A glimpsing account for the benefit of simulated combined acoustic and electric hearing

Ning Li and Philipos C. Loizou^{a)}

Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688

(Received 27 July 2007; revised 29 November 2007; accepted 8 January 2008)

The benefits of combined electric and acoustic stimulation (EAS) in terms of speech recognition in noise are well established; however the underlying factors responsible for this benefit are not clear. The present study tests the hypothesis that having access to acoustic information in the low frequencies makes it easier for listeners to glimpse the target. Normal-hearing listeners were presented with vocoded speech alone (V), low-pass (LP) filtered speech alone, combined vocoded and LP speech (LP+V) and with vocoded stimuli constructed so that the low-frequency envelopes were easier to glimpse. Target speech was mixed with two types of maskers (steady-state noise and competing talker) at -5 to 5 dB signal-to-noise ratios. Results indicated no advantage of LP+V in steady noise, but a significant advantage over V in the competing talker background, an outcome consistent with the notion that it is easier for listeners to glimpse the target in fluctuating maskers. A significant improvement in performance was noted with the modified glimpsed stimuli over the original vocoded stimuli. These findings taken together suggest that a significant factor contributing to the EAS advantage is the enhanced ability to glimpse the target.

© 2008 Acoustical Society of America. [DOI: 10.1121/1.2839013]

PACS number(s): 43.66.Ts, 43.71.Ky [RYL]

Pages: 2287–2294

I. INTRODUCTION

A recent development in cochlear implants is to implant an electrode array only partially into the cochlea so as to preserve the residual acoustic hearing (20–60 dB HL up to 750 Hz and severe to profound hearing loss at 1000 Hz and above) that many patients still have at the low frequencies (von Ilberg *et al.*, 1999; Kiefer *et al.*, 2005; Gantz and Turner, 2003; Gantz *et al.*, 2006). Low frequency information is provided to these patients via a hearing aid and high-frequency (>1000 Hz) speech information is provided via a cochlear implant. Thus, these patients perceive speech via a combined electric and acoustic stimulation (EAS) mode.

The benefit of EAS in terms of better speech recognition in noise has been well documented in the literature and demonstrated by studies involving EAS patients (Kiefer *et al.*, 2005; Gantz and Turner, 2003; Turner *et al.*, 2004; Kong *et al.*, 2005; Gantz *et al.*, 2006) as well as studies involving normal-hearing listeners listening to vocoded speech (Qin and Oxenham, 2006; Dorman *et al.*, 2005; Chang *et al.*, 2006; Kong and Carlyon, 2007). EAS patients fitted with a short-electrode array were shown by Turner *et al.* (2004) to receive a 9 dB advantage in multitalker background when compared to a group of traditional patients who were matched to speech scores in quiet. Kong *et al.* (2005) demonstrated significant improvements (by an average of 8–20 percentage points) on both speech recognition (in competing talker background) and melody recognition tasks for patients with low-frequency (<1000 Hz) residual hearing in the ear contralateral to the cochlear implant. These patients were wearing a hearing aid in one ear and a cochlear implant (CI)

in the other. Several studies based on acoustics simulations of cochlear implants were conducted to probe the mechanisms and benefits obtained with simulated EAS conditions (Qin and Oxenham, 2006; Dorman *et al.*, 2005; Kong and Carlyon, 2007; Chang *et al.*, 2006). All studies reported large improvements on speech recognition, particularly in noise. Qin and Oxenham (2006) showed that including acoustic information below 600 Hz improved the speech reception threshold (SRT) by 6 dB in the presence of a competing talker and by 4 dB in the presence of speech-shaped noise compared to the vocoded-only condition. The results from both vocoder simulations and real EAS patients (Kiefer *et al.*, 2005; Kong *et al.*, 2005; Kong and Carlyon, 2007) confirmed a “superadditive” effect with combined acoustic and electric stimulation. That is, performance obtained in the EAS condition exceeded the performance obtained with either the acoustic information alone, electric information alone, or the sum of the two.

