Factors influencing glimpsing of speech in noise

Ning Li and Philipos C. Loizou[a)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688

(Received 15 December 2006; revised 4 April 2007; accepted 22 May 2007)

The idea that listeners are able to "glimpse" the target speech in the presence of competing noise has been supported by many studies, and is based on the assumption that listeners are able to glimpse pieces of the target speech occurring at different times and somehow patch them together to hear out the target speech. The factors influencing glimpsing in noise are not well understood and are examined in the present study. Specifically, the effects of the frequency location, spectral width, and duration of the glimpses are examined. Stimuli were constructed using an ideal time-frequency (*T-F*) masking technique that ensures that the target is stronger than the masker in certain *T-F* regions of the mixture, thereby rendering certain regions easier to glimpse than others. Sentences were synthesized using this technique with glimpse information placed in several frequency regions while varying the glimpse window duration and total duration of glimpsing. Results indicated that the frequency location and total duration of the glimpses had a significant effect on speech recognition, with the highest performance obtained when the listeners were able to glimpse information in the $F1/F2$ frequency region (0–3 kHz) for at least 60% of the utterance. © *2007 Acoustical Society of America.* [DOI: 10.1121/1.2749454]

PACS number(s): 43.72.Dv, 43.72.Ar [DOS] Pages: 1165–1172

I. INTRODUCTION

The notion that listeners can "glimpse" the target speech when listening in noise dates back to an early study by Miller and Licklider (1950) on the intelligibility of interrupted speech masked by noise. Miller and Licklider (1950) assessed the intelligibility of interrupted speech produced by gating the speech signal on and off at a range of modulation frequencies. They found that high levels of speech understanding can be obtained when the modulation frequency (rate of interruption) was around 10 Hz even though 50% of the signal was gated off. Miller and Licklider (1950) concluded that listeners were able to piece together glimpses of the target speech available during the uninterrupted ("on" segments) portions of speech. In this study, listeners had access to the full spectrum during the uninterrupted portion. Other studies (e.g., Howard-Jones and Rosen, 1993; Buss *et al.*, 2003) used a "checkerboard" type of noise masker to investigate whether listeners were able to integrate asynchronous glimpses present in disjoint segments of the spectrum. Howard-Jones and Rosen (1993) showed that listeners were able to piece together asynchronous glimpses, provided the spectral region of the glimpses was wide enough.

Evidence of glimpsing was also reported in intelligibility studies investigating the difference in performance between identifying words in the presence of steady-state noise and in the presence of a single competing talker. Several studies (e.g., Festen and Plomp, 1990; Miller, 1947) confirmed that performance is lower in steady-state noise than in a single competing talker, and the difference in speech reception threshold (SRT) was large (6–10 dB). This difference was attributed to the fact that listeners were exploiting the silent gaps or waveform "valleys" in the competing signal to recognize the words in the target sentence. These gaps presumably enabled listeners to "glimpse" entire syllables or words of the target voice, since the local SNR is quite favorable during those gaps.

The listening-in-the-gaps account of speech segregation falls apart, however, when there are large numbers (more than 4) of competing voices present since the masker waveform becomes nearly continuous, leaving no silent gaps in the waveform (Miller, 1947). A different view of glimpsing was proposed by Cooke (2003, 2005) extending and generalizing the above idea of listening in the gaps. This new view was based on a different definition of what constitutes a *glimpse*: "a time-frequency region which contains a reasonably undistorted 'view' of local signal properties" (Cooke, 2005). Useful signal properties may include signal energy or presence of reliable $F0$ and/or formant frequency information. Glimpses of speech in background noise might, for instance, comprise of all time-frequency (*T-F*) bins or regions having a local SNR exceeding a certain threshold value (e.g., 0 dB). This definition of glimpse is henceforth adopted in the present study. The assumption is that listeners are able to first detect "useful" glimpses of speech, possibly occurring at different times and occupying different regions of the spectrum, and then somehow integrate those glimpses to hear out the target speech.

Computational models of glimpsing were developed for computational auditory scene analysis (CASA) algorithms and for robust automatic speech recognition by modifying the recognition process to allow for the possibility of "missing data" (Cooke *et al.* 1994, 2001). Despite the attractive appeal of glimpsing as a means of speech segregation in competing noise sources, there remain several issues to be resolved. Foremost among those issues is the question of what constitutes a useful glimpse and whether glimpses contain sufficient information to support identification of the tar-

[a)]Electronic mail: loizou@utdallas.edu

get signal. Several studies (Roman *et al.*, 2003; Roman and Wang, 2006; Cooke, 2006; Brungart *et al.*, 2006; Anzalone *et al.*, 2006) have attempted to answer these questions and demonstrated that speech synthesized from the ideal binary mask is highly intelligible even when extracted from multisource mixtures (Roman *et al.*, 2003) or under reverberant conditions (Roman and Wang, 2006). The ideal binary "mask" takes values of 0 and 1, and is constructed by comparing the local SNR in each *T-F* unit against a threshold (e.g., 0 dB). The ideal mask is commonly applied to the *T-F* representation of a mixture signal and eliminates portions of a signal (those assigned to a "0" value) while allowing others (those assigned to a "1" value) to pass through intact. Roman *et al.* (2003) assessed the performance of an algorithm that used location cues and an ideal time-frequency binary mask to synthesize speech. Large improvements in intelligibility were obtained from partial spectro-temporal information extracted from the ideal time-frequency mask. Similar findings were also reported by Brungart *et al.* (2006), for a range of SNR thresholds (−12 to 0 dB) used for constructing the ideal binary mask. A different method for constructing the ideal binary mask was used by Anzalone *et al.* (2006) based on comparisons of the speech energy detected in various bands against a preset threshold. The threshold value was chosen such that a fixed percentage (99%) of the total energy contained in the entire stimulus was above this threshold. Results with the ideal speech energy detector indicated significant reductions in speech reception thresholds (SRTs) for both normal-hearing and hearing-impaired listeners. Cooke (2006) used a computational model of glimpsing along with behavioral data collected from normal-hearing listeners on a consonant identification task. Several different glimpsing models were tested differing in the local SNR used for detection, the minimum glimpse size, and the use of information in the masked regions. Close fits to listener's performance on a consonant task were obtained with local SNR thresholds in the range of −2 to 8 dB.

