cues. Specifically, we generated sounds whose log-energy time-frequency decompositions had spectrotemporal correlations like those measured in natural sounds such as spoken words and animal vocalizations (McDermott *et al.*, 2011). Time-frequency decompositions were generated using an analysis-synthesis subband transform (Crochiere *et al.*, 1976) based on an auditory filter bank with bandwidths that scaled like the bandwidths of human auditory filters. We used 39 filters with center frequencies ranging from 40–3967 Hz, equally spaced on an ERB_N scale (Glasberg and Moore, 1990). Each filter had a frequency response that rose from 0 to 1 and then back again as half a cycle of a sinusoid (i.e., the portion of a sinusoid covering the interval $[0\ \pi]$). Adjacent filters overlapped by 50%, such that the center frequency of one filter was the lower/upper absolute cutoff for each of its neighbors. The resulting subbands (i.e., filtered versions of the original signal) were divided into overlapping time windows by multiplying each subband by multiple 20 ms-long windows, each of which was a single cycle of a raised cosine function (adjacent windows overlapped by 50%). We hereafter refer to a single time window of a single subband as a "cell" in the time-frequency decomposition, and to the array of the log rms amplitude of each cell in the decomposition as the spectrogram.

We modeled the spectrogram as a multivariate Gaussian random variable, defined by a mean spectrogram and a covariance matrix consisting of the covariance between each pair of cells in the spectrogram. The value of each cell in the mean spectrogram was set proportional to the filter bandwidth, so that the average stimulus power spectrum was white. The covariance matrix was generated from exponentially decaying correlation functions resembling those measured in natural sounds: the correlation between pairs of

time-frequency cells fell with increasing separation in both time and frequency (decay constants of −0.075 per filter and −0.065 per time window; McDermott *et al.*, 2011). Each synthetic source was produced with the following steps. First, a sample spectrogram was drawn from the multivariate Gaussian generative distribution. Second, a sample of Gaussian white noise was drawn and its time-frequency decomposition was generated. Third, the amplitude in each time-frequency cell of the noise decomposition was adjusted (preserving the fine structure) so that the energy matched that of the corresponding sample spectrogram. The altered time-frequency decomposition was then inverted to yield a synthetic signal (i.e., the contents of each time window in each subband were summed to generate new subbands, the new subbands were passed through the auditory filter bank to ensure they remained bandlimited, and then the filtered subbands were summed to generate a full-bandwidth sound signal).

Stimuli were 310 ms in duration with 10 ms-long, half-Hanning-window onset and offset ramps. Spectrograms of example stimuli generated this way are shown in Fig. 1. The second-order spectrogram correlations induce structure that allows different stimuli to be discriminated from each other, but fail to capture other aspects of natural sound structure. In particular, although some degree of comodulation is induced across frequencies, the strong bottom-up grouping cues of abrupt common onset and harmonicity are lacking in our stimuli. See McDermott *et al.* (2011) for further details.

On each trial, a synthetic target source was generated with the procedure described above. In the fixed-distractor condition, a second stimulus, termed the "distractor," was randomly generated with the same procedure used for the target. The target and distractor were summed and the resulting mixture was presented six times, back-to-back. In the varying-distractor condition, six independent distractors were generated and each summed with a different presentation of the target; these target-distractor mixtures were concatenated and presented to the listener (an example stimulus sequence for this condition is depicted in the bottom row of Fig. 1). In both fixed- and varying-distractor conditions, the sequence of six mixtures of target + distractor(s) was followed by a 500 ms silent gap, after which a "probe" stimulus was presented. This probe was either identical to the target ("matching" probe) or was a new stimulus, distinct from both target and distractors ("non-matching" probe), with equal likelihood.

The non-matching probes were designed to test whether listeners had correctly estimated the spectrotemporal content of the target by segregating it from the mixture. To prevent listeners from distinguishing non-matching probes from matching probes on some basis other than the perceptual similarity of the probe and the perceived target, non-matching probes had statistics like those of the target sounds and had a similar acoustic relationship to the mixture. Specifically, non-matching probes were designed to never exceed the mixture's level at any point in the time-frequency decomposition, but also to be equal to the mixture over a portion of the decomposition, as both these properties were true of the target. The latter property reflects the fact that two sources generally have substantially different levels at most points in a time-frequency decomposition (Ellis, 2006); as a result, one of the sources is likely to dominate the mixture in any given time-frequency cell. To this end, we generated non-matching probes by first fixing a randomly chosen time slice (1/8 the duration of the stimulus, or 62.5 ms) to be equal to the corresponding time-slice of the mixture. We then sampled the remaining points in the time-frequency decomposition from the multivariate Gaussian distribution described above, conditioned on the fixed slice of the time-frequency decomposition having the same values as that of the mixture. In this way the non-matching probe retained largely the same covariance structure as the target. Any energy value of the resulting spectrogram that exceeded that of the mixture was set equal to the mixture value, so that the resulting sound was physically consistent with the mixture. Non-matching probes were not re-normalized following this procedure; their overall level was generally slightly lower than that of the target as a result, albeit by an amount that was never more than 1 dB (on average, the non-matching probes were 0.4 dB lower in level than the target). To avoid non-matching probes that were more similar to the mixture than were the targets, probes whose average difference from the mixture (computed point-wise from their spectrograms) was less than 7 dB were rejected. Enforcing these properties prevented listeners from performing the task successfully in conditions when the mixture was subjectively perceived as one source (McDermott *et al.*, 2011).

2. Task

Subjects were asked to judge whether or not the probe stimulus was identical to the target sound in the preceding mixtures. They responded by clicking one of four buttons, marked "yes," "maybe yes," "maybe no," and "no," indicating their level of confidence in their response. Subjects were instructed to attempt to use all four responses roughly equally throughout a given session.

Experiment 1 tested all combinations of three stimulus manipulations. The first of these manipulations was to either repeat the target with the same distractor for each of the six presentations, or to choose a new distractor for each presentation (McDermott *et al.*, 2011). The second manipulation was to vary the ITDs present in the target and distractor. We assumed that absolute location has relatively little effect on stimulus identification compared to spatial separation of stimuli, and thus manipulated only the relative position of the sources in three conditions: target left (−333 μs ITD) and distractor right (+333 μs ITD), target right and distractor left, or both diotic (0 μs ITD). The third manipulation involved the presence or absence of a visual cue (a red dot on the GUI) indicating the location of the target as left, right, or center (for the three spatial conditions above, respectively). The cue was presented prior to the sound presentation so that listeners would know where to attend to hear the target, maximizing the possible benefit of spatial separation (Kidd *et al.*, 2005; Best *et al.*, 2007). Listeners completed 32 trials (16 with a matching probe and 16 with a non-matching probe) for each of the 12 combinations of these conditions,

as well as an additional 24 control trials in which the distractor was absent, for a total of 408 trials. The control trials served to ensure subjects could perform the task. Any subject with an average performance in control trials below 0.75 was to be excluded; however, no subject met this exclusion criterion.