While the benefits of EAS are indisputable, the reasons for the large contribution of low-frequency acoustic information to speech intelligibility in noise are less clear. It is worth noting that only F0 and F1 information is present in the low-frequency range (<500 Hz), but the presence of F1 information alone cannot support high levels of speech recognition (Kiefer *et al.*, 2005; Kong *et al.*, 2005), at least in noise. This raises the important question: What is so special about the information in the low frequencies that enables EAS patients to better communicate in noisy environments? Some speculated (Qin and Oxenham, 2006; Turner *et al.*, 2004) that it is the improved access to F0 information that is readily available to the listeners. Voice pitch (F0) has long been thought to be a critical cue in the perceptual segregation of speech in competing-talker listening environments (e.g., Brox and Nooteboom, 1982). The pitch perception abilities

^{a)}Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu

of cochlear implant listeners are generally poor as evidenced by the poor (chance level) performance of CI users on melody recognition tasks (e.g., Gfeller *et al.*, 2002). Hence, as suggested by Kong *et al.* (2005) it is possible that EAS patients are able to somehow combine (perhaps more effectively) the salient pitch information available in the low-frequencies via acoustic stimulation with the relevantly weak pitch available in the envelope modulations via electric stimulation to enhance speech segregation between the target and the masker.

While the above-noted F0-centric hypotheses are highly plausible and reasonable, the evidence is not overly convincing. Qin and Oxenham (2005) demonstrated that normal-hearing listeners are unable to benefit from F0 differences between competing vowels in a concurrent-vowel paradigm despite the good F0 difference limens (<1 semitone) obtained with 8- and 24-channel vocoder processing. A similar outcome was noted by Stickney *et al.* (2007) with cochlear implant users listening to target and competing sentences with an F0 separation ranging from 0 to 15 semitones. Small improvements were obtained in a subsequent study by Qin and Oxenham (2006) when a five-channel vocoder was supplemented with low-frequency (<600 Hz) information which also included F1 information. It was unclear from that study however, as to whether it was access to improved F0 representation, F1 information, or access to both that contributed to the small improvement in performance with EAS. The improvement was noted primarily over the range of 0–2 semitones, where beating occurs between adjacent harmonics of the two vowels allowing listeners to use spectral cues other than F0 to identify the vowels (Culling and Darwin, 1994). In brief, the evidence from the vocoder simulation studies by Qin and Oxenham (2003, 2005) in support of improved F0 representation as the factor contributing to the EAS advantage was not strong. A recent study by Kong and Carlyon (2007) showed that the F0 information present in the low-frequency acoustic range is neither necessary nor sufficient to obtain an advantage with EAS. Their study showed that the EAS advantage persisted at low signal-to-noise ratio (SNR) levels (5 dB) even when the F0 cues were removed from the low-passed stimulus.¹ Furthermore, the combined EAS advantage disappeared at high SNR levels when F0 cues were preserved but low-frequency phonetic cues were eliminated. This outcome indicated that the low-passed acoustic stimulus contained information other than F0 that is integrated with information in the electrical stimulation to enhance speech recognition in noise. Kong and Carlyon concluded that those cues may include voicing and/or glimpsing information that EAS patients use to segregate the target. No evidence, however, was provided in that study in support of the glimpsing or voicing hypothesis.

The present study examines the hypothesis that a glimpsing mechanism is responsible for the benefit seen with EAS. More specifically, the present study considers and tests the hypothesis that having access to acoustic information in the low frequencies makes it easier for listeners to glimpse the target. We therefore expect that the EAS advantage will be diminished (or eliminated) if the masker signal is steady-state noise, since this masker lacks the waveform dips typi-

cally present in competing talker or modulated noise backgrounds. It is well established (e.g., Festen and Plomp, 1990) that listeners are able to exploit the waveform “dips” in fluctuating maskers to glimpse the target, since the SNR is more favorable during those periods. Only a few studies (e.g., Turner *et al.*, 2004) tested EAS subjects with steady-state noise, and those studies reported speech recognition performance in terms of SRT. We cannot infer, however, from a single SRT value how subjects perform as a function of SNR, since the SRT value represents a single point on the psychometric function. To further test the above-mentioned hypothesis, we modified the electric-only stimuli to allow for better glimpsing. This was accomplished using a signal-processing technique (Li and Loizou, 2007) that ensures that the target envelope amplitudes are larger than the masker envelopes in the low-frequency region, thereby rendering this region easier to glimpse. Note that we no longer restrict the definition of glimpsing to the temporal (time) domain wherein “dips” are present in the waveform but rather extend it more generally to time-frequency regions where the local SNR is favorable, i.e., the target is stronger than the masker (Cooke, 2006; Li and Loizou, 2007). If the listeners are able to glimpse the low-frequency envelope information in the modified electric-only stimuli, then we would expect to see an improvement in performance comparable to that attained with the simulated EAS stimuli.

II. EXPERIMENT: EFFECT OF LOW-FREQUENCY GLIMPING

A. Methods

1. Subjects

Seven 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 graduate students at 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). Two types of masker were used. The first was continuous (steady-state) noise, which had the same long-term spectrum as the test sentences in the IEEE corpus. The second masker was a competing-talker (female) recorded in our lab. The female talker produced a long sentence taken from the IEEE database. This was done to ensure that the target signal was always shorter (in duration) than the masker.