The ideal time-frequency mask used in the above intelligibility studies for synthesizing speech makes the implicit assumption that all *T-F* units falling below a prescribed SNR threshold (e.g., 0 dB) are not detectable and should therefore be eliminated. While this assumption is valid in situations wherein there is little or no spectral overlap between the masker and the target signal in individual *T-F* units, it is not valid for speech babble or other broadband type of maskers where there exists a great deal of spectral overlap between the masker and the target. It is very likely that the masker has enough energy to distort the signal, but not to the point that it makes the target signal undetectable. Nonsimultaneous masking effects, for instance, are not taken into account when zeroing out the *T-F* units falling below the SNR threshold. Furthermore, it is known from intelligibility studies (Drullman, 1995) that the weak elements of speech lying below the noise level do contribute to some extent (up to −4 dB) to intelligibility and should therefore be preserved.

A different approach is taken in this paper to address the above limitations of using the ideal binary mask as a tool to study speech segregation or auditory scene analysis. In the proposed approach, rather than eliminating completely any *T-F* unit falling below the SNR threshold, we consider retaining those units. The *T-F* mask is no longer binary but takes real values. In the proposed approach, speech is synthesized by retaining all *T-F* units falling below the local SNR threshold while carefully controlling the duration and frequency region of the *T-F* units above the SNR threshold. The synthesized stimuli better approximate the acoustic stimuli encountered by normal-hearing listeners in a real-world noisy scenario. Under this framework, the present study aims to answer the question of what is a useful glimpse and examine the various factors that could potentially influence glimpsing in noise.

The total duration of glimpsing is one of many factors hypothesized to influence performance. In most CASA-based methods, it is assumed that glimpsing opportunities are available throughout the utterance. In practice, only a portion of the signal might be glimpsed, which in turn raises the question: What is the minimum duration of glimpsing required to achieve high levels of performance? An experiment is conducted in the present study to answer this question. In the study by Miller and Licklider (1950), 50% of the stimulus was uninterrupted and available for glimpsing, with performance steadily improving as the total duration increased. Listeners, however, had access to the full spectrum during the uninterrupted portions of speech, an assumption that generally does not hold in a complex listening situation. Only a portion of the spectrum is typically available to listeners for glimpsing in noisy environments depending on the temporal/spectral characteristics of the masker. This, in turn, raises another question: What is the influence of the location and/or width of the frequency region that is available for glimpsing? Clearly, the glimpse window width (i.e., glimpse window duration) will affect the answer to this question, and for that reason we examine systematically in experiment. 1 the influence of glimpse window width for different frequency regions of glimpsing. Previous studies showed that listeners can exploit glimpse window widths lasting as long as a phoneme for sentence/word recognition tasks (e.g., Miller and Licklider, 1950), and as short as 10 ms for a double-vowel identification task (Culling and Darwin, 1994). In most of these studies, however, listeners had either access to the full spectrum or disjoint segments of the spectrum (i.e., "checkerboard" noise) occurring periodically in time. These conditions might not reflect the true scenario in noisy environments faced by listeners wherein glimpsing opportunities may occur randomly in both time and frequency.

The findings from the present study have important implications for CASA and speech enhancement algorithms aiming to improve speech intelligibility. In many of the above studies, it is assumed that an ideal binary mask is available throughout the utterance and across the whole spectrum. In a practical system, the binary mask needs to be estimated from the noisy data, and that is a challenging task, particularly in adverse noisy conditions. Since it is practically impossible to compute accurately the ideal binary mask for all frames and all frequencies, it is of interest to determine at the very least the region in the spectrum that is perceptually most important and also the minimum duration of glimpsing required to synthesize highly intelligible

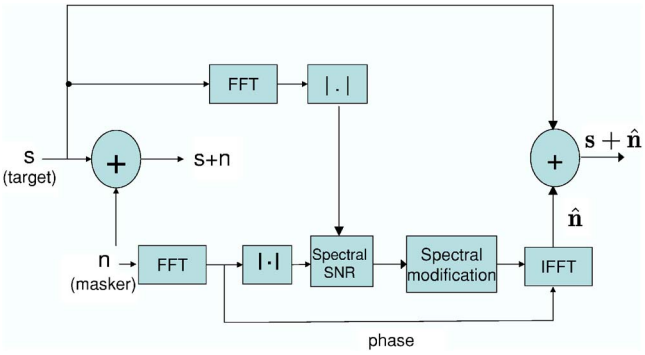

FIG. 1. (Color online) Block diagram of the signal processing technique used for constructing stimuli with glimpse information injected in prescribed frequency bands.

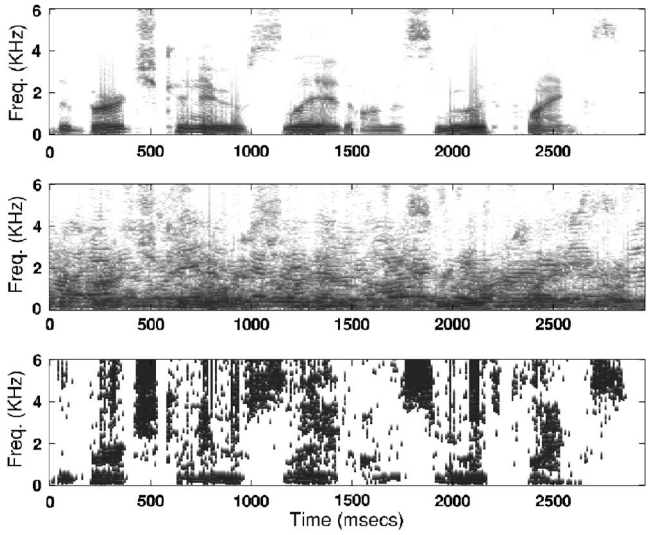

FIG. 2. Top panel shows the spectrogram of a sentence in quiet from the IEEE corpus. Middle panel shows the spectrogram of the sentence embedded in multitalker babble at −5-dB SNR. Bottom panel shows the ideal binary mask using an SNR threshold of 0- dB, with white pixels indicating a 0 (target weaker than the masker) and black pixels indicating a 1 (target stronger than the masker).

speech. These questions are addressed in the present paper.

II. EXPERIMENT 1: EFFECT OF GLIMPSE WINDOW WIDTH AND FREQUENCY LOCATION ON SPEECH INTELLIGIBILITY

A. Methods

1. Subjects

Nine normal-hearing listeners participated in this experiment. All subjects were native speakers of American English, and were paid for their participation. Subject's age ranged from 18 to 40 years, with the majority being undergraduate students from the University of Texas at Dallas.

2. Stimuli

The speech material consisted of sentences taken from the IEEE database (IEEE, 1969). All sentences were produced by a male speaker. The sentences were recorded in a sound-proof booth (Acoustic Systems) in our lab at a 25-kHz sampling rate. Details about the recording setup and copies of the recordings are available in Loizou (2007). The IEEE database consists of 72 phonetically balanced lists, each consisting of ten sentences. The sentences were corrupted by a 20-talker babble (Auditec CD, St. Louis) at −5-dB SNR. This SNR level was chosen to avoid floor effects (i.e., performance near zero).

3. Signal processing

To create stimuli with glimpses present in certain frequency regions, we spectrally modified the masker signal according to the diagram shown in Fig. 1. Our definition of glimpse is similar to that used by Cooke (2006): a time-frequency (*T-F*) region wherein the speech power is greater than the noise power by a specific threshold value (see the example in Fig. 2). In our study, we used a threshold of 0 dB, which is the threshold typically used for constructing ideal binary masks (Wang, 2005). Different SNR thresholds are considered later.

As shown in Fig. 1, the masker signal (20-talker babble) is first scaled (based on the rms energy of the target) to obtain a desired −5-dB SNR level. The target and scaled masker signals are segmented (using rectangular windows with no overlap) into 20-ms frames. A fast Fourier transform (FFT) is applied to each frame of the scaled masker to obtain the magnitude masker spectrum. Two different types of scaling are done to the masker spectrum depending on whether the *T-F* units fall within a prescribed region of the spectrum (i.e., the glimpse region) or outside the glimpse region. For all *T-F* units falling within the prescribed frequency region (glimpse region), the masker spectrum is appropriately scaled to ensure that the target *T-F* units are greater or equal (since the SNR threshold is 0 dB) in magnitude to the masker *T-F* units. The scaling (see the Appendix A for more details) is done independently in all *T-F* units in the masker spectrum which are in the glimpse region and are larger in magnitude than the corresponding target *T-F* units. No spectral modifications are done to individual *T-F* units in the masker spectrum if the target *T-F* units happen to be larger in magnitude than the masker *T-F* units. For all *T-F* units falling outside the prescribed frequency region (i.e., outside the glimpse region), the masker spectrum is appropriately scaled to ensure that the target *T-F* units are smaller in magnitude than the masker *T-F* units [note that in other studies (e.g., Brungart *et al.*, 2006), spectral components falling below the SNR threshold are set to zero]. No spectral modifications are done to individual *T-F* units in the masker spectrum if the target *T-F* units happen to be smaller in magnitude than the masker *T-F* units. The two types of scaling done to the masker spectrum ensure that only the prescribed frequency band contains glimpsing information. Following the masker magnitude modification, an inverse FFT is applied to the modified magnitude spectrum to obtain the masker signal in the time domain. The original phase spectrum of the masker is used in the reconstruction. The modified masker signal is finally added in the time domain to the clean speech signal to obtain the desired stimulus with glimpses present in a prescribed frequency band (see Appendix A for more details).

Three different frequency bands were considered: a low-

frequency (LF) band (0–1 kHz), a middle-frequency (MF) band (1–3 kHz), and a high-frequency (HF) band (>3 kHz). These bands were chosen to assess the individual contribution of formant frequencies ($F1$ and $F2$) on glimpsing in noise. The LF band contains primarily $F1$ information and the MF band contains $F2$ information. In addition to the above three bands, we also considered a low-to-mid-frequency (LF+MF) band: 0–3 kHz. This band was included as it contains both $F1$ and $F2$ information critically important for speech recognition. For comparative purposes, we also considered the following two conditions: (1) a condition spanning the full (FF) signal bandwidth, and (2) a condition, termed RF, in which the LF, MF and HF bands were randomly selected in each frame with equal probability.

To assess the effect of number of glimpses (i.e., the number of glimpse opportunities) on speech recognition, we created stimuli with different glimpse window widths (i.e., glimpse window durations). More specifically, we created stimuli with glimpse window widths of 20, 200, 400, and 800 ms spanning the duration of a phoneme to a few words. The glimpse window width is defined here as the total duration of a single glimpse spanning multiple, and neighboring in time, frames of speech. For instance, a single 200-ms glimpse is composed of ten consecutive frames (20 ms each) all containing glimpse information in a prescribed frequency band. Similarly, one 400-ms glimpse is composed of 20 consecutive frames, and one 800-ms glimpse is composed of 40 consecutive frames. The total duration of all glimpses introduced over the whole utterance was fixed to 800 ms. This number was chosen as it corresponds approximately to 33% of the total duration of most sentences in the IEEE database (average duration of sentences in the IEEE corpus was 2.4 s with a standard deviation of 0.3 s). Cooke (2005) observed that speech corrupted by eight talkers contains approximately 30% glimpses (based on a −3-dB SNR threshold). Since the signal processing involved is based on spectrally modifying the masker spectrum on a frame-by-frame basis, which is 20 ms in our experiments, we chose 20 ms as the smallest window width (duration) to be evaluated. Pilot data showed that glimpse window widths between 20 and 200 ms yielded comparable performance. Given that the total duration of all glimpses across the whole utterance was fixed at 800 ms, we created stimuli that had either 40 20-ms window glimpses, four 200-ms window glimpses, two 400-ms window glimpses, or one 800-ms window glimpse. The time location of each glimpse within the utterance was selected randomly. For comparative purposes, we also constructed stimuli in which the glimpses were present throughout the whole duration of each utterance.

In summary, we created stimuli which had low-frequency (LF) glimpse information, middle-frequency (MF) glimpse information, high-frequency (HF) glimpse information, low-to-mid frequency (LF+MF) glimpse information, randomly selected frequency (RF) information, and full-bandwidth (FF) glimpse information. For each of the above spectral regions, the glimpse window width was set to 20, 200, 400, 800 ms, and the whole utterance. To assess the potential gain in intelligibility introduced by glimpsing, we also included as a baseline condition the unmodified noisy sentences (−5-dB SNR). Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions.

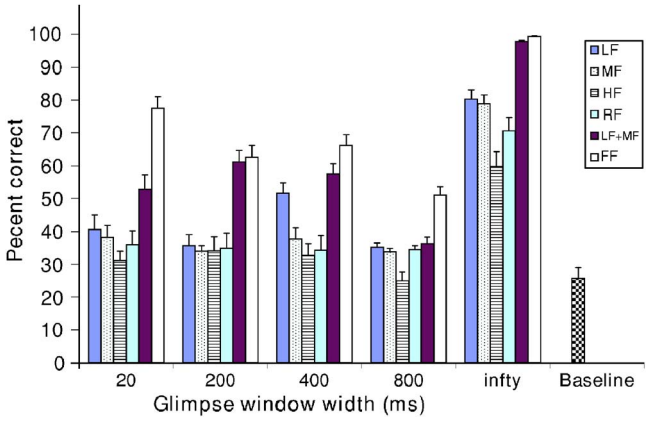

FIG. 3. (Color online) Mean subject recognition performance as a function of glimpse window width (in ms) for different frequency bands. The "infty" condition corresponds to the condition in which the indicated frequency bands were glimpsed throughout the whole utterance. The baseline condition corresponds to the unprocessed stimuli embedded in −5-dB SNR. Error bars indicate standard errors of the mean.

4. Procedure

The experiments were performed in a sound-proof room (Acoustic Systems, Inc) using a PC connected to a Tucker-Davis system 3. Stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. Prior to the test, each subject listened to a set of noisy sentences to get familiar with the testing procedure. During the test, the subjects were asked to write down the words they heard. The order of the test conditions was randomized across subjects.

B. Results and discussion

The mean scores for all conditions are shown in Fig. 3. Performance was measured in terms of percent of words identified correctly (all words were scored). The mean baseline score of the unprocessed stimuli was 25.8% correct (s.d.=9.2%). Two-way ANOVA (repeated measures) indicated a significant effect of glimpse window width ($F[4,12]=193.9$, $p<0.0005$), a significant effect of frequency band location ($F[5,15]=122.9$, $p<0.0005$), and a significant interaction ($F[20,60]=7.75$, $p<0.0005$).

Protected *posthoc* tests (Fisher's LSD) were run to examine whether there were any differences in performance between the various glimpse window widths. This analysis aims to answer the question whether it is more beneficial to have multiple, but short, glimpse opportunities or few, but long, glimpse opportunities. Separate analysis was performed for each frequency band. For the LF band, and considering only glimpse window widths from 20 to 800 ms, performance peaked at 400 ms. That is, performance at 400 ms was significantly ($p<0.05$) higher than performance at 20, 200, or 800 ms. A different pattern emerged for the other frequency bands. For the MF, HF, LF+MF, and RF bands, performance remained relatively flat across all

glimpse window widths (20–800 ms). That is, there was no statistically significant ($p>0.05$) difference in performance between the 20, 200, or 800-ms conditions. When the full bandwidth (FF) was available for glimpsing, performance peaked at 20 ms. This suggests that it is more beneficial to have multiple, but short (20 ms), glimpse opportunities rather than few, but long (400–800 ms), glimpse opportunities. This finding applies only to the full-bandwidth (FF) condition, which does not reflect the realistic scenario of listening in noise. It does, however, have important implications for speech enhancement algorithms. If an enhancement algorithm improves the spectral SNR across the whole signal bandwidth, and does so for at least 33% of the utterance duration (which is the duration used in experiment 1), then there is a good likelihood that the algorithm will significantly improve speech intelligibility. In practice, it is extremely challenging to improve the spectral SNR at all frequencies; hence, it is more practical to look for frequency bands that perform as well (or nearly as well) as when glimpsing the full signal bandwidth (more on this follows).

Next, we examined the effect of frequency band location on glimpsing in noise. We were interested in knowing whether a particular frequency band offers more benefit than others (in terms of intelligibility); hence, we ran protected *posthoc* analysis (Fisher's LSD) on the data for a fixed glimpse-window width. Results indicated the LF+MF band performed significantly ($p<0.05$) better than the other bands (LF, MF, RF) in nearly all conditions. The exception was in the 400 and 800-ms conditions wherein performance with the LF band was not statistically different ($p>0.05$) from the performance obtained with the LF+MF band. Comparison between the performance obtained with the LF+MF band and the full bandwidth (FF) condition indicated that the intelligibility scores did not differ significantly ($p>0.05$) in three of the five conditions tested. More specifically, performance with the LF+MF band in the 200-ms, 400-ms, and whole utterance glimpse conditions was the same as that obtained with the FF band (whole bandwidth), and was significantly ($p<0.05$) lower than the FF condition only in the 20 and 800-ms conditions. The finding that the LF+MF band condition performed the best and attained in nearly all cases the upper bound in performance (i.e., was as good as FF) is not surprising given that the LF+MF band contains $F1$ and $F2$ information critically important for speech recognition. The implications of this finding for speech enhancement and CASA applications is that in order to improve speech intelligibility is it extremely important to improve at the very least the spectral SNR in the region of 0–3 kHz (LF+MF band), which is the region containing $F1$ and $F2$ information.

Finally, we assessed the gain in speech intelligibility introduced by glimpsing in the various frequency bands. This gain is assessed in reference to the baseline noisy condition (−5-dB SNR). Figure 4 plots the difference in score between the scores reported in Fig. 3 and the baseline score (26.8% correct). Protected *posthoc* tests (Fisher's LSD) were run to examine whether there were any significant differences between the scores obtained with and without glimpsing (i.e., baseline score). Asterisks in Fig. 4 indicate the presence of statistically significant differences. Results indicated that introducing glimpses in the LF band produced small (about 10%–5%), but statistically significant ($p<0.05$), improvement in performance. This outcome is consistent with the findings by Anzalone *et al.* (2006), who applied, in one condition, the ideal speech energy detector only to the lower frequencies (70–500 Hz). Significant reductions in SRT were obtained by both normal-hearing and hearing-impaired listeners when the ideal speech detector was applied only to the lower frequencies (Anzalone *et al.*, 2006).

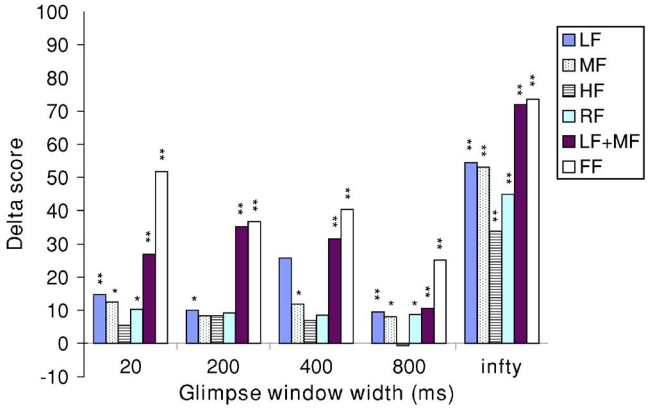

FIG. 4. (Color online) Difference in performance between that reported in Fig. 3 with glimpsed stimuli, and the baseline performance (26.8% correct). Asterisks ($*p<0.05$, $**p<0.005$) indicate statistically significant differences between the performance obtained with glimpsed stimuli and baseline stimuli. The "infty" condition corresponds to the condition in which the indicated frequency bands were glimpsed throughout the whole utterance.

Considerably larger (25%–35%), and significant ($p<0.005$), improvements were obtained in our study when glimpses were introduced in the LF+MF region. No significant ($p>0.05$) gain in intelligibility was observed when the glimpses were introduced in the HF band in any of the conditions (20–800 ms).[1] Also, no significant gain was observed when glimpses were introduced in the MF band (200 ms) or in the RF band (200, 400 ms). As one might expect, large improvements (>50%) were observed when glimpses were introduced in all frames throughout the utterance. Performance in the RF condition was consistently poor in nearly all conditions. This suggests that it is more difficult for listeners to integrate glimpses available in different frequency regions at different times, than to integrate glimpses available in the same region across time. It should be pointed out that the glimpses in the RF condition appeared randomly in time and frequency and differed in this respect to the checkerboard type of noise used in other studies (e.g., Buss *et al.*, 2003; Howard-Jones and Rosen, 1993) which appeared periodically.

The local SNR threshold used for defining the glimpses in the present experiment was fixed at 0 dB, and its value can understandably influence the outcome of the experiment. Interested to know whether a different pattern of results would be obtained with different SNR threshold values, we ran a follow-up experiment in which we varied the SNR threshold from −6 to 12 dB. Five new subjects were recruited for this experiment. The same signal-processing technique described in Sec. II A 3 (see Fig. 1) was adopted to construct stimuli with glimpses available in the LF+MF

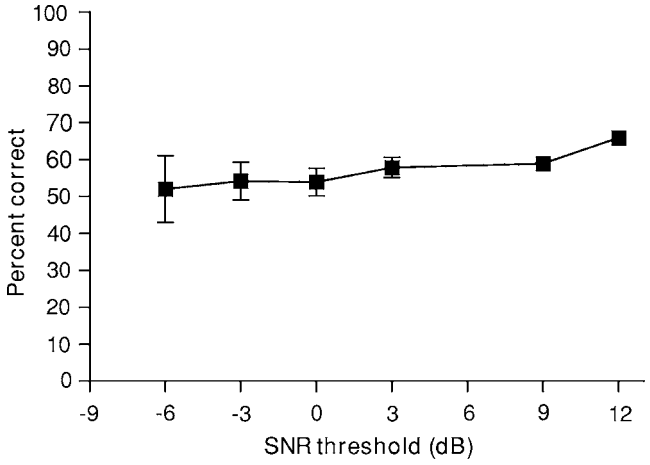

FIG. 5. Mean subject recognition performance as a function of the local SNR threshold for stimuli glimpsed in the LF+MF band. Error bars indicate standard errors of the mean.

band. This band was chosen as it performed nearly as well as the FF condition (full spectrum available). The glimpse window width was set to 20 ms. The procedure outlined in Sec. II A 4 was followed. The results, plotted in terms of percent correct, are shown in Fig. 5 as a function of the local SNR threshold. ANOVA (with repeated measures) indicated a nonsignificant ($F[5,20]=1.78$, $p=0.163$) effect of SNR threshold on performance. Performance increased slightly, but nonsignificantly, as the SNR threshold increased, and remained the same for negative values of the SNR threshold. It is worth noting that the plateau in performance seen in Fig. 5 is partially consistent with that observed by Brungart *et al.* (2006) using the ideal binary mask. The main difference between our study and that of Brungart *et al.* (2006) is that in our case performance remained flat even for positive SNR thresholds, whereas in Brungart *et al.* (2006), performance dropped precipitously for SNR thresholds above 0 dB. This difference is attributed to the fact that in Brungart *et al.* (2006) all *T-F* units falling below the SNR threshold were zeroed out; hence, the number of retained *T-F* units progressively decreased as the SNR threshold increased. In contrast, in our study all *T-F* units falling below the SNR threshold were retained [see Eq. (A5) in the Appendix A].

In summary, the results from the present experiment indicate that the glimpse window width as well as the SNR threshold had only a minor effect on performance. Glimpsing in noise was primarily affected by the location of the frequency band containing glimpses. High gains in intelligibility were obtained when glimpse information was available in the $F1$-$F2$ region (0–3 kHz).

III. EXPERIMENT 2: EFFECT OF TOTAL GLIMPSE DURATION ON SPEECH INTELLIGIBILITY

In the previous experiment, we fixed the total glimpse duration to 800 ms, corresponding roughly to 33% of the total duration for most utterances in the IEEE corpus. As shown in Fig. 3, large improvements in intelligibility were observed when the total glimpsing duration increased from 33% to 100% (compare the "infty" condition against all other conditions). This suggests that the total glimpse duration can have a significant effect on intelligibility. For that reason, we examine next the effect of total glimpse duration on performance.

A. Methods

1. Subjects and material

Nine new normal-hearing listeners participated in this experiment. All subjects were native speakers of American English, and were paid for their participation. Subjects age ranged from 18 to 40 years, with the majority being undergraduate students from the University of Texas at Dallas. The speech material consisted of sentences taken from the IEEE database (IEEE, 1969). As in experiment 1, the sentences were corrupted by a 20-talker babble masker (Auditec CD, St. Louis) at −5-dB SNR.

2. Signal processing

The method used to introduce glimpses in the time-frequency plane was the same as that used in experiment 1 (see Fig. 1). Given the relatively weak effect of glimpse window width on performance, we set the glimpse window width to 20 ms for this experiment. Unlike experiment 1, we varied the total glimpse duration to 20%, 30%, 50%, 60%, 70%, 80%, and 100% of the whole utterance. In the 50% condition, for instance, glimpses were introduced in half of the (20-ms) frames in the utterance. The time placement of the glimpses was random. Glimpses were introduced in two different bands, the LF band (0–1 kHz) and the LF+MF band (0–3 kHz). These two bands were chosen as they were found in experiment 1 to yield significant gains in intelligibility (see Fig. 4). To assess any potential gain in intelligibility introduced by glimpsing, we also included as a baseline condition the unmodified noisy sentences (−5-dB SNR). Two sentence lists were used per condition, and none of the lists were repeated.

3. Procedure

The procedure was identical to that used in experiment 1.

B. Results and discussion

The mean scores for all conditions are shown in Fig. 6. Performance was measured in terms of percent of words identified correctly. Two-way ANOVA (repeated measures) indicated a significant effect of total glimpse duration ($F[6,24]=81.5$, $p<0.0005$), a significant effect of frequency band location ($F[1,4]=269.7$, $p<0.0005$), and a significant interaction ($F[6,24]=16.54$, $p<0.0005$).

As expected, performance improved as more glimpses were introduced in both LF and LF+MF conditions. Protected *posthoc* tests (Fisher's LSD) were run to examine at which point (glimpse duration) performance reached an asymptote. Results indicated that, when the glimpses were introduced in the LF band, performance reached an asymptote at 80% of utterance duration. That is, scores obtained with 80% glimpse duration did not differ significantly ($p=0.981$) from those obtained with 100% duration (i.e., whole

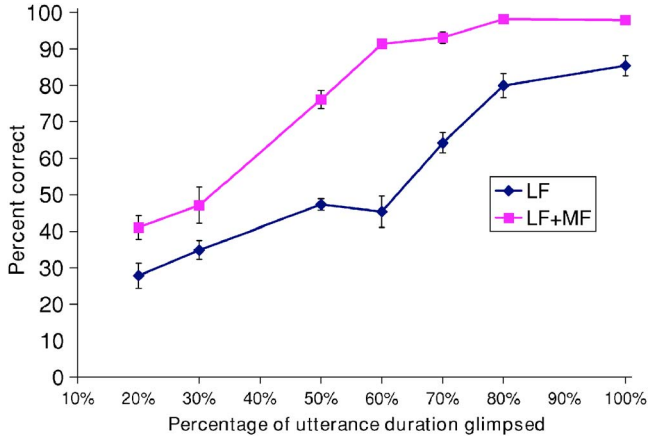

FIG. 6. (Color online) Mean subject recognition performance as a function of the percentage of the utterance glimpsed for two frequency bands. Error bars indicate standard errors of the mean.

utterance) and were significantly ($p<0.05$) higher than all other conditions ($<80\%$). In stark contrast, analysis of the LF+MF scores indicated that the asymptote occurred when glimpsing 60% of the utterance. Performance with glimpsing 100% duration (whole utterance) did not differ significantly ($p=0.093$) from that obtained with glimpsing 60% of the utterance.

The findings of experiment 2 are in close agreement with those of Miller and Licklider (1950). Near-perfect identification was achieved when only 50% of the signal was available for glimpsing during the uninterrupted portions. In their study, the listeners had access to the full clean spectrum of the target signal during the "on" segments of the signal. In our case, listeners had access to the full noisy spectrum but only the LF+MF band was above the SNR threshold and presumably available for glimpsing. For this type of stimuli containing partially masked spectral information, listeners required at least 60% of the total duration of the utterance to obtain high levels of speech understanding.

The results from the present experiment suggest that the extent of the benefit introduced by glimpsing relies heavily on both the total duration of glimpsing and the frequency band glimpsed. This suggests that, in order for CASA and enhancement algorithms to improve speech intelligibility, glimpsing in the LF+MF band needs to occur more than 50% of the time.

IV. CONCLUSIONS

A signal processing technique (Fig. 1) was proposed that can be used as a tool for studying auditory scene analysis and speech segregation in the presence of various types of maskers. Unlike the time-frequency masks used in the previous studies (e.g., Roman *et al.*, 2003; Brungart *et al.*, 2006), the proposed time-frequency mask is not binary but takes real values.

The present study primarily focused on identifying factors that may influence glimpsing speech in noise with the proposed time-frequency mask. Experiment 1 investigated the effect of glimpse window width and frequency location of the glimpse for a fixed duration (33% of utterance) of glimpsing. Experiment 2 investigated the effect of total glimpse duration for two frequency bands. From the results of these two experiments, we can draw the following conclusions:

(1) The frequency location of the glimpses had a significant effect on speech recognition, with the highest performance obtained for the LF+MF band and the lowest for the HF band. Performance with the LF+MF band was found to be as good as performance with the FF band in the majority of the conditions tested.
(2) The glimpse window width and SNR threshold had a relatively minor effect on performance (see Figs. 3 and 5), at least for the range of values considered.
(3) Relative to the unprocessed stimuli (−5-dB SNR), small (10%–15%), but statistically significant, improvements in intelligibility were obtained when the glimpses were available in the LF band, and comparatively larger (20%–30%) improvements were obtained when the glimpses were available in the LF+MF band containing $F1$ and $F2$ information.
(4) Listeners were able to integrate glimpsed information more easily when the glimpses were consistently taken from the same frequency region over time. Performance with the RF band (randomly chosen bands) was significantly lower than performance obtained with the other frequency bands.
(5) The total glimpse duration had the strongest effect in performance. High levels of speech understanding were obtained when more than 60% of the utterance duration was glimpsed in the LF+MF band, at least for the masker (multitalker babble) considered in this study. Relative to the unprocessed sentences (−5-dB SNR), this corresponds to an improvement of 64 percentage points (from 26% to 90%).

The above results have strong implications for speech enhancement and CASA algorithms aiming to improve intelligibility of speech embedded in multitalker babble. For these algorithms to improve speech intelligibility, it is extremely important to improve the spectral SNR in the region of 0–3 kHz (LF+MF band), which is the region containing $F1$ and $F2$ information. Furthermore, it is not necessary to improve the spectral SNR in all frames (i.e., whole utterance), but in at least 60% of the utterance.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH.

APPENDIX A: A TECHNIQUE FOR INTRODUCING GLIMPSES

In this appendix, we describe the signal processing technique used for modifying the masker magnitude spectra to obtain glimpses in specific regions of the spectrum.

We start by expressing the noisy speech spectrum in the frequency domain as follows:

$$Y(\tau,\omega_k) = X(\tau,\omega_k) + N(\tau,\omega_k), \tag{A1}$$

where $Y(\tau,\omega_k)$, $X(\tau,\omega_k)$, $N(\tau,\omega_k)$ are the complex FFT spectra of the noisy speech, clean speech, and masker, respectively, obtained at time (frame) τ and frequency bin ω_k (in our case, multitalker babble was added in experiment 1 to the speech signal at −5-dB SNR). The spectral SNR in time-frequency unit $\{\tau,\omega_k\}$ is given by

$$\xi(\tau,\omega_k) = 10\log_{10}\frac{|X(\tau,\omega_k)|^2}{|N(\tau,\omega_k)|^2}, \tag{A2}$$

where $|\cdot|$ indicates the magnitude spectrum. For all *T-F* units falling within the prescribed frequency region (i.e., the glimpse region), the spectral SNR $\xi(\tau,\omega_k)$ in time-frequency unit $\{\tau,\omega_k\}$ is compared against a threshold, T, and the masker magnitude spectrum is modified accordingly if $\xi(\tau,\omega_k) < T$ or left unaltered if $\xi(\tau,\omega_k) \geq T$. More precisely,

if $\xi(\tau,\omega_k) \geq T$

$Y(\tau,\omega_k) = X(\tau,\omega_k) + N(\tau,\omega_k)$

else

$Y(\tau,\omega_k) = X(\tau,\omega_k) + \hat{N}_M(\tau,\omega_k)$

end, (A3)

where $\hat{N}_M(\tau,\omega_k)$ is the modified masker spectrum, given by

$$\hat{N}_M(\tau,\omega_k) = N(\tau,\omega_k)\cdot 10^{(\xi(\tau,\omega_k)-T)/20}, \tag{A4}$$

and T is the SNR threshold given in decibels. In experiment 1, T was set to 0 dB. The operation described in Eq. (A3) is applied to all *T-F* units falling within the glimpse region. For all *T-F* units falling outside the glimpse region, the following operation is applied to ensure that the spectral SNR $\xi(\tau,\omega_k)$ of the remaining target *T-F* units is below the SNR threshold T:

if $\xi(\tau,\omega_k) < T$

$Y(\tau,\omega_k) = X(\tau,\omega_k) + N(\tau,\omega_k)$

else

$Y(\tau,\omega_k) = X(\tau,\omega_k) + \hat{N}_M(\tau,\omega_k)$

end, (A5)

where $\hat{N}_M(\tau,\omega_k)$ is given by Eq. (A4). The two types of scaling done to the masker spectrum by Eq. (A3) and Eq. (A5) ensure that only the prescribed frequency band contains glimpsing information. After applying Eq. (A3) to all *T-F* units inside the glimpse region and Eq. (A5) for all *T-F* units outside the glimpse region, we reconstruct the noisy speech in frame τ by taking inverse Fourier transform of $Y(\tau,\omega_k)$.

[1]Note that we cannot directly compare the outcome obtained in the HF condition in the present study with that obtained by Anzalone *et al.* (2006). This is because the high-frequency condition tested in the study by Anzalone *et al.* (2006) included all frequencies above 1.5 kHz, whereas in the present study the HF condition included all frequencies above 3 kHz.

Anzalone, M., Calandruccio, L., Doherty, K., and Carney, L. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**(5), 480–492.

Brungart, D., Chang, P., Simpson, B., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**(6), 4007–4018.

Buss, E., Hall, J. W., and Grose, J. H. (**2003**). "Spectral integration of synchronous and asynchronous cues to consonant identification," J. Acoust. Soc. Am. **115**, 2278–2285.

Cooke, M.P., Green, P.D., and Crawford, M.D. (**1994**). "Handling missing data in speech recognition," *Proc. 3rd Int. Conf. Spok. Lang. Proc.*, pp. 1555–1558.

Cooke, M. (**2003**). Glimpsing speech. J. Phonetics **31**, 579–584.

Cooke, M. (**2005**). "Making sense of everyday speech: A glimpsing account," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Dordrecht), pp. 305–314.

Cooke, M. P., Green, P. D., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and uncertain acoustic data," Speech Commun. **34**, 267–285.

Cooke, M.P. (**2006**). "A glimpse model of speech perception in noise," J. Acoust. Soc. Am. **119**(3), 1562–1573.

Culling, J., and Darwin, C. (**1994**). "Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating," J. Acoust. Soc. Am. **95**, 1559–1569.

Drullman, R. (**1995**). "Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796–1798.

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

Howard-Jones, P.A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**(5), 2915–2922.

IEEE. (**1969**). "IEEE Recommended Practice for Speech Quality Measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246.

Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Taylor Francis Group, Boca Raton, FL).

Miller, G. (**1947**). "The masking of speech," Psychol. Bull. **44**(2), 105–129.

Miller, G.A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**(2), 167–173.

Roman, N., Wang, D., and Brown, G. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.

Roman, N., and Wang, D. (**2006**). "Pitch-based monaural segregation of reverberant speech," J. Acoust. Soc. Am. **120**, 458–469.

Wang, D. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Dordrecht), pp. 181–187.