To familiarize themselves with the task prior to the experiment, subjects first completed a few 40-trial practice runs with feedback. In these runs, if the subject answered correctly (defined as "maybe yes" or "yes" if the probe matched the target, "maybe no" or "no" for a non-matching probe), the response GUI would display the word "correct" for half a second before moving on to the next trial. If the subject answered incorrectly, the word "incorrect" was displayed, and the target and probe were then played separately. Subjects then could opt to play the two stimuli again, or to continue to the next trial. Subjects were allowed to repeat these practice sessions until comfortable with the task. Most subjects performed one or two such runs. No feedback was provided in the main experimental sessions.

3. Analysis

For each of the 12 trial conditions ($2 \times 2 \times 3$, for fixed/varying distractor, cue present/absent, and left/right/center target location), a receiver operating characteristic (ROC) was generated from the responses (MacMillan and Creelman, 1991). The area under the ROC was used to quantify performance (0.5 represents chance performance and 1 represents perfect discrimination between matching and non-matching probes). Prior to performing statistical tests, the ROC areas were transformed with the inverse logistic function $f(x) = \log(x/(1-x))$ to make their distribution closer to Gaussian. Although this function results in extreme values for inputs close to 0 or 1, ROC areas outside the range [0.01, 0.99] were not observed in our data except in control conditions, which were not used in statistical tests.

4. Subjects

Thirty subjects participated in experiment 1. All subjects, ranging in age from 18 to 37, had clinically normal hearing (15 dB HL or better) as verified by pure-tone audiometry for frequencies between 250 Hz and 8 kHz. Subjects for each of the three experiments detailed in this manuscript gave written consent (overseen by the Boston University Charles River Campus IRB), and were paid an hourly wage in compensation for their efforts.

B. Results

Figure 2 shows means and standard errors of performance (ROC areas) averaged across all 30 subjects for each individual condition. All three factors produced significant effects [repeated measures ANOVA: $F(1,29) = 93.09$, $p<0.001$ for distractor variation, $F(1,29) = 4.79$, $p=0.04$ for visual cue, and $F(2,58) = 10.62$, $p<0.001$ for ITD]. Performance was better when the distractors varied, when target and distractors were given different ITD, and when the target's location was cued. The effect of varying the distractors, however, was considerably larger than that of either ITD separation or spatial cueing.

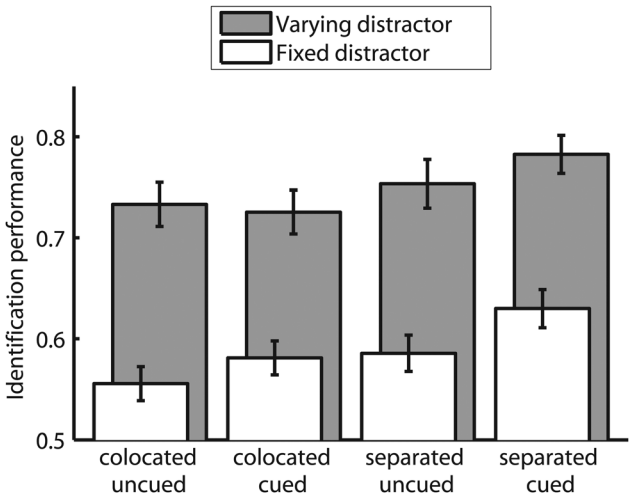

FIG. 2. Mean and standard errors of performance (ROC area) across 30 subjects in experiment 1, where the task was to judge if the probe was the target or not. The target was repeated six times, and the distractor was either fixed across the six presentations or varied. Performance improved when varying distractors were used (gray bars) compared to fixed distractors (white bars). Performance was only slightly better when the target and distractor were given different ITDs (right-most bars) than when the target and distractor were diotic (left-most bars). A visual cue indicating the target location also modestly improved performance.

Performance was poor, although significantly above chance [$t(29) = 3.17$; $p=0.004$] in the fixed-distractor conditions with no spatial separation. Subjects performed better in diotic, varying-distractor trials than for diotic, repeated-distractor trials (consistent with McDermott *et al.*, 2011). Averaged over all conditions including diotic and spatially separated trials, the mean performance benefit of varying distractors was 0.16.

In contrast, giving the target and distractors different ITDs had a modest effect: mean performance increased by only 0.025 and 0.053 for conditions without and with a visual cue, respectively, with a significant interaction between ITD and visual cue ($F(2,58) = 4.4$, $p=0.017$). This interaction suggests that ITD separation of targets was more useful when a visual cue indicated the location of the target (telling listeners where to direct spatial attention) than without such knowledge. One interpretation of this result could be that subjects could segregate the target and masker from the mixture using ITD, but that without the visual cue they may have selected the incorrect token to compare to the probe. However, the poor performance even when the visual cue was present indicates that much of the difficulty cannot be simply explained by such a failure of selection. Overall, while the effect of ITD separation was significant, it was quite small.

Subjects' average performance in control trials, in which no distractor was present, was very good (mean = 0.94, ranging from 0.79 to 1). These results, as well as results from the varying-distractor condition and from the other conditions in McDermott *et al.* (2011), indicate that the near-chance performance in the repeated distractor conditions is not due to lack of discriminability of the target/probe stimuli

themselves, but rather reflects difficulties in segregating the target from the target + distractor mixtures well enough to allow identification of the target content. ITD separation failed to produce a large improvement in performance both in conditions where it was poor to begin with (fixed distractors) and where it was well above chance (varied distractors).

III. EXPERIMENT 2

Experiment 2 aimed to determine if the weak benefit of ITD separation in experiment 1 depended on the degree of spectrotemporal overlap of the target and distractor. Using time-frequency masks (Brungart *et al.*, 2006), we manipulated the spectrotemporal overlap between the target and distractor and assessed the effect on performance.

A. Methods

Methods in experiment 2 were based on those of experiment 1. Only the differences between the two experiments are described here.

1. Stimuli

To control target-distractor overlap, we generated stimulus sets with three conditions: normal-overlap (as in experiment 1), min-overlap, and max-overlap (McDermott *et al.*, 2011; see also Brungart, 2001; Brungart *et al.*, 2006), examples of which are shown in Fig. 3. We first generated distractor stimuli for each target with the constraint that the target and distractor always overlapped in some spectrotemporal regions, but that the distractor also had energy in some spectrotemporal regions in which the target did not. These "root" distractors were used as-is for the "normal" overlap condition (second row of Fig. 3). As a result, they overlapped the target on average more than did the distractors of experiment 1, which were not so constrained. For the minimal overlap condition, we generated distractors by zeroing the regions of the root distractors that overlapped the target (fourth row of Fig. 3); and for the max overlap we zeroed regions that did not overlap the target (third row of Fig. 3). Distractors masked in this way were then renormalized so that the rms energy was equal to that of the original root distractor.

Specifically, to ensure that root distractors would produce suitable distractors for both the min- and max-overlap conditions, each of the following two criteria had to be met in at least 25% of time-frequency cells: (1) the target cell had "sufficient energy" (defined as having energy within 40 dB of the maximum in the spectrogram) but the corresponding mixture cell exceeded it in level by 5 dB (i.e., the distractor added energy in that cell that was likely to energetically mask the target in that cell); (2) the target cell did not have sufficient energy but mixture cell did. To generate a min-overlap distractor, we zeroed all distractor cells that contained masking energy (defined conservatively as having energy no less than 10 dB below that of the corresponding target cell for any target cells that had sufficient energy). To generate a max-overlap distractor, we zeroed distractor cells that had sufficient energy but for which the corresponding target cell did not, leaving energy only in places where the target had energy as well. Non-matching probes were generated as in experiment 1. Example time-frequency decompositions for normal, max-overlap, and min-overlap distractors are shown in Fig. 3 (left panels of second, third, and fourth rows, respectively).

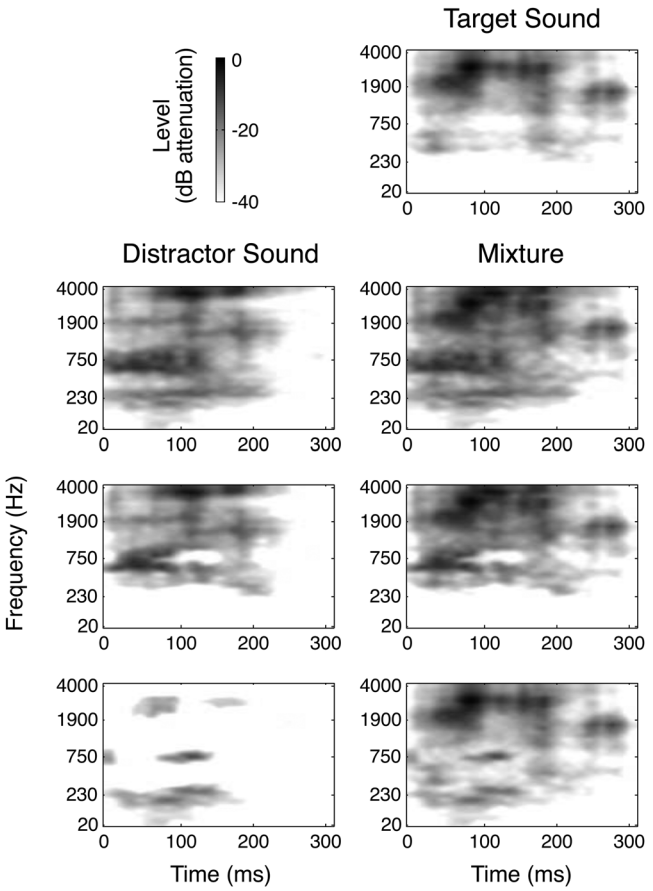

FIG. 3. Spectrograms of example stimuli from experiment 2. A target sound is shown at the top. The bottom three rows show the three distractor types (normal, max overlap and min overlap) for that target sound, created by masking the normal distractor (see experiment 2: Methods). To the right of each distractor is the spectrogram of the mixture of the target with that distractor. Aspect ratio and dB scale range differ from Fig. 1 to make the differences between distractor types easier to see.

2. Task

Because we were principally interested in exploring the role of ITD in this experiment, we did not include the varied distractor condition. Subjects were presented with a single mixture of a target and distractor, followed by a probe. There were two independent variables: (1) stimulus overlap, and (2) ITD separation. To maximize the possible benefit of ITD separation, on separated trials, the locations of target and distractor were always the same, with the target from the right (+333 μs ITD) and the distractor from the left (−333 μs ITD). Subjects were told beforehand where the targets would be located. On colocated trials, both target and distractor were presented diotically. In addition, every trial contained a visual cue indicating the target direction (either right for separated conditions, or center for diotic conditions).

As in experiment 1, subjects completed 32 trials (16 with matching probes and 16 with novel probes) of each of

the six possible conditions (three overlap conditions × two spatial configurations). We also included twelve control trials in which there was no distractor present, and 16 "easy" trials per condition in which the distractor was attenuated by 10 dB. This easy condition was included to help maintain motivation, as all of the non-control trials were expected to be difficult. It also served as an additional means to confirm that subjects were paying attention and were not simply guessing randomly. These conditions resulted in 300 trials per session, randomly ordered.

3. Subjects

Thirty-five normal-hearing subjects, 17 of whom previously completed experiment 1, participated in the experiment. Five of these subjects were excluded from analysis because their performance in control trials (in which the distractor was absent) was below 0.75.

B. Results

Performance on control trials for the 30 subjects included in the analysis ranged from 0.79 to 1 (mean = 0.92). Mean performance in "easy" trials (in which the distractor was attenuated by 10 dB), pooled across conditions, was 0.75, significantly above chance in each condition ($t(29) \geq 8.34$, $p < .001$ in all conditions).

Figure 4 displays the average results in each of the primary experimental conditions. Performance was poor overall, although it was better for the high overlap conditions, consistent with (McDermott *et al.*, 2011), producing a main effect of overlap (2-way repeated-measures ANOVA, $F(2,58) = 15.06$, $p < 0.001$). The reason for the improvement in performance in the high overlap condition is likely a result of differences in the constraints governing generation of the distractors in the different conditions, detailed above, which caused differences in the similarity between matching and non-matching probes in the different conditions.

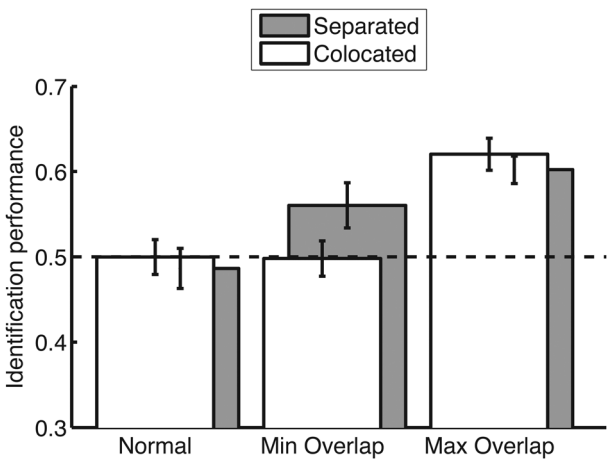

FIG. 4. Mean and standard error of performance (ROC area) in experiment 2, in which spectrotemporal overlap between the target and distractor was manipulated. The target was presented only once, along with a distractor. The dashed line represents chance performance. The effect of ITD was not significant, but there was a significant interaction between overlap and ITD (see experiment 2: Results).

As expected, we found a significant interaction between overlap and ITD ($F(2,58) = 3.24$, $p = 0.046$). This interaction was driven by a small ITD benefit for the min-overlap condition ($t(29) = 1.95$, $p = 0.03$, one-tailed; mean performance increased by 0.06), consistent with the intuition that whatever benefit ITD provides is enhanced when stimulus overlap is reduced. As the distractors in the other conditions were explicitly selected to overlap the target, we expected the benefit of ITD would be smaller in these conditions. Consistent with this expectation, we found no significant benefit of ITD separation for either the normal or max-overlap conditions. As a result, there was no main effect of ITD, unlike in experiment 1. These results confirm that time-frequency overlap reduces the utility of ITD cues for segregating the target and distractor. However, even in the minimal overlap condition, the effect of ITD was quite small in absolute terms, on par with what we observed in experiment 1. This may be because the min-overlap condition did not eliminate the potential effects of forward masking and suppression, which could have corrupted internal computations of ITD in much the same way that physical overlap does. The weak effect of ITD contrasts with the relatively large effect of target repetition in the varying-distractor conditions of experiment 1. While we did not test varying-distractor conditions in experiment 2, McDermott and colleagues observed ROC areas of ~0.8 for varying distractor conditions in their experiments with stimuli generated like those used here (McDermott *et al.*, 2011, Fig. S3).

IV. EXPERIMENT 3

Experiments 1 and 2 failed to show a substantial effect of ITD separation on the identification of targets embedded in mixtures. However, listeners nonetheless consistently reported that trials with spatially separated target and distractors were heard as two distinct sound sources at different spatial locations. Experiment 3 was designed to confirm this subjective observation. Listeners reported how many sources were perceived in a mixture, and then localized the perceived sources. We expected listeners to perceive two sources when target and distractor had different ITDs, even though ITDs alone did not allow listeners to accurately perceive the spectrotemporal content of targets in the mixtures.

A. Methods

1. Stimuli

Stimuli were generated as in experiment 1. As in experiment 1, on each trial the target + distractor mixture was repeated six times. We included only the fixed-distractor condition from experiment 1, as the varying-distractor mixtures were heard as two sources even without spatial separation.

2. Task

Subjects were instructed to localize the stimuli by clicking on a GUI that displayed an image of a cartoon head centered inside an arc spanning ±90°. For simplicity, we used a fixed linear mapping of ITD to spatial angle (500 μs mapped

to 70°) rather than obtaining an individualized ITD-to-angle mapping for each subject. No stimuli were presented from absolute ITDs greater than 500 μs, corresponding to 70° on the arc, but in responding subjects were permitted to click any location on the arc between ±90°. To help listeners reliably map perceptual locations to responses on this GUI, on every fifth trial during training and every tenth trial during the experiment, we presented a randomly selected stimulus in a moving sequence, with ITDs from −500 μs to +500 μs in steps of 167 μs (four samples), simultaneously displaying a yellow dot at the corresponding angle on the arc. No response was requested for these reminder sequences.

When subjects clicked on the GUI to indicate a response, a red dot would appear on the arc indicating their response. After every training trial, feedback was given: the stimulus was repeated, and a yellow dot was shown at the "true" location (based on the ITD-to-angle relationship as described) with the subject's red dot still visible. Each training session consisted of 80 trials.

There were two types of training sessions: single-source and two-source sessions. For single-source training sessions, the target was repeated four times, back-to-back on each trial. For two-source training sessions, two different sources were randomly selected and assigned distinct locations. The first source was played four times back-to-back, followed by four repetitions of the second source. Sequential presentation allowed a clear, unambiguous, "correct" location for each source so that feedback could be given during training. Subjects confirmed their answer for a given trial by clicking a button on the GUI that, during the training session, was enabled only when the correct number of locations was selected. Prior to clicking this button, subjects were free to adjust their responses until satisfied.

All subjects ran at least three single-source and two two-source training sessions. Single-source training sessions typically took between 10 and 15 min, and two-source sessions between 15 and 20 min. Subjects trained for approximately 1.5 h on their first day, ending the day with a 30-trial practice session of the main experiment. Subjects then came back for a second day in which they typically repeated a single-source training session and a 30-trial practice session, and then performed the full 192-trial experiment session (some subjects did additional training sessions until comfortable with the task).

For the main experiment session, each stimulus (either a single Gaussian-generated sound, as used in experiment 1, or a mixture of two such sounds) was repeated six times. There were three conditions: single-source, two-source colocated, and two-source separated. For each condition, the first source was given a random ITD from a uniform distribution between ±500 μs (discretized in integer sample delays). In the two-source separated condition, the second source's ITD was constrained to enforce a minimum separation between the two sources in the mixture: the second ITD was selected from a uniform distribution spanning from 208 μs (5 samples) contra-laterally away from the first-source ITD to a maximum of 500 μs on the side contralateral to the first source. For example, if the first source had an ITD of −290 μs (−7 samples), the second source's ITD was constrained to fall between −83 μs and +500 μs (−2 to +12 samples). If the first source had an ITD of 0 μs, the second source's ITD could be between −208 μs and −500 μs or between +208 μs and +500 μs, chosen randomly.

In two-source mixtures, the mixture was normalized to have the same RMS energy as a single stimulus. To further discourage listeners from attempting to use level as a cue to decide whether one or two sources were present, the level was roved by up to ±5 dB for each trial of each condition, and subjects were informed that level would be uninformative.

Subjects were told to (1) place either one or two points on the GUI to indicate whether they heard one or two sound sources, and (2) to place the points at the perceived locations of the source(s). Subjects were instructed to focus primarily on getting the number of sources correct, even if they had to simply guess the location of one or both of the sources. They were told that in two-source trials, these sources could be separated or colocated; if they heard two colocated sources they should click twice on the same location of the arc. The GUI at all times displayed a number telling the subjects how many objects were entered on the arc, so that they could easily tell if they had entered one or two objects at a given location.

B. Results

Figure 5 displays box plots summarizing, for each condition, the percentage of trials in which the number of sources was correctly identified [Fig. 5(a)], and the RMS localization error [Fig. 5(b)]. Subjects were generally able to correctly identify the number of sources in single-source trials and two-source-separated trials, averaging 91.1% and 89.5% correct, respectively. However, subjects typically responded with only a single location for two-source

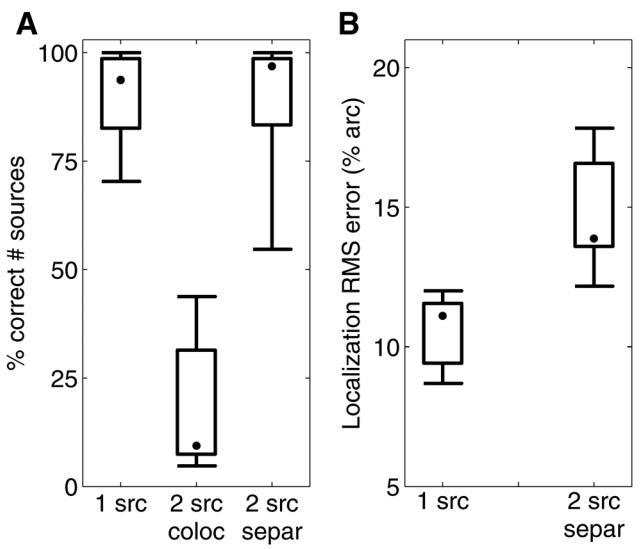

FIG. 5. Results of experiment 3, in which the task was to identify the location(s) of the one or two sound sources presented on a trial. (a) Box plots of the percentage of trials in which subjects correctly identified the number of stimuli contributing to the mixture. (b) Box plots of the mean RMS localization error for trials in which the subject correctly identified the number of stimuli in the mixture. The two-source-colocated condition is omitted as there were few trials in which the subject correctly identified the number of stimuli. Box plots show median (dot), the middle 75th percentile (5 of 7 subjects; boxes), and full range (whiskers).

colocated trials, and thus rarely responded correctly (12.5% of trials on average). This difference in percent correct between two-source colocated and two-source separated trials was significant (sign test, $p<0.02$), showing that ITD separation allows listeners to perceptually segregate a mixture of two synthetic sources.

Localization error between subjects' responses (measured as an angle on the arc) and the actual location of the source (in μs ITD) was computed by converting both numbers into a percentage of the total range represented by the arc ($\pm$90 degrees and $\pm$643 μs, respectively, mapped to $\pm$100%) and determining the difference between the response and the true location. When computing RMS localization error for the single-source or two-source-separated conditions, we considered only trials in which the correct number of responses was given. For two-source trials, the RMS error was computed by assuming the right-most response was an estimate of the location of the right-most stimulus, and likewise for the left. This method may underestimate the actual RMS error if for some trials the estimated positions of the stimuli were reversed.

The first author of this paper (whose data are not included in any other analyses) as well as the first subject participated in several pilot training sessions (seven sessions and ten sessions, respectively) with single sources only, each over multiple days, to gauge the expected accuracy of subjects using the localization GUI. The first author's lowest RMS error for a single session was 8.2%, while the first subject's lowest error was 13.9%. Other subjects' lowest localization error in single-source training sessions ranged from 8.1% to 10.4%. This error represents a smaller fraction of the arc than the minimum distance enforced between spatially separated stimuli in the experiment (16% of the arc, or 5 samples ITD), and provides a baseline to which we compare localization errors in the main session; we are not explicitly interested in absolute localization acuity.

The localization errors from the main experiment are summarized in Fig. 5(b). The RMS localization error was greater in the two-source-separated condition than the single-source condition (medians of 13.9% and 11.1%, respectively). These errors are again less than the minimum separation between stimuli, suggesting that our subjects were able to effectively use our interface to simultaneously localize two sources. For the two-source colocated trials, we did not analyze localization errors, as there were very few trials in which subjects answered with the correct number of sources (less than 10 trials for many subjects).

A closer look at the error for the two-source-separated condition revealed a distinctive pattern: the perceived location of stimuli with an ITD close to zero showed a systematic bias away from the other stimulus' ITD. This type of localization "repulsion" between elements of an auditory scene has been reported previously (Best *et al.*, 2005; Braasch and Hartung, 2002; Lee *et al.*, 2009; Schwartz and Shinn-Cunningham, 2010), and is associated with segregation of elements into distinct objects; when two sound elements group together, "attraction" typically occurs instead of repulsion (Lee *et al.*, 2009). We quantified repulsion for a given response by the localization error in the direction away from the other stimulus (positive numbers represent repulsion; negative numbers represent attraction).

Figure 6 displays the magnitude of the repulsion effect, plotted separately for stimuli located centrally and laterally (absolute ITDs less than 200 μs or greater than 300 μs, respectively). Consistent with prior findings that repulsion is stronger for stimuli localized around midline (Lee *et al.*, 2009), the mean repulsion magnitude was positive for all subjects for centrally located sources, but was close to zero for stimuli located laterally (Fig. 6). This repulsive effect at least partially explains why localization errors increased when two separated sources were present in the mixture, and is additional evidence that ITD produced perceptual segregation of the sources.

V. DISCUSSION

A. ITDs alone are ineffective at segregating source content

The primary purpose of these experiments was to determine if ITD separation would permit accurate segregation of a mixture of two complex sounds in the absence of other strong grouping cues. The novel contribution of our method was to isolate the effects of ITD on the segregation of unfamiliar stimuli with some degree of naturalistic complexity, such that the content of a perceived source could be probed in detail. We found that ITD separation produced little improvement in our target identification task. Although

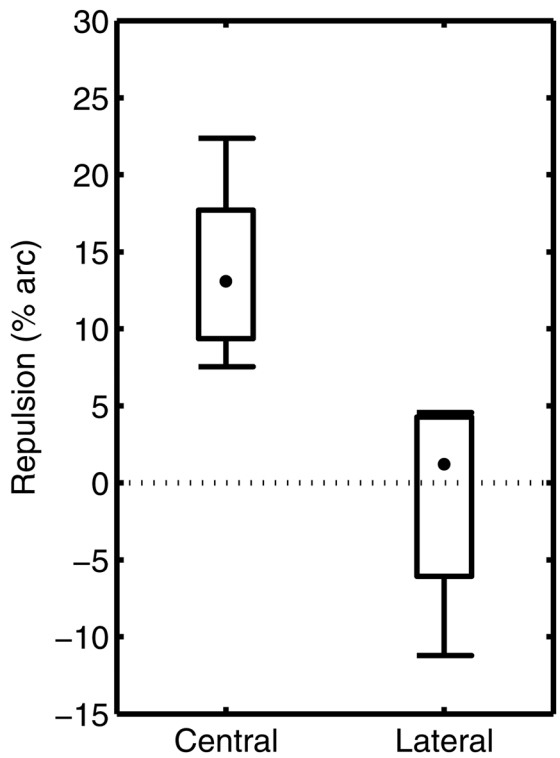

FIG. 6. Box plots of mean repulsion for each subject, defined as the localization error in the direction away from the other stimulus' ITD in trials with two spatially separated sources. Mean repulsion is plotted separately for stimuli from central (stimuli with ITD within $\pm$200 μs) and lateral (stimuli with absolute ITD greater than 300 μs) locations. Box plots show median (dot), the middle 75th percentile (5 of 7 subjects; boxes), and full range (whiskers).

giving two concurrent sources distinct ITDs produced an improvement in identification that was statistically significant, the effect was small in an absolute sense even when the direction of the target was known. Particularly given the ability of subjects to reliably identify targets when they were presented with distractors that varied over time (experiment 1), our results indicate that ITD alone does not enable accurate segregation of sound sources from mixtures.

By contrast, when subjects in previous studies have been asked to identify speech tokens embedded in a target sentence within a mixture of speakers, ITDs on their own have raised performance from below 70% correct to above 90% (Gallun *et al*., 2005). Benefit from spatial cues can also be measured as the difference in the target level needed for the target to be accurately identified. Measured this way, ITD can provide "spatial release from masking" (SRM) of anywhere from several dB to tens of dB (e.g., see Edmonds and Culling, 2005; Kidd *et al*., 2010). An ITD benefit is typically obtained even with speech filtered into low- or high-frequency bands, indicating that both conventional low-frequency ITD as well as high-frequency envelope ITD can contribute to SRM (Edmonds and Culling, 2005; Kidd *et al*., 2010). Differences in the tasks used in such studies as well as the different metrics used to measure performance make it difficult to quantitatively compare the effects seen in previous studies to those reported here. However, the benefit of ITD that we found was much smaller than that of varying distractors across repetitions of a target. Viewed this way, ITD was clearly a weak contributor to the identification of our novel, complex stimuli within mixtures, unlike what has been found for the identification of speech utterances from mixtures.

Why were our results with complex synthetic sounds different from those in previous studies with speech? In principle, ITDs and other spatial cues could benefit performance in a segregation task either by facilitating the segregation of a target source from background sources, or by making it easier to focus attention on the "correct" source amongst sources that are already well segregated (Gallun *et al*., 2005; Shinn-Cunningham, 2005; Ihlefeld *et al*., 2007; Shinn-Cunningham, 2008). Many of the previously documented benefits of spatial cues on understanding speech in sound mixtures likely derive from their influence on object selection. Natural speech contains many grouping cues that support segregation; the bottleneck in identifying a speech utterance from a mixture is thus often primarily in correctly focusing selective attention rather than in object formation. In contrast, our stimuli do not possess other strong grouping cues, and performance in our task appears to have been primarily limited by poor segregation rather than difficulties with selection. In both the visual-cue condition of experiment 1 and in experiment 2, knowing where to listen had only a modest impact on performance, indicating that selection was not at issue. Our results are thus compatible with previous findings showing a benefit of ITD on source selection, while suggesting that ITD cues alone have a weak influence on segregation.

Why was ITD of so little benefit for the segregation of our stimuli? The most likely explanation is that ITD cues are unreliable when sounds at different locations are mixed together. When concurrent sound sources overlap in time and frequency, the ITD cues at a particular point in time and frequency often do not reliably identify either source's direction (Shinn-Cunningham *et al*., 2005; Roman and Wang, 2008; Mandel *et al*., 2010). Perceptual estimates of target content derived exclusively from ITD are thus likely to contain errors. Under this hypothesis, the sources that subjects perceived are likely to have contained some, but not all, of the spectrotemporal content of the true target, as well as some of the content of the distractor, which could explain the weak effects of ITD in our identification task. The results of experiment 2 are consistent with this explanation, as the ITD benefit was greater (though still small) when overlap was reduced, conditions in which ITD information should be more accurate. Moreover, although the minimal-overlap conditions reduced the physical time-frequency overlap of the target and distractor, additional interference between target and distractor could occur within the auditory system. This additional stimulus interaction due to peripheral auditory processing (e.g., forward masking, suppression, etc.) may further corrupt ITD cues, increasing the percentage of time-frequency cells with mixture-corrupted ITDs, and potentially explaining the weak effect of ITD even in our condition of minimal physical overlap. Source overlap in natural auditory scenes might have less effect on other grouping cues, such as common onset or harmonicity, which could explain why listeners rely very little on ITDs to segregate simultaneous sound sources. Our results are thus consistent with the idea that ITDs are a weak contributor to concurrent sound segregation, and are not particularly useful on their own for deriving accurate representations of sound source content.

Could the limited benefit of ITD be specific to the ITD-only manipulation that we used? Classically, ILD was thought to be the primary high-frequency lateralization cue (as in the "duplex" theory of sound localization, first proposed by Lord Rayleigh around 1900). This might seem to suggest that manipulating ITD alone with broadband sounds would be a weak test of spatial cues. However, although ILDs are influential at high frequencies, ITDs in the envelopes of high-frequency sounds also influence lateralization, even when ILDs are zero, as in our stimuli (Henning, 1974; McFadden and Pasenen, 1976; Nuetzel and Hafter, 1976; Macpherson and Middlebrooks, 2002; Bernstein and Trahiotis, 1994; Best and Shinn-Cunningham, 2007). Such high-frequency ITDs in fact aid speech recognition in mixtures (Edmonds and Culling, 2005; Kidd *et al*., 2010). Moreover, low-frequency ITD (encoded by phase-locked neural responses to acoustic fine structure) typically dominates sound lateralization of broadband signals (e.g., see Wightman and Kistler, 1992; Macpherson and Middlebrooks, 2002). There is thus reason to think that an ITD manipulation would affect the segregation of a broadband sound. That said, the relative contributions of low- and high-frequency ITDs to segregation have yet to be examined in detail. It is possible that the influence of spatial cues on performance would have been greater if we had used stimuli with restricted frequency content, or if we had manipulated other spatial attributes, including ILDs and spectral cues.

B. ITDs alone are effective for determining competing source locations

Despite the weak effect of ITD on segregation as measured by an identification task, subjects in experiments 1 and 2 reported the perception of two distinct sources when the target and distractor were spatially separated. We verified these subjective reports in experiment 3, in which subjects reliably judged the presence and location of two sources in a mixture when those two sources were given distinct ITDs, but not otherwise. Furthermore, subjects demonstrated repulsion between the perceived locations of the two stimuli when they were near the midline, providing further evidence that they perceived two competing objects in these mixtures (Lee *et al.*, 2009). We conclude that ITD on its own produces perceptual segregation of stimuli in an auditory scene, but that the resulting estimates of sound source content are inaccurate when no other grouping cue helps listeners segregate the competing sounds.

When stimuli in a mixture contain other grouping cues, as in most natural scenes, the effect of spatial separation is highlighted in tasks when grouping cues for a particular sound element are ambiguous (e.g., Shinn-Cunningham *et al.*, 2007). We used synthetic sources with weak grouping cues to maximize the potential role of ITD in source segregation, and yet still found that ITD alone was insufficient for accurate segregation. Given that natural sounds generally contain other, stronger cues for grouping simultaneous sources, it is thus unlikely that ITD make an important contribution to our ability to accurately segregate simultaneous sounds in natural auditory scenes.

VI. SUMMARY AND CONCLUSIONS

(1) Complex synthetic sound sources with impoverished grouping cues were used to study sound segregation. Consistent with past results using such sounds (McDermott *et al.*, 2011), subjects were able to identity a target sound source from multiple presentations of unique target-distractor mixtures, but not from a single combination of target and distractor.
(2) ITD separation of the target and distractor produced only a weak improvement in the identification of the spectrotemporal content of the target, whether the distractor varied or not. The improvement was largest, though still small in an absolute sense, when a visual cue indicated the direction of the target.
(3) Spectrotemporal overlap of target and distractor reduced the benefit derived from ITD, but the improvement to target identification due to ITD separation was still modest even when the distractors were engineered to have little physical spectrotemporal overlap with the target.
(4) When the task was to localize, rather than identify, the stimuli in the mixture, ITD separation produced perceptual segregation as expected: subjects accurately reported two distinct object locations, and demonstrated repulsion between the two objects, consistent with perceptual segregation.
(5) Overall, our results indicate that ITD can promote segregation, but is insufficient to accurately estimate the spectrotemporal content of sound sources in auditory scenes.

ACKNOWLEDGMENTS

This work was supported by grant NIDCD ROI DC009477 (to B.S.C.), training grant NIDCD T32 DC00038 (supporting A.S.), and the Howard Hughes Medical Institute (supporting J.H.M.).

Akeroyd, M. A. (**2004**). "The across frequency independence of equalization of interaural time delay in the equalization-cancellation model of binaural unmasking," J. Acoust. Soc. Am. **116**, 1135–1148.
Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.
Bernstein, L. R., and Trahiotis, C. (**1994**). "Detection of interaural delay in high-frequency sinusoidally amplitude-modulated tones, two-tone complexes, and bands of noise," J. Acoust. Soc. Am. **95**, 3561–3567.
Best, V., Gallun, F. J., Carlile, S., and Shinn-Cunningham, B. G. (**2007**). "Binaural interference and auditory grouping," J. Acoust. Soc. Am. **121**, 1070–1076.
Best, V., Gallun, F. J., Ihlefeld, A., and Shinn-Cunningham, B. G. (**2006**). "The influence of spatial separation on divided listening," J. Acoust. Soc. Am. **120**, 1506–1516.
Best, V., Ozmeral, E. J., and Shinn-Cunningham, B. G. (**2007**). "Visually-guided attention enhances target identification in a complex auditory scene," J. Assoc. Res. Otolaryngol. **8**, 294–304.
Best, V., Schaik, A. van, Jin, C., and Carlile, S. (**2005**). "Auditory spatial perception with sources overlapping in frequency and time," Acustica **91**, 421–428.
Braasch, J., and Hartung, K. (**2002**). "Localization in the presence of a distracter and reverberation in the frontal horizontal plane. I. Psychoacoustical data," Acustica **88**, 942–955.
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), pp. 1–773.
Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acustica **86**(1), 117–128.
Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109.
Brungart, D. S., P. S. Chang, and Simpson, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**(6), 4007–4018.
Buell, T. N., and Hafter, E. R. (**1991**). "Combination of binaural information across frequency bands," J. Acoust. Soc. Am. **90**(4), 1894–1900.
Carlyon, R. P. (**2004**). "How the brain separates sounds," Trends Cogn. Sci. **8**(10), 465–471.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and two ears," J. Acoust. Soc. Am. **25**(5), 975–979.
Cohen, M. F., and Schubert, E. D. (**1987**). "The effect of cross-spectrum correlation on the detectability of a noise band," J. Acoust. Soc. Am. **81**, 721–723.
Crochiere, R. E., Webber, S. A., and Flanagan, J. L. (**1976**). "Digital coding of speech in sub-bands," Bell Syst. Tech. J. **55**, 1069–1085.
Culling, J. F. (**2007**). "Evidence specifically favoring the equalization-cancellation theory of binaural unmasking," J. Acoust. Soc. Am. **122**, 2803–2813.
Culling, J. F., Hawley, M. L., and Litovsky, R. Y. (**2004**). "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," J. Acoust. Soc. Am. **116**, 1057–1065.
Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**(2), 785–797.
Cutting, J. E. (**1975**). "Aspects of phonological fusion," J. Exp. Psychol. **104**, 105–120.
Darwin, C. J. (**1981**). "Perceptual grouping of speech components different in fundamental frequency and onset-time," Q. J. Exp. Psychol. **3A**, 185–207.

Darwin, C. J. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**, 327–333.
Darwin, C. J., and Hukin, R. W. (**1997**). "Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity," J. Acoust. Soc. Am. **102**, 2316–2324.
Darwin, C. J., and Hukin, R. W. (**1999**). "Auditory objects of attention: The role of interaural time differences," J. Exp. Psychol. **25**(3), 617–629.
de Cheveigne, A., S. McAdams, Laroche, J., and Rosenberg, M. (**1995**). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. **97**(6), 3736–3748.
Drennan, W. R., Gatehouse, S., and Lever, C. (**2003**). "Perceptual segregation of competing speech sounds: the role of spatial location," J. Acoust. Soc. Am. **114**, 2178–2189.
Drullman, R., and Bronkhorst, A. W. (**2000**). "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," J. Acoust. Soc. Am. **107**, 2224–2235.
Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. **35**, 1206–1218.
Dye, R. H. (**1990**). "The combination of interaural information across frequencies: Lateralization on the basis of interaural delay," J. Acoust. Soc. Am. **88**(5), 2159–2170.
Edmonds, B. A., and Culling, J. F. (**2005**). "The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay," J. Acoust. Soc. Am. **117**, 3069–3078.
Ellis, D. P. W. (**2006**). "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, edited by D. Wang and G. J. Brown (John Wiley and Sons, Hoboken, NJ), pp. 115–146.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.
Gallun, F. J., Mason, C. R., and Kidd, J. (**2005**). "Binaural release from informational masking in a speech identification task," J. Acoust. Soc. Am. **118**, 1614–1625.
Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**(1–2), 103–138.
Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectrotemporal pattern analysis," J. Acoust. Soc. Am. **76**, 50–56.
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 interferer," J. Acoust. Soc. Am. **115**, 833–843.
Henning, G. B. (**1974**). "Detectability of interaural delay in high-frequency complex waveforms," J. Acoust. Soc. Am. **55**, 84–90.
Ihlefeld, A., and Shinn-Cunningham, B. G. (**2007**). "Spatial release from energetic and informational masking in a selective speech identification task," J. Acoust. Soc. Am. **123**(6), 4369–4379.
Kidd, G., Jr., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**). "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**, 422–431.
Kidd, G., Jr., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (**2005**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**, 3804–3815.
Kidd, G., Mason, C. R., Best, V., and Marrone, N. (**2010**). "Stimulus factors influencing spatial release from speech-on-speech masking," J. Acoust. Soc. Am. **128**, 1965–1978.
Kidd, G. J., Mason, C., and Deliwala, P. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**(6), 3475–3480.
Lee, A. K. C., Deane-Pratt, A., and Shinn-Cunningham, B. G. (**2009**). "Localization interference between components in an auditory scene," J. Acoust. Soc. Am. **126**(5), 2543–2555.
MacMillan, N. A., and C. D. Creelman (**1991**). *Detection Theory: A User's Guide* (Cambridge University Press, New York), pp. 51–77.
Mandel, M. I., Bressler, S., Shinn-Cunningham, B. G., and Ellis, D. (**2010**). "Evaluating source separation algorithms with reverberant speech," IEEE Trans. Audio Speech Language Process. **18**, 1872–1883.
McDermott, J. H. (**2009**). "The cocktail party problem," Curr. Biol. **19**, R1024–R1027.
McDermott, J. H., Wrobleski, D., and Oxenham, A. (**2011**). "Recovering sound sources from embedded repetition," Proc. Natl. Acad. Sci. **108**, 1188–1193.
Macpherson, E. A., and Middlebrooks, J. C. (**2002**). "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," J. Acoust. Soc. Am. **111**, 2219–2236.
McFadden, D., and Pasanen, E. G. (**1976**). "Lateralization at high frequencies based on interaural time differences," J. Acoust. Soc. Am. **59**, 634–639.
Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (**1986**). "Thresholds for hearing mistuned partials as separate tones in harmonic complexes," J. Acoust. Soc. Am. **80**, 479–483.
Nuetzel, J. M., and Hafter, E. R. (**1976**). "Lateralization of complex waveforms: Effects of fine structure, amplitude, and duration," J. Acoust. Soc. Am. **60**, 1339–1346.
Roberts, B., and J. M. Brunstrom (**1998**). "Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes," J. Acoust. Soc. Am. **104**(4), 2326–2338.
Roman, N., and Wang, D. (**2008**). "Binaural tracking of multiple moving sources," IEEE Trans. Audio Speech Language Process. **16**(4), 728–739.
Schooneveldt, G. P., and Moore, B. C. J. (**1987**). "Comodulation masking release CMR: Effects of signal frequency, flanking-band frequency, masker bandwidth, flanking-band level, and monotic versus dichotic presentation of the flanking band," J. Acoust. Soc. Am. **82**, 1944–1956.
Schwartz, A. H., and Shinn-Cunningham, B. G. (**2010**). "Dissociation of perceptual judgments of 'what' and 'where' in an ambiguous auditory scene," J. Acoust. Soc. Am. **128**, 3041–3051.
Shamma, S. A. and C. Micheyl (**2010**). "Behind the scenes of auditory perception," Curr. Opin. Neurobiol. **20**, 361–366.
Shinn-Cunningham, B. G. (**2005**). "Influences of spatial cues on grouping and understanding sound," in *Proceedings of Forum Acusticum*, Budapest, Hungary.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**(5), 182–186.
Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, and Larson, E. (**2005**). "Bottom-up and top-down influences on spatial unmasking," Acta Acust. Acust. **91**, 13.
Shinn-Cunningham, B. G., Kopco, N., and Martin, T. (**2005**). "Localizing sources in a classroom: Binaural room impulse responses," J. Acoust. Soc. Am. **117**, 3100–3115.
Shinn-Cunningham, B. G., Lee, A. K. C., and Oxenham, A. J. (**2007**). "A sound element gets lost in perceptual competition," Proc. Natl. Acad. Sci. **104**(29), 12,223–12,227.
Stellmack, M. A., and Dye, J. (**1993**). "The combination of interaural information across frequencies: The effects of number and spacing of components, onset asynchrony, and harmonicity," J. Acoust. Soc. Am. **93**(5), 2933–2947.
Wan, R., Durlach, N. I., and Colburn, H. S. (**2010**). "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," J. Acoust. Soc. Am. **128**, 3678–3690.
Wightman, F. L., and Kistler, D. J. (**1992**). "The dominant role of low-frequency interaural time differences in sound localization," J. Acoust. Soc. Am. **91**, 1648–1661.