3. Signal processing

The stimuli were presented in four different processing conditions. The first processing condition was designed to simulate the effects of eight-channel electrical stimulation, and used an eight-channel sinewave-excited vocoder (Loizou *et al.*, 1999). Signals were first processed through a pre-

TABLE I. Filter cutoff (−3 dB) frequencies for the V and LP+V vocoder simulations.

Channel	V		LP+V	
	Low (kHz)	High (kHz)	Low (kHz)	High (kHz)
1	0.080	0.221	Unprocessed (0.080–0.600)	
2	0.221	0.426		
3	0.426	0.724		
4	0.724	1.158	0.724	1.158
5	1.158	1.790	1.158	1.790
6	1.790	2.710	1.790	2.710
7	2.710	4.050	2.710	4.050
8	4.050	6.000	4.050	6.000

emphasis filter (2000 Hz cutoff), with a 3 dB/octave rolloff, and then bandpassed into eight frequency bands between 80 and 6000 Hz using sixth-order Butterworth filters. The Cambridge filter spacing (Glasberg and Moore, 1990) was used to allocate the eight channels in the specified bandwidth. This filter spacing was identical to that used by Qin and Oxenham (2006) and is shown in Table I. The envelope of the signal was extracted by full-wave rectification and low-pass filtering (second-order Butterworth) with a 400 Hz cutoff frequency. Sinusoids were generated with amplitudes equal to the rms energy of the envelopes (computed every 4 ms) and frequencies equal to the center frequencies of the bandpass filters. The sinusoids of each band were finally summed and the level of the synthesized speech segment was adjusted to have the same rms value as the original speech segment.

The second processing condition was designed to simulate the acoustic stimulation alone. Signal was low-pass (LP) filtered to 600 Hz using a sixth-order Butterworth filter. The 600 Hz cutoff was chosen as it closely reflects the situation with EAS patients who have residual hearing up to approximately 500–750 Hz and precipitous hearing loss thereafter (Turner *et al.*, 2004; Kiefer *et al.*, 2005; von Ilberg *et al.*, 1999; Gantz *et al.*, 2006). The third processing condition was designed to simulate combined electric and acoustic stimulation. To simulate the effects of EAS with residual hearing below 600 Hz, we combined the LP stimulus with the upper five channels of the eight-channel vocoder [note that the low-pass cutoff frequency may vary in true EAS patients depending on the electrode array used and the extent (in frequency) of residual hearing in individual users].

The fourth processing condition was designed to assess the effect of low-frequency glimpsing for the electric-only stimuli. The technique is similar to that used in Li and Loizou (2007), although adapted in the present study to operate in the filter-bank (eight channels) domain rather than the Fourier transform domain. As mentioned in Sec. I, the definition of “glimpse” adopted here 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. In our study, we used a threshold of 0 dB, which is the threshold typically used for constructing ideal binary masks (Wang, 2005). The masker signal is first scaled (based on the rms energy of the target) to obtain the desired SNR level. The target and masker signals

are independently bandpass filtered as before into eight channels (same frequency spacing), and envelopes are extracted by low-pass filtering (400 Hz cutoff) the rectified waveforms. The masker envelopes in the first three channels (<600 Hz) are appropriately scaled² to ensure that the target envelopes are greater or equal (since the SNR threshold is 0 dB) than the masker envelopes. No scaling is done to the masker envelopes if the target envelopes happen to be larger in magnitude than the masker envelopes. Following the scaling of the masker envelopes, the masker envelopes are added to the target envelopes to obtain the modified mixture envelopes of channels 1–3. The mixture envelopes of the five higher frequency channels are obtained by vocoder processing the original mixture signals. Following the scaling of the low-frequency envelopes, the signal is synthesized as before as a sum of eight sine waves with amplitudes set to the envelopes and frequencies set to the center frequencies of the bandpass filters. The scaling done to the masker envelopes of channels 1–3 ensures that the low-frequency envelopes (<600 Hz) are easier to glimpse than the high-frequency envelopes. Note that the modified stimuli do not contain the original target (clean) low-frequency envelopes, but rather the envelopes which are constructed to have positive SNR (i.e., so that the target envelopes are stronger than the masker envelopes). The assumption is that listeners will be able to better glimpse low-frequency envelopes that have positive SNR.

We will be referring to the above-mentioned processing conditions as: vocoded (V), low-passed (LP), combined vocoded and low-passed acoustic (LP+V), and vocoded with low-frequency glimpsing (V+G).

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 vocoded (eight channels) speech to become familiar with the stimuli. The training session lasted for about 15–20 min. During the test, the subjects were asked to write down the words they heard. Subjects participated in a total of 24 conditions (=3 SNR levels $\times$ 4 algorithms $\times$ 2 maskers). Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions. Sentences were presented to the listeners in blocks, with 20 sentences/block in each condition. The different conditions were run in random order for each listener.

B. Results

The mean scores for all conditions are shown in Fig. 1. Performance was measured in terms of percent of words identified correctly (all words were scored). The corresponding SRT values (SNR level corresponding to 50% correct), computed by interpolating the scores in Fig. 1, were 2.2, 0.5, and 1.5 dB for the V, V+G, and LP+V stimuli, respectively, in steady-state noise. The computed SRT values in the

FIG. 1. Mean speech recognition scores as a function of SNR level for two types of background interference. The error bars denote ± 1 standard errors of the mean.

female-talker background were 3, 0, and -2.5 dB for the V, V+G, and LP+V stimuli, respectively. The SRT values for the V and LP+V stimuli are similar to those obtained by Qin and Oxenham (2006) with eight channels of stimulation and a low-pass filter of 600 Hz. The SRT values for the V stimuli, for instance, were approximately 3.5 dB for the competing-talker background and 2 dB for steady-state noise (Qin and Oxenham, 2006).

For the conditions in the steady-state noise background, two-way analysis of variance (ANOVA) (with repeated measures) indicated significant effect of SNR ($F[2,12]=426.4$, $p<0.0005$), significant effect of processing condition ($F[3,18]=72.3$, $p<0.0005$), and significant interaction ($F[6,36]=20.4$, $p<0.0005$). Similarly, for the female-talker conditions, two-way ANOVA (with repeated measures) indicated significant effect of SNR ($F[2,12]=205.6$, $p<0.0005$), significant effect of processing condition ($F[3,18]=22.8$, $p<0.0005$), and significant interaction ($F[6,36]=5.4$, $p<0.0005$).

Multiple paired-comparisons (with Bonferroni correction) were run between the scores obtained with LP+V and V stimuli in steady-state noise at the various SNR levels. The comparisons indicated no statistically significant (Bonferroni corrected $p>0.016$, $\alpha=0.05$) differences between the LP+V and V scores at any of the three SNR levels, suggesting no advantage of LP+V in steady-state noise. The scores obtained with the V+G stimuli at -5 dB SNR were significantly higher ($t(6)=3.65$, $p=0.011$) than those obtained with the LP+V stimuli but did not differ at higher SNR levels.

The pattern in performance differed in the female-talker background. Performance with the LP+V stimuli was significantly higher ($p<0.005$) than performance with the V stimuli at all SNR levels. This outcome is consistent with prior studies (Turner *et al.*, 2004; Kong *et al.*, 2005; Kong and Carlyon, 2007; Qin and Oxenham, 2006). The performance of the V+G stimuli was significantly higher ($p<0.016$) than the performance of the V stimuli at -5 and 0 dB SNR, but not at 5 dB SNR. The performance of the LP+V stimuli was significantly higher ($p<0.016$) than the performance of the V+G stimuli at -5 and 0 dB SNR, and was marginally ($p=0.016$) higher at 5 dB SNR.

The difference in performance between the V and V+G stimuli indicate that listeners are able to receive significant benefit from the glimpsed envelopes in the low frequencies. This suggests that the EAS advantage is partly due to the enhanced ability of listeners to glimpse the target. This ability is diminished when the low-frequency envelope information is provided by the vocoded stimuli (V), but can be enhanced by manipulating the low-frequency envelopes to provide better glimpsing, as done with V+G processing.

III. DISCUSSION AND CONCLUSIONS

The results of the experiments described here indicate that the LP+V stimuli did not provide significant intelligibility advantages over the V stimuli in steady-state noise. This outcome is consistent with the findings by Turner *et al.* (2004) with normal-hearing and cochlear implant listeners. No significant advantages in intelligibility were observed in steady-state noise when speech was processed via an EAS simulation and compared against a sixteen-channel vocoder (i.e., electric-only simulation). Also, the difference in performance in steady-state noise between a group of patients implanted with a traditional 20-electrode array and another group (matched with the first group in terms of consonant scores in quiet) utilizing low-frequency acoustic information (EAS) supplemented with a short-electrode array was very small and nonsignificant. In contrast, the difference in performance between the EAS and electric-only stimuli was striking when speech was presented in the background of two competing talkers. Hence, while the steady-state noise affected the two groups of patients the same way, the use of competing talkers as maskers provided significantly more benefit to the EAS patients. A similar pattern in performance was also observed in the study by Qin and Oxenham (2006) with normal-hearing listeners.

The EAS performance, and benefit, were clearly affected by the type of masker used. The fact that performance in the LP+V condition in steady-state noise background was the same as that in the V condition, but was significantly higher in a competing talker background suggests that low-frequency glimpsing played a critical role. Low-frequency glimpsing allowed the listeners in the present study to hear out the target in the LP+V stimuli corrupted by fluctuating maskers but not the target in the stimuli corrupted by steady-state noise. This is expected given that steady-state noise lacks the temporal envelope “dips” which allow listeners to glimpse the target. Listeners were probably able to glimpse

FIG. 2. (Color online) Magnitude spectra of the target and masker signals extracted from a voiced segment of an IEEE sentence. The spectra are shown prior to mixing the signals at 0 dB SNR.

the low-frequency envelope information in the V stimuli; however the glimpsed information was not effectively integrated with the high-frequency envelope information. We thus speculate that two factors played a critical role in receiving the EAS benefit when LP information is supplemented with higher-frequency vocoded information: ability to detect glimpses and ability to integrate the glimpsed information (more on this in Sec. III A).

Overall, the present study points to a glimpsing mechanism that contributes to the benefit of EAS in noise and this is discussed next along with the implications in cochlear implants.

A. Glimpsing: Suggested mechanism underlying benefit of EAS

The benefit introduced by glimpsing raises the following question: Which underlying factors or cues present in the low-frequency acoustic stimulus contribute to or facilitate glimpsing? Put differently, what is so special about the low-frequency region in the acoustic stimulus that enables EAS patients to perform better in competing talker backgrounds?

The answer to both questions and key contributing factor is the low-frequency SNR advantage. When speech is corrupted by interferers with low-pass spectral characteristics (e.g., a competing talker), the low-frequency region is masked to a lesser degree compared to the high-frequency region, at least during voiced speech segments (e.g., vowels, semivowels). The LF region is shielded to a certain extent from distortion and noise because of the low-frequency dominance of the long-term speech spectrum, (Assmann and Summerfield, 2004; Loizou, 2007, Chap. 4). The 250–500 Hz region, in particular, contains prominent speech energy with a dominant peak near 500 Hz, a characteristic of the long-term spectrum of speech that is common across 12 different languages (Byrne *et al.*, 1994). Figure 2 shows example spectra of the target and masker signals, prior to mixing them at 0 dB SNR. Note that despite the fact that the long-term rms SNR (measured across the whole sentence) of

FIG. 3. Average band SNR values of signals embedded in speech-shaped noise (top panel) and female talker (bottom panel) masker at various SNR levels. The average was computed across 20 IEEE sentences (~ 1 min) and the SNR calculation was restricted only to voiced speech segments.

the stimulus is 0 dB, the target is stronger than the masker in the low frequencies (< 500 Hz) but not in the high frequencies. Consequently, it is reasonable to expect that the SNR in the low-frequency region will be, on average, larger than the SNR in the high frequencies, thereby observing a low-frequency SNR advantage. To further assess that, we computed the average spectral SNR (in various bands) of signals embedded in female-talker and speech-shaped noise at various SNR levels. The SNR computation was restricted only to voiced speech segments,³ where the SNR advantage is to be expected. Figure 3 plots the average band SNR values obtained by filtering the stimuli in 19 bands, spaced according to the critical-band scale (Table I, ANSI, 1997), and computing the SNR in each band. It is clear that the band SNR values in the low frequencies are always larger than the SNR values in the high frequencies, at least for voiced speech segments. The difference between the SNR values in the low frequencies (< 600 Hz) and the values at higher frequencies can be as large as 10 dB (see Fig. 3). SNR analysis of the unvoiced segments revealed that the band SNR values of the unvoiced segments were substantially lower than the corresponding SNR values in voiced segments, particularly in the low frequencies (< 3 kHz). For instance, in the 0 dB SNR stimuli corrupted by the female masker, the difference was about 15–20 dB in the low frequencies (< 3 kHz) and about 1 dB in the high frequencies. Unlike the low-pass nature of the SNR distribution of the voiced segments (Fig. 3), the

SNR distribution of the unvoiced segments was found to be somewhat uniform across all frequencies, and highly variable across sentences.

The low-frequency SNR advantage (Figs. 2 and 3) is critical for several reasons. First, it provides access to a better F0 and F1 representation, and second, it provides a better glimpsing capability as the target will likely be stronger than the masker in the low-frequency region. Speech harmonics falling in the low frequencies will be affected less than the high-frequency harmonics, and listeners will thus have access to reliable F0 cues. Listeners will also have access to reliable F1 information critical for vowel and stop-consonant identification. The study by [Parikh and Loizou \(2005\)](#) demonstrated, via acoustic analysis, that F1 is preserved to a certain degree in noise even at extremely low SNR levels (−5 dB). Based on acoustic analysis of a large vowel database, F1 was identified reliably 60% of the time, whereas F2 was identified only 30% of the time when the vowels were embedded in −5 dB SNR multitalker babble. F1 information is important not only for vowel perception but also for stop-consonant perception as it conveys voicing information. The F1 onset time following the release of prevocalic stops is known, for instance, to be one of the major cues to stop voiced-unvoiced distinction (e.g., [Liberman et al., 1958](#)).

Access to a better SNR in the low-frequency region makes it easier for listeners to segregate the target in complex listening situations. Evidence of the advantage introduced by glimpsing the low-frequency region was provided in the study by [Li and Loizou \(2007\)](#). Stimuli were constructed using an ideal T – F masking technique (similar to the V+G 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. When the glimpses were introduced in the low-frequency band (0–1 kHz) for a fraction (30%) of the utterance duration, small but statistically significant improvements in performance were observed, over that attained by the unprocessed (noisy) stimuli. Considerably larger improvement (about 50 percentage points) was observed when the low-frequency glimpses were available throughout the utterance. A similar outcome was reported by [Anzalone et al. \(2006\)](#) who applied, in one condition, the ideal speech energy detector only to the lower frequencies (70–1500 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. The outcome in these studies is consistent with the benefit seen with the V+G stimuli, despite the poor spectral resolution of the stimuli used in the present study.

In summary, the LP+V stimuli contain several phonetic cues that enable listeners to better segregate the target and we contend that the underlying mechanism responsible for the EAS benefit is glimpsing. Glimpsing involves a two-stage process: detection of the target followed by integration of cues contained in the glimpses. The detection process is facilitated by a favorable SNR. Time–frequency regions with positive SNR (i.e., regions wherein the target is stronger than the masker) are easier to detect than regions with negative SNR. As argued earlier and shown in Fig. 3, it is easier to

FIG. 4. Scatter plot of low-frequency band SNR values (averaged over the 0–600 Hz range) against intelligibility scores obtained with the LP+V stimuli in the various conditions.

detect the target in the low frequencies than in the high frequencies because the low-frequency region has a more favorable SNR. This suggests a positive correlation between the effective SNR in the low frequencies and the intelligibility scores, at least for the LP+V stimuli which contain intact (nonvocalized) acoustic information in the low frequencies. Figure 4 plots the SNR values computed in the low-frequency region (averaged over the 0–600 Hz frequency range) against the intelligibility scores obtained with the LP+V stimuli in all conditions. The resulting correlation coefficient was high ($r=0.84$, $p=0.034$) consistent with our glimpsing hypothesis. A similar outcome was also found by [Cooke \(2006\)](#) when he computed the correlation between intelligibility scores of VCV consonants and the proportion of the target speech in which the local SNR exceeded 3 dB. The resulting coefficient was 0.95, albeit he considered in his study T – F regions spanning the whole spectrum rather than the low-frequency region.

The output of the detection process produces glimpses, which are somehow patched together in the second stage, namely the integration stage. The latter stage involves higher level (central) processing. Multiple cues are likely involved in the integration stage and may include F0 and F1 information, voicing cues, onset cues, and/or other auditory grouping cues ([Bregman, 1990](#)). We cannot exclude the possibility that listeners used F0 cues in the present study to segregate the target when the masker was a female talker. There is evidence, however, from another glimpsing study ([Li and Loizou, 2007](#)) that F0 cues are not always necessary depending on the task. In the study by [Li and Loizou \(2007\)](#), listeners were able to glimpse successfully the target amidst a background of 20 talkers, suggesting that cues other than F0 may be utilized in the integration process.

The favorable low-frequency SNR was present in both V and LP+V stimuli, yet the LP+V stimuli were recognized more accurately than the V stimuli. We believe that this is because the LP information facilitated better (and perhaps more effective) integration of the glimpses detected in the low (acoustic) frequency regions with the information contained in the high frequency (vocalized) regions. Alternatively, or perhaps equivalently, we can say that the glimpses in the LP stimulus provided information about the target that was missing from the V stimulus. The information (about the target) extracted from the LP stimulus was subsequently in-

tegrated with the V stimulus to yield higher speech recognition scores than those obtained with the V stimulus alone.

The V+G stimuli improved performance of the V stimuli, but did not yield the same level of performance as the LP+V stimuli. We believe that this is because the V+G processing enhanced the ability to detect the target (rendering it easier to glimpse) but the detected (vocoded) glimpses were not integrated as easily and effectively as were the (acoustic) glimpses detected from the LP stimuli. Perhaps the glimpsed envelopes provided some additional bits of information about the target that was missing from the V stimulus, but did not provide all the information available in the LP stimulus. Arguably, it is more difficult to patch together the sparse representation (three channels) of the vocoded target glimpsed in the low frequencies with the envelope information contained in the high-frequency channels. This is akin to the differences in difficulty encountered when putting together two different jigsaw puzzles comprised of either large or small pieces. It is easier to put together a jigsaw puzzle when the individual pieces are large (and smaller in number) than when the individual pieces are tiny (and larger in number). What constitutes a “useful” glimpse in terms of duration and frequency extent/location remains an open question. Modeling studies by [Cooke \(2005\)](#) showed good agreement with human listener’s performance when the glimpses were about 6 equivalent-rectangular-bandwidths (ERBs) wide in frequency.

The glimpsing account provides an explanation for the outcomes in the study by [Chang *et al.* \(2006\)](#). No EAS benefit was observed in that study when high-frequency acoustic information was introduced and combined with low-frequency vocoded information. As shown in Fig. 3, the high-frequency region has on average a nonfavorable SNR (<0 dB), suggesting that it is unlikely that the target will be glimpsed in this region, at least during voiced speech segments. As a result, the detection process will result in fewer (if any) glimpses and the integration process will be ineffective, resulting in no benefit in intelligibility.

B. Implications for cochlear implants

It is important to note that our simulations assumed “ideal” low-frequency residual hearing with normal cochlear function. Some EAS patients may indeed have good residual hearing (<30 dB HL) but others might have moderate-to-severe hearing loss (30–60 dB HL) in the 0–1000 Hz range ([Kiefer *et al.*, 2005](#)). In reality, EAS patients will have a low-frequency hearing loss and will be wearing a hearing aid. The acoustic low-pass cutoff frequency may vary (200–1000 Hz) depending on the patient and device used, and the acoustic and electric filter allocations may (or not) overlap in frequency or place. Despite the above-noted differences between the true EAS patients and the present vocoder simulations, studies ([Turner *et al.*, 2004](#); [Dorman *et al.*, 2005](#); [Kong *et al.*, 2005](#)) have shown that the outcomes of vocoder simulations have been consistent with those observed with EAS patients. With this caveat in mind, we next discuss some of the implications of our study in cochlear implants.

The obvious implication of the present study is that in order to get traditional implant users (fitted with long electrode arrays) to perform as well as EAS users in noise we need to improve the spectral resolution in the low-frequency region. As argued earlier, this will enhance the integration process and provide better glimpsing of the target. A plausible solution is to place more filters in the low frequencies at the expense of sacrificing resolution in the high frequencies ([Mckay and Henshall, 2002](#); [Fourakis *et al.*, 2004](#); [Loizou, 2006](#)). Such an approach was taken by [Fourakis *et al.* \(2004\)](#) and [Loizou \(2006\)](#) and has been found to benefit vowel recognition. The study by [Mckay and Henshall \(2002\)](#) showed that in quiet when more filters were allocated in the low frequencies, the transmission of vowel information improved but the consonant information degraded. When listening in noise, a significant benefit was noted when the low-frequency range (up to 2.6 kHz) was allocated across the nine instead of the five standard electrodes. The latter outcome is consistent with our glimpsing hypothesis, in that having access to a better (finer) spectral representation in the low frequencies will facilitate better glimpse integration. In the context of combined electric and acoustic hearing, it is not clear how many low-frequency channels would be required to achieve the same level of performance as EAS, and further experiments are warranted to investigate that.

The present study demonstrated with the use of the V+G stimuli that there exists an alternative method, other than improving spectral resolution (a challenging task), for improving the ability of listeners to glimpse low-frequency envelope information. Indeed, significant improvements in performance can be obtained with the V+G stimuli. These stimuli enabled listeners to glimpse low-frequency envelope information without having to increase the spectral resolution (note that the V+G and V stimuli had the same spectral resolution, i.e., eight channels). In fact, the performance with the V+G stimuli was significantly higher than the performance of the LP+V stimuli at extremely low SNR levels (-5 dB) in steady-state noise. For both steady-state noise and competing talker backgrounds (-5 dB SNR), performance improved by approximately 10–15 percentage points relative to the performance obtained with the V stimuli, and provided a 2–3 dB reduction in SRT. Construction of the V+G stimuli requires, however, access to the ideal binary mask or equivalently access to the true SNR in each frequency band. Several techniques do exist for estimating the ideal binary mask (e.g., [Hu and Wang, 2004](#); [Wang and Brown, 2006](#)) or the (instantaneous) spectral SNR ([Hu *et al.*, 2007](#)). One possibility is to use an algorithm to estimate the instantaneous SNR in each channel. By retaining the channel envelopes with positive SNR and discarding the channel envelopes with negative SNR, we can effectively enhance the detection of glimpses, as done with V+G processing. An algorithm for estimating the instantaneous SNR in each channel was presented in [Hu *et al.* \(2007\)](#) and is amenable to real-time implementation. Aside from devising techniques to improve glimpsing of low-frequency envelope information, different techniques can be devised to improve access to voicing information and/or to improve the F1 representation. Such techniques will likely hold promise for improving speech

recognition in noise by CI users who have no residual hearing and therefore no access to low-frequency acoustic information.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH. We would like to thank Dr. Fan-Gang Zeng and two anonymous reviewers for the valuable comments they provided.

¹It should be pointed out that in some conditions the LP stimuli in the Kong and Carlyon (2007) study were not corrupted by the masker, but the voiced stimuli were.

²The scaling of the masker envelopes of the V+G stimuli increases the global (long-term rms) SNR of the stimulus, but only by a small amount (~0.5 dB).

³In our analysis, voiced segments included not only vowels, but also all consonants with vowel-like characteristics (e.g., nasals, semivowels). The onsets following the release of voiced stop consonants (e.g., /b/, /d/) were also included in the analysis.

American National Standards Institute (1997). "Methods for calculation of the speech intelligibility index," ANSI S3.5-1997 (American National Standards Institute, New York).

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

Assmann, P., and Summerfield, Q. (2004). "The perception of speech under adverse conditions," in *Speech Processing In The Auditory System*, edited by S. Greenberg, W. Ainsworth, A. Popper, and R. Fay (Springer, New York), pp. 231–308.

Bregman, A. (1990). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Brox, J., and Nooteboom, S. (1982). "Intonation and perception of simultaneous voices," *J. Phonetics* **10**, 23–26.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M., Nasser, N., El Kholy, W., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G., Westerman, S., and Ludvigsen, C. (1994). "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.* **96**, 2108–2120.

Chang, J., Bai, J., and Zeng, F.-G. (2006). "Unintelligible low-frequency sound enhances stimulated cochlear-implant speech recognition in noise," *IEEE Trans. Biomed. Eng.* **53**, 2598–2601.

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. (2006). "A glimpse model of speech perception in noise," *J. Acoust. Soc. Am.* **119**, 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.

Dorman, M., Spahr, A., Loizou, P., Dana, C., and Schmidt, J. (2005). "Acoustic simulations of combined electric and acoustic hearing (EAS)," *Ear Hear.* **26**, 371–380.

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.

Fourakis, M., Hawks, J., Holden, L., Skinner, M., and Holden, T. (2004). "Effect of frequency boundary assignment on vowel recognition with the Nucleus 24 ACE speech coding strategy," *J. Am. Acad. Audiol.* **15**, 281–289.

Gantz, B., and Turner, C. (2003). "Combining acoustic and electric hearing," *Laryngoscope* **113**, 1726–1730.

Gantz, B. J., Turner, C., and Gfeller, K. E. (2006). "Acoustic plus electric speech processing: Preliminary results of a multicenter clinical trial of the Iowa/Nucleus Hybrid implant," *Audiol. Neuro-Otol.* **11**, 63–68.

Gfeller, K., Turner, C., Mehr, M., Woodworth, G., Fearn, R., Knutson, J., Witt, S., and Stordahl, J. (2002). "Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults," *Cochlear Implant Int.* **3**, 29–53.

Glasberg, B., and Moore, B. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.

Hu, G., and Wang, D. (2004). "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.* **15**, 1135–1150.

Hu, Y., Loizou, P., Li, N., and Kasturi, K. (2007). "Use of a sigmoidal-shaped function for noise attenuation in cochlear implants," *J. Acoust. Soc. Am.* **122**, EL128–EL134.

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

Kiefer, J., Pok, M., Adunka, O., Sturzebecher, E., Baumgartