Influences of auditory object formation on phonemic restoration[a)]

Barbara G. Shinn-Cunningham[b)]
Hearing Research Center, Department of Cognitive and Neural Systems and Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02421 and Speech and Hearing Bioscience and Technology Program, Harvard-MIT Division of Health Sciences and Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139

Dali Wang
Hearing Research Center and Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02421

(Received 17 August 2007; revised 6 October 2007; accepted 8 October 2007)

In phonemic restoration, intelligibility of interrupted speech is enhanced when noise fills the speech gaps. When the broadband envelope of missing speech amplitude modulates the intervening noise, intelligibility is even better. However, this phenomenon represents a perceptual failure: The amplitude modulation, a noise feature, is misattributed to the speech. Experiments explored whether object formation influences how information in the speech gaps is perceptually allocated. Experiment 1 replicates the finding that intelligibility is enhanced when speech-modulated noise rather than unmodulated noise is presented in the gaps. In Experiment 2, interrupted speech was presented diotically, but intervening noises were presented either diotically or with an interaural time difference leading in the right ear, causing the noises to be perceived to the side of the listener. When speech-modulated noise and speech are perceived from different directions, intelligibility is no longer enhanced by the modulation. However, perceived location has no effect for unmodulated noise, which contains no speech-derived information. Results suggest that enhancing object formation reduces misallocation of acoustic features across objects, and demonstrate that our ability to understand noisy speech depends on a cascade of interacting processes, including glimpsing sensory inputs, grouping sensory inputs into objects, and resolving ambiguity through top-down knowledge. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804701]

PACS number(s): 43.66.Pn, 43.71.Rt, 43.66.Ba, 43.71.An [JCM] Pages: 295–301

I. INTRODUCTION

In everyday settings, the speech we hear is often partially masked by other sound sources, such as other talkers and events (Cherry, 1953). Our ability to communicate using noisy, ambiguous speech can be attributed in part to the redundancy in meaningful speech, which allows us to fill in masked or missing portions of the attended signal (Cooke, 2006). For instance, over a range of interruption rates, listeners are able to understand speech relatively well when half of the speech signal is replaced by silence (Miller and Licklider, 1950; Powers and Wilcox, 1977).

Speech intelligibility is even better when the speech gaps are filled in by unmodulated, steady-state noise, presumably because perceptual "filling in" of an interrupted speech signal is more automatic and complete than when there are sudden, audible silences (Warren, 1970; Powers and Wilcox, 1977; Bashford *et al.*, 1992). This filling in is informed by our expectations of the likely content of meaningful speech at every level of analysis, from continuity of spectrotemporal energy in the sound to lexical, linguistic, and semantic constraints (Warren, 1970; Bashford *et al.*, 1992; Warren *et al.*, 1994, 1997; Petkov *et al.*, 2007). While some of these expectations are learned (e.g., filling in a missing phoneme to generate a meaningful word in a given sentence), others may be hard-wired (e.g., perceiving a frequency glide interrupted by noise as if the glide is continuous; see Bashford *et al.*, 1992; Bailey and Herrmann, 1993; Darwin, 2005; Petkov *et al.*, 2007).

When the broadband temporal amplitude of missing speech is used to amplitude modulate the noise presented in speech gaps (henceforth referred to as speech-modulated noise), intelligibility is enhanced compared to when the noise is unmodulated (Bashford *et al.*, 1996). At first glance, this result is unsurprising—providing more speech-derived information in the input stimulus enhances intelligibility.

However, the speech and speech-modulated noise are perceived as distinct auditory objects (Bregman, 1990). Evidence suggests that listeners actively attend to one auditory object at a time in most situations (e.g., see Best *et al.*, 2006), consistent with the biased-competition model of visual attention (Desimone and Duncan, 1995). Thus, the improvement in speech intelligibility must come about because a feature of the noise (its modulation) is incorrectly bound with a competing object (the speech), an example of an "illusory conjunction" (Treisman and Gelade, 1980; Dyson and

[a)] Portions of this work were presented at the 2007 Mid-Winter Meeting of the Association for Research in Ogo-larynology, Denver, Colorado.

[b)] Author to whom correspondence should be addressed. Electronic mail: shinn@bu.edu

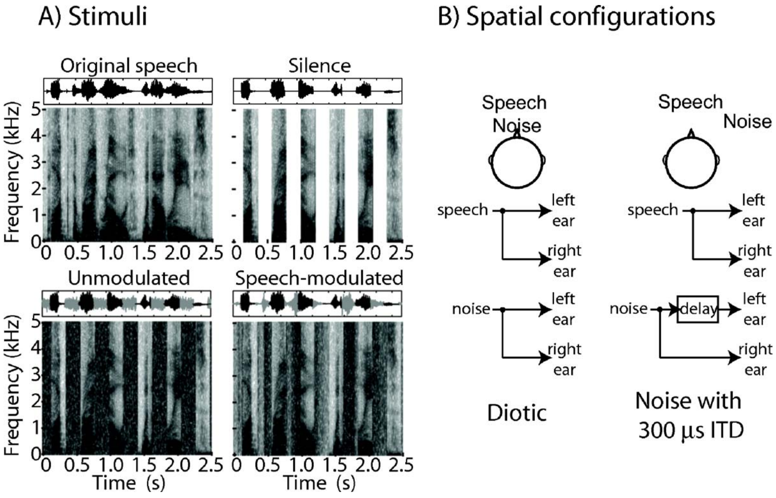

FIG. 1. (Color online) Stimuli and stimulus configurations. (A) Example of original speech and periodically interrupted speech when interruptions are filled with silence, unmodulated noise, and speech-modulated noise. Gray-scaled panels are spectrograms, showing the relative energy as a function of time (abscissa) and frequency (ordinate), with black corresponding to high energy and white to low energy. Small insets above each spectrogram plot the time domain wave forms, with black showing the unprocessed signal and gray showing the inserted noise wave forms. (B) Spatial configurations in Experiment 2 varied the ITD of the noise so that it was either heard at midline (left) or to the right of midline (right).

Quinlan, 2003). Illusory conjunctions of features in nonspeech stimuli are most likely to occur when the cues driving auditory object formation are ambiguous (Hall *et al.*, 2000). This suggests that manipulating acoustic grouping cues to increase the perceptual segregation of the speech and the speech-modulated noise might affect speech intelligibility.

We reasoned that strengthening the perceptual segregation of the noise and speech should reduce the likelihood that a noise feature would be misattributed to the speech. Increased segregation of speech and noise should not have any impact on intelligibility of interrupted speech presented with unmodulated noise in the gaps, which provides no information about how to complete missing speech. Instead of possessing a feature derived from the speech, unmodulated noise simply serves as a plausible masker of the missing speech, encouraging perceptual filling in (Warren, 1970; Hall *et al.*, 2000; Darwin, 2005). In contrast, if low-level auditory object formation affects speech perception, improved segregation of the speech-modulated noise and interrupted speech should decrease the likelihood of integrating the modulation (a noise feature) with the speech, and should therefore degrade speech intelligibility.

The dominant cues driving auditory object formation are spectrotemporal (Bregman, 1990; Darwin and Carlyon, 1995); however, spatial cues play a larger role in object formation when other cues are ambiguous (Darwin and Hukin, 1997, 1998; Freyman *et al.*, 2001; Shinn-Cunningham *et al.*, 2007), as when interrupted speech is presented with speech-modulated noise. We therefore manipulated the perceived spatial separation of interrupted speech and noise to affect the perceptual segregation of the speech and noise, and to see if segregation affected speech intelligibility.

We found that these low-level cues affected speech understanding when the interfering noise was modulated by the speech envelope, but not when the noise was unmodulated. These results demonstrate that low-level auditory cues affect speech just as they affect other, less specialized acoustic signals.

II. METHODS

A. Subjects

Nine normal-hearing subjects performed the tasks (five in Experiment 1 and four in Experiment 2). Subjects were recruited through on-campus advertisement, and all were students at Boston University (between ages 23 and 35). None had prior experience with psychophysical tasks, or with the corpus of test materials employed. All participants had puretone thresholds of 20 dB HL or better at all frequencies in the range from 250 to 8000 Hz, in both ears, and their threshold at 500 Hz was 15 dB HL or better. All subjects gave informed consent to participate in the study, as overseen by the Boston University Charles River Campus Institutional Review Board.

B. Equipment

Stimuli were processed in MATLAB (Mathworks, Natick, MA) using a sampling rate of 25 kHz. The stimuli were processed in MATLAB and sent to Tucker-Davis Technologies hardware for D/A conversion and attenuation before presentation over Sennheiser HD580 headphones. Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. MATLAB was used to control the stimulus presentation, to record responses, and to analyze results.

C. Stimuli

Speech sentences were from the Harvard IEEE corpus (IEEE, 1969). Sentences were periodically interrupted so that 50% of each signal was replaced by silence. This was accomplished by multiplying each sentence with a square wave (ranging between zero and one) with a 50% duty cycle. The periodicity of the periodic square wave was chosen to match rates that in past studies elicited large improvements in speech intelligibility when unmodulated noise filled in the speech gaps (Powers and Wilcox, 1977). Three rates, equal to 1.5, 2.2, and 3.0 Hz, were used.

In some conditions, the silent speech gaps were filled in, either with unmodulated white noise or speech-modulated noise. The speech-modulated noise was generated by multiplying unmodulated white noise by the Hilbert envelope of the speech that was missing in the gap (see Fig. 1). The average long-term, broadband root-mean-square intensity of the speech and noise were matched across the stimulus set; however, the spectra were not matched.[1]

The sentences used in each condition were chosen randomly from the 720 sentences making up the corpus. No sentence was presented more than once to any subject.

D. Spatial cues and stimulus conditions

In all conditions of both experiments, the interrupted speech was presented diotically.

Experiment 1 compared intelligibility of interrupted speech with silent gaps, unmodulated noise, and speech-modulated noise [see Fig. 1(a)]. Like the interrupted speech, both unmodulated noise and speech-modulated noise were presented diotically in Experiment 1, so that both speech and intervening noise were heard in the center of the head.

Experiment 2 compared intelligibility of interrupted speech with unmodulated and speech-modulated noise, with the speech and noise either collocated or perceived from different directions. In the collocated conditions, all stimuli were diotic, while in the spatially separated conditions, the noise (either unmodulated or speech modulated) was presented with an interaural time difference of 300 μs leading to the right ear. Thus, in the collocated configurations, both speech and noise were heard at the same, midline location [see the left-hand side of Fig. 1(b)]. In the spatially separated configurations, the speech was perceived at midline and the noise to the right of the listener [see the right-hand side of Fig. 1(b); subjectively, the off-midline perceptual locations of the unmodulated and speech-modulated noises with the 300 μs ITD were indistinguishable, and to the right of midline].

E. Procedure

Each listener performed four experiment sessions (at most one per day), each of which lasted approximately 1 h. Within each session, multiple experimental blocks were presented. In each block, the stimulus type and spatial configuration were fixed, but the interruption rate was randomly chosen on a trial-by-trial basis (with all three rates presented an equal number of times in a block). In each session, listeners performed one block of trials for each combination of stimulus type and spatial configuration, in random order (different in each session and for each listener). Thus, in each session, a listener performed an equal number of all possible combinations of stimulus type, spatial configuration, and interruption rate. There was no evidence that performance improved from session to session, and the randomization of the order of conditions across subjects and sessions ensured that any such learning effects would be averaged out, if they did exist.

In Experiment 1, which had three stimulus types and one spatial configuration, each listener performed three blocks of 60 trials each in each of the four sessions. In each block, listeners performed 20 trials at each interruption rate. Across the four experimental sessions, each subject performed a total of 80 repetitions (20 trials/session × 4 sessions) for each combination of stimulus type (three—silent gaps, unmodulated noise, and speech modulated noise), spatial configuration (only collocated in Experiment 1), and repetition rate (three: 1.5, 2.2, and 3.0 Hz).

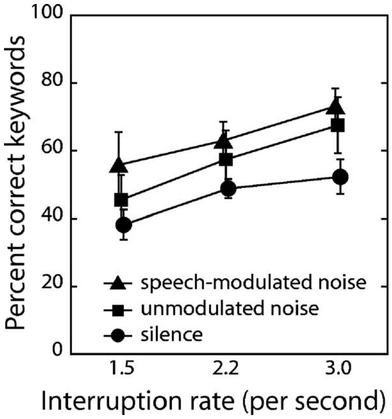

FIG. 2. Speech intelligibility is best when speech gaps are filled with speech-modulated noise and worst when gaps are silent. Average percent-correct performance on key words is plotted as a function of speech interruption rate for silent gaps, gaps filled with unmodulated noise, and gaps filled with speech-modulated noise. Error bars show the 95% confidence interval around the across-subject mean in performance.

In Experiment 2, there were two stimulus types (unmodulated and speech-modulated noise) and two spatial configurations (collocated and separated), for a total of four different combinations of stimulus and configuration. In this experiment, listeners performed four blocks of 45 trials each in each of the four sessions. In each block, listeners performed 15 trials of at each interruption rate. Across the four experimental sessions, each subject performed a total of 60 repetitions (15 trials/session × 4 sessions) for each combination of stimulus type (two—unmodulated and speech modulated noise), spatial configuration (two—collocated and separated), and repetition rate (three—1.5, 2.2, and 3.0 Hz).

F. Scoring

Each sentence contained three to five key words (adjectives, adverbs, nouns, and verbs). After each sentence was presented, listeners typed in the words they heard. The percentage of key words correctly reported was scored as a measure of speech intelligibility for each combination of stimulus type, spatial configuration, and repetition rate.

III. RESULTS

A. Experiment 1: Speech-modulated noise, unmodulated noise, or silence

Experiment 1 was designed to replicate previous findings showing that intelligibility is enhanced when speech-modulated noise fills in the speech gaps compared to silent gaps and to unmodulated noise (Bashford *et al.*, 1996). The across-subject average of the raw percent-correct key words is plotted in Fig. 2 (error bars show the 95% confidence interval around the across-subject average, in percent correct). Because individual subjects showed the same pattern as the across-subject average, only the average is shown.

Raw results verify that the interrupted speech is least intelligible when the speech gaps are silent (circles fall below other symbols in Fig. 2), most intelligible when the gaps are filled with speech-modulated noise (triangles fall above other symbols), and intermediate when the gaps are filled

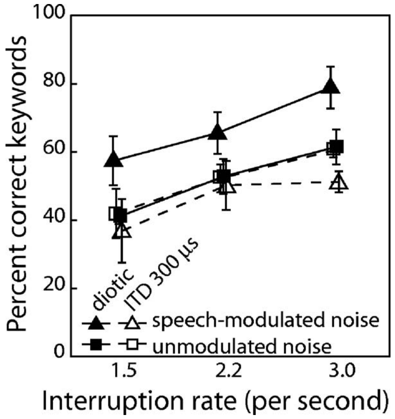

FIG. 3. Spatial configuration does not affect speech intelligibility for interrupted speech with unmodulated noise, but affects performance when the speech envelope modulates the intervening noise. Percent-correct performance on key words is plotted as a function of speech interruption rate. Speech gaps are filled with unmodulated or speech-modulated noise that are either at the same midline location as the interrupted speech or to the right of midline, with a 300 μs ITD. Error bars show the 95% confidence interval around the across-subject mean in performance.

with unmodulated noise (squares fall in between circles and triangles). As in previous studies, performance improves as the interruption rate increases (Powers and Wilcox, 1977).

A two-way ANOVA with factors of interruption rate and stimulus type supported the above-mentioned observations. Both the main effects of stimulus condition and interruption rate were significant [$F(2,36)=26.4$, $p<0.0001$ for stimulus condition; $F(2,36)=27.2$, $p<0.0001$ for interruption rate]. However, the interaction term was not statistically significant [$F(4,36)=0.673$, $p=0.62$]. Post-hoc tests with Bonferroni correction showed that compared to the silent-gap condition, performance was significantly better for both the unmodulated noise ($p<0.0001$) and the speech-modulated noise conditions ($p<0.0001$). In addition, performance with speech-modulated noise was significantly better than with unmodulated noise ($p=0.014$). Performance improved significantly with increasing interruption rate, and all three pairwise comparisons of interruption rates were significant ($p<0.0001$ for 1.5 Hz vs 2.2 Hz; $p<0.0001$ for 1.5 Hz vs 3.0 Hz; $p=0.007$ for 2.2 Hz vs 3.0 Hz).

B. Experiment 2: Effects of spatial separation for different noise types

Experiment 2 investigated the hypothesis that the across-object misallocation of amplitude modulation that produced better performance for speech-modulated noise than for unmodulated noise occurs only when the perceptual separation of the speech and noise objects is weak. We used perceived location to affect the perceptual segregation of the speech and noise by manipulating the interaural time difference (ITD) of the noise signals so that the noise was either perceived at the same midline location as the interrupted speech or to the right of midline.

Figure 3 plots mean percent correct scores (error bars represent 95% confidence interval of the across-subject mean).

Consistent with our hypothesis, perceived spatial separation between the interrupted speech and the noise did not have a noticeable impact on results for unmodulated noise: Percent-correct performance was essentially equal in both spatial configurations using unmodulated noise (compare closed and open square symbols in Fig. 3). As in Experiment 1 and previous reports (Bashford *et al.*, 1996), performance was better for speech-modulated noise than unmodulated noise when speech and noise collocated (the closed triangles are above the squares in Fig. 3). Finally, intelligibility was degraded when the interrupted speech was diotic and the speech-modulated noise was perceived to the right (the open triangles fall below the squares in Fig. 3).

Again, these conclusions were supported by statistical tests. A two-way ANOVA with factors of spatial condition and interruption rate found that both main effects were significant [$F(3,36)=42.8$, $p<0.0001$ for spatial condition; $F(2,36)=64.9$, $p<0.0001$ for interruption rate]. However, the interaction between interruption rate and condition was not statistically significant [$F(6,36)=0.777$, $p=0.593$ for the interaction]. Post-hoc tests with Bonferroni correction revealed that there was no significant difference between performance for the unmodulated noise, collocated condition and the unmodulated noise, spatially separated condition ($p>0.999$). All other conditions were significantly different at the $p=0.05$ level. Performance was significantly better in the speech-modulated noise, collocated condition than in the speech-modulated noise, spatially separated condition ($p<0.0001$); in the speech-modulated noise, collocated condition than in the unmodulated noise, collocated condition ($p<0.0001$); in the speech-modulated noise, collocated condition than in the unmodulated noise, spatially separated condition ($p<0.0001$); in the unmodulated noise, spatially separated condition than in the speech-modulated noise, spatially separated condition ($p=0.0457$); and in the unmodulated noise, collocated condition than in the speech-modulated noise, spatially separated condition ($p=0.0186$).

IV. DISCUSSION

Experiment 1 confirmed that intelligibility of interrupted speech is enhanced when the broadband envelope of the speech is used to amplitude modulate noise in the speech gaps. Although it may not initially appear surprising that adding information about the speech improves intelligibility, a more careful consideration of what is taking place is warranted. The information provided by the broadband speech envelope in the speech-modulated noise is rudimentary, providing no information about the spectral content of the missing speech. The only speech information present in the speech-modulated stimuli is crude prosodic and voicing information, which co-vary with overall speech amplitude. That such reduced information provides any improvement in speech intelligibility is a testament to how efficiently listeners use any snippet of evidence they hear in order to resolve ambiguity about the content of noisy, interrupted speech.

Moreover, it is especially surprising that these simple amplitude modulation cues aid speech intelligibility given that all listeners report that they perceive the modulation as part of the noise. In other words, the modulation is heard as an attribute of the noise, yet still contributes to the intelligibility of the collocated interrupted speech. It is likely that the noise modulation contributes to speech intelligibility because it partially matches the modulations that are expected to be present in the speech during the gaps, based on the speech glimpses the listener hears. That is, the current results suggest that knowledge of the likely spectrotemporal structure of speech causes a form of perceptual competition between the speech and noise, each of which has some evidence that it is the proper "owner" of the modulation.

Perceived location has little effect on speech intelligibility when the intervening noise simply serves as a plausible masker of the missing speech. Only when a feature of the noise provides partial information about the missing speech does the spatial relationship between the speech and noise affect intelligibility. Importantly, in our spatial manipulations, we only manipulated ITDs (e.g., we did not simulate changes in the level or spectral content of the stimulus at the ears that arise with changes in source location), which should have no direct effect on intelligibility other than strengthening the low-level auditory organization of the scene and the perceptual segregation of the competing speech and noise. Indeed, because only ITD was manipulated, there were no differences in the relative energy of the speech and noise signals in the collocated and spatially separated conditions. Nonetheless, perceived location had a large effect on intelligibility when the noise contained a feature derived from the missing speech.

In the current experiment, it appears that the perceptual grouping of the scene is ambiguous when speech-modulated noise and interrupted speech are spatially collocated because the modulation of the noise fits expectations of what should be present in the missing speech. As a result of competition for the modulation, the modulation is perceived as a feature of the noise, yet still contributes to the intelligibility of the interrupted speech. Perceived location can tip the balance for how to resolve the ambiguity about how to perceptually allocate the modulation in the noise, simply by providing additional evidence that the modulation belongs to the noise rather than the interrupted speech.

When ITDs promote hearing the speech-modulated noise and interrupted speech as separate objects, intelligibility is actually worse than in the two (collocated and spatially separated) conditions using unmodulated noise. This result likely reflects the fact that the modulations reduce the overall energy in the speech-modulated noise compared to the unmodulated noise. When the modulations do not contribute to perception of the interrupted speech because ITDs better segregate noise and speech, enhancements in intelligibility come about because the noise serves as a plausible masker of the missing speech. However, the speech-modulated noise is less effective as a possible masker of the missing speech, and thus produces less automatic filling in of the missing speech. If the modulations are perceived as coming from the same direction as the interrupted speech, they are partially attributed to the speech and enhance intelligibility, so the "plausibility" of the intervening noise as a masker of the missing speech is irrelevant. However, when interrupted speech and speech-modulated noise are spatially distinct, the modulation is perceptually allocated only to the speech-modulated noise, and the only role of the noise in speech intelligibility is to act as a plausible masker of the missing speech. Because the noise is modulated and contains temporal gaps and less overall energy, it is less effective in this role than the unmodulated noise. Thus, the fact that the intelligibility for speech-modulated noise is worse than for unmodulated noise when the interrupted speech and noises are perceived as coming from different directions further emphasizes the fact that the way a scene is organized into objects has a direct impact on the ability to understand the interrupted speech.

Many past studies suggest that the automatic filling in of missing speech caused by intervening noise only occurs when the noise is sufficiently intense to ensure that it would have masked the speech if it were continuous (e.g., see Verschuure and Brocaar, 1983). In the current study, informal reports of the subjects suggest that the speech was not perceived as continuous (although we did not test this formally). Consistent with these subjective reports, we set the noise level to have the same broadband rms as the missing speech, and used a white (not speech-shaped) spectrum. As a result, the noise is unlikely to have masked the missing speech (if it had been present) at all frequencies;[1] the speech-to-noise ratio is 0 dB, averaged across frequency, which is greater than is typically required to achieve perceived continuity. Thus, we observe perceptual filling in of the missing speech, even though the missing speech would have been audible, if it were present. It is likely the improvement in intelligibility afforded by the unmodulated noise would have been even greater with a more intense, speech-shaped noise that did produce an illusion of continuity in the speech (Verschuure and Brocaar, 1983). Nonetheless, the unmodulated noise was sufficient to encourage automatic filling in, leading to improvements in speech intelligibility over interrupted speech presented alone.

A handful of past studies have explored the degree to which perceived continuity of an interrupted signal is affected by spatial attributes of the signal and the plausible masking signal (Hartmann, 1984; Kashino and Warren, 1996; Darwin *et al.*, 2002). These studies show that perceived continuity is stronger when the signal and plausible masker have the same spatial cues rather than different spatial cues, consistent with binaural processing reducing the level of the signal that would have been masked when the interaural cues in the interrupted signal and candidate masker differ from one another. The fact that the unmodulated noise was equally effective in improving intelligibiilty when it was diotic and when it was to the side is somewhat surprising in light of these studies. Specifically, one might expect poorer speech intelligibility for the unmodulated noise with the nonzero ITD than for the diotic, unmodulated noise. However, as discussed earlier, our listeners did not perceive the speech as continuous even in the collocated, unmodulated noise condition. These results suggest that there is a less direct link between perceived continuity of an interrupted signal (which appears to be sensitive to the binaural parameters of the interrupted signal and candidate masker) and the amount of automatic filling in of the missing speech content (which appears to be less sensitive to the exact level of the candidate masker employed, at least in the current study) than some past studies suggest.

The influence of low-level auditory processes on speech perception has been a source of some debate. Some argue that the general rules governing auditory object formation do not apply to the perceptual organization of speech because there is a special, speech-specific phonetic system that is independent of the auditory system (Bentin and Mann, 1990; Whalen and Liberman, 1996; Remez, 2005). The phenomenon of "duplex perception" (Repp *et al.*, 1983), in which listeners integrate information from a frequency glide with information from other elements defining a vowel while still perceiving the glide as a distinct auditory object, is often cited as evidence supporting this kind of specialized phonetic processor (Liberman *et al.*, 1981; Repp *et al.*, 1983; Repp, 1984; Bentin and Mann, 1990; Whalen and Liberman, 1996).

In this sense, the current results resemble those of duplex perception experiments. In both paradigms, a spectrotemporal element (here, the amplitude modulation in the noise; in duplex perception, the frequency glide) could logically belong to either a speech object or a competing object, and ends up contributing perceptually to both. However, the same phenomenon is observed for an ambiguous element that logically could belong to two different nonspeech objects (e.g., Darwin and Ciocca, 1992; Bailey and Herrmann, 1993; Darwin, 1995; Hukin and Darwin, 1995; Hill and Darwin, 1996; Darwin and Hukin, 1997; 1998).

In all of these examples, ambiguous or conflicting grouping cues appear to lead to an element contributing to two different objects (e.g., when the spectrotemporal structure of a speech object supports hearing an element as part of the speech element, while other grouping cues support hearing the element as part of a separate object). These results suggest that the perceptual organization of both speech and nonspeech sounds depends on the majority of all evidence available to the listener, from low-level features (common onsets, harmonicity, comodulation, etc.) to higher-order cues such as expectations about speech structure. Any manipulation of grouping or streaming cues that alters the balance of competition for an ambiguous element can change the degree to which that element contributes to the objects in the mixture, while "sharing" of an element typically occurs only if the evidence is conflicting or ambiguous.

How an ambiguous feature or element is allocated across the objects in a sound mixture does not obey intuitively appealing rules of energy trading, wherein the total perceived content of an element is divided between competing objects (McAdams, 1989; Darwin, 1995; Shinn-Cunningham *et al.*, 2007). This seemingly paradoxical result is consistent with the idea that attention alters how an auditory scene is perceptually organized (Carlyon *et al.*, 2001; Sussman *et al.*, 2007), as if how an ambiguous scene is organized into objects depends on what object is the focus of attention (Shinn-Cunningham *et al.*, 2007). Current results are consistent with the view that perceptual organization of an ambiguous auditory scene depends on high-level factors, including listener expectations and goals.

In everyday settings, the problem of how to determine what sound energy belongs to what sound source is a significant challenge. It is often claimed that human listeners are very good at separating sound sources. Yet, often, as in the current experiments, there is a great deal of perceptual uncertainty about how to separate the sound energy in a mixture into constituent sources.

Ultimately, the goal for listeners is not segregating the sources, but understanding them. Rather than being good at estimating exactly what source produced what sound energy, listeners may simply excel at analyzing a noisy, ambiguous source and extracting its meaning, using all available evidence including a range of cues that affect source separation.

We conclude that our robust ability to understand noisy signals does not derive from an exceptional ability to perceptually separate sound sources in a mixture. Instead, our ability to understand noisy signals relies on integrating bottom-up sensory information with top-down knowledge of the likely source content (including knowledge of speech structure), taking into account all kinds of evidence that a particular sound feature belongs to a particular object. In this view, separating a source from a mixture and understanding it are intrinsically linked, rather than stages in a single, hierarchical, feed-forward process.

V. CONCLUSIONS

When interrupted speech and noise objects are imperfectly segregated, plausible modulations in the noise, derived from the missing speech, can be "borrowed" by the speech to enhance intelligibility. However, when perceptual segregation of speech and noise is strengthened through a manipulation of ITDs, this enhancement disappears. In contrast, ITD had no effect on perception when the intervening noise does not contain any speech-derived attributes.

These results show a direct interaction between auditory object formation and speech perception, at odds with claims that low-level auditory processes do not affect perception of speech. However, the current results also hint that expectations about speech content influence grouping. Together, these results suggest that perception of any complex auditory signal presented in a sound mixture depends on reciprocal, competitive interactions between low-level auditory processes and high-level knowledge about the likely spectrotemporal content of the sources making up the mixture.

ACKNOWLEDGMENTS

This work was supported by grants from the National Science Foundation. Virginia Best, Chris Darwin, John Middlebrooks, and an anonymous reviewer provided helpful comments on the manuscript.

[1] The intervening noise was white, rather than speech shaped in its spectra. As a result, and because the speech is sparse in frequency, the intervening unmodulated noise presented during the speech gaps was not optimal for eliciting perceived continuity of the speech. Specifically, within a given narrow-band frequency range that contained significant speech energy going into a speech gap, there was typically a drop of energy at the gap onset even when noise was presented in the gap. This choice was made in part because we were more interested in how intelligibility was affected by the presence of different intervening noises than in any illusory continuity of

the speech induced by the noises. During piloting, intelligibility enhancements were obtained using white noise, so we used white noise during our formal tests.

Bailey, P. J., and Herrmann, P. (**1993**). "A reexamination of duplex perception evoked by intensity differences," Percept. Psychophys. **54**, 20–32.

Bashford, J. A., Jr., Riener, K. R., and Warren, R. M. (**1992**). "Increasing the intelligibility of speech through multiple phonemic restorations," Percept. Psychophys. **51**, 211–217.

Bashford, J. A., Jr., Warren, R. M., and Brown, C. A. (**1996**). "Use of speech-modulated noise adds strong 'bottom-up' cues for phonemic restoration," Percept. Psychophys. **58**, 342–350.

Bentin, S., and Mann, V. (**1990**). "Masking and stimulus intensity effects on duplex perception: A confirmation of the dissociation between speech and nonspeech modes," J. Acoust. Soc. Am. **88**, 64–74.

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.

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

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (**2001**). "Effects of attention and unilateral neglect on auditory stream segregation," J. Exp. Psychol. Hum. Percept. Perform. **27**, 115–127.

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

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Darwin, C. J. (**1995**). "Perceiving vowels in the presence of another sound: A quantitative test of the 'Old-plus-New' heuristic," in *Levels in Speech Communication: Relations and Interactions: A Tribute to Max Wajskop*, edited by J. C. Sorin, H. Meloni, and J. Schoenigen, Elsevier, Amsterdam, the Netherlands, pp. 1–12.

Darwin, C. J. (**2005**). "Simultaneous grouping and auditory continuity," Percept. Psychophys. **67**, 1384–1390.

Darwin, C. J., Akeroyd, M. A., and Hukin, R. W. (**2002**). "Binaural factors in auditory continuity," International Conference on Auditory Displays, Kyoto, Japan.

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *Hearing*, edited by B. C. J. Moore, Academic Press, San Diego, CA, pp. 387–424.

Darwin, C. J., and Ciocca, V. (**1992**). "Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component," J. Acoust. Soc. Am. **91**, 3381–3390.

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. (**1998**). "Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony," J. Acoust. Soc. Am. **103**, 1080–1084.

Desimone, R., and Duncan, J. (**1995**). "Neural mechanisms of selective visual attention," Annu. Rev. Neurosci. **18**, 193–222.

Dyson, B. J., and Quinlan, P. T. (**2003**). "Feature and conjunction processing in the auditory modality," Percept. Psychophys. **65**, 254–272.

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.

Hall, M. D., Pastore, R. E., Acker, B. E., and Huang, W. (**2000**). "Evidence for auditory feature integration with spatially distributed items," Percept. Psychophys. **62**, 1243–1257.

Hartmann, W. M. (**1984**). "A search for central lateral inhibition," J. Acoust. Soc. Am. **75**, 528–535.

Hill, N. I., and Darwin, C. J. (**1996**). "Lateralization of a perturbed harmonic: Effects of onset asynchrony and mistuning," J. Acoust. Soc. Am. **100**, 2352–2364.

Hukin, R. W., and Darwin, C. J. (**1995**). "Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification," Percept. Psychophys. **57**, 191–196.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Kashino, M., and Warren, R. M. (**1996**). "Binaural release from temporal induction," Percept. Psychophys. **58**, 899–905.

Liberman, A. M., Isenberg, D., and Rakerd, B. (**1981**). "Duplex perception of cues for stop consonants: Evidence for a phonetic mode," Percept. Psychophys. **30**, 133–143.

McAdams, S. (**1989**). "Segregation of concurrent sounds. I. Effects of frequency modulation coherence," J. Acoust. Soc. Am. **86**, 2148–2159.

Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.

Petkov, C. I., O'Connor, K. N., and Sutter, M. L. (**2007**). "Encoding of illusory continuity in primary auditory cortex," Neuron **54**, 153–165.

Powers, G. L., and Wilcox, J. C. (**1977**). "Intelligibility of temporally interrupted speech with and without intervening noise," J. Acoust. Soc. Am. **61**, 195–199.

Remez, R. E. (**2005**). "Perceptual organization of speech," in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez, Blackwell, Oxford, UK, pp. 28–50.

Repp, B. H. (**1984**). "Against a role of 'chirp' identification in duplex perception," Percept. Psychophys. **35**, 89–93.

Repp, B. H., Milburn, C., and Ashkenas, J. (**1983**). "Duplex perception: Confirmation of fusion," Percept. Psychophys. **33**, 333–337.

Shinn-Cunningham, B. G., Lee, A. K., and Oxenham, A. J. (**2007**). "A sound element gets lost in perceptual competition," Proc. Natl. Acad. Sci. U.S.A. **104**, 12223–12227.

Sussman, E. S., Horvath, J., Winkler, I., and Orr, M. (**2007**). "The role of attention in the formation of auditory streams," Percept. Psychophys. **69**, 136–152.

Treisman, A. M., and Gelade, G. (**1980**). "A feature-integration theory of attention," Cogn. Psychol. **12**, 97–136.

Verschuure, J., and Brocaar, M. P. (**1983**). "Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise," Percept. Psychophys. **33**, 232–240.

Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.

Warren, R. M., Bashford, J. A., Jr., Healy, E. W., and Brubaker, B. S. (**1994**). "Auditory induction: Reciprocal changes in alternating sounds," Percept. Psychophys. **55**, 313–322.

Warren, R. M., Hainsworth, K. R., Brubaker, B. S., Bashford, J. A., Jr., and Healy, E. W. (**1997**). "Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps," Percept. Psychophys. **59**, 275–283.

Whalen, D. H., and Liberman, A. M. (**1996**). "Limits on phonetic integration in duplex perception," Percept. Psychophys. **58**, 857–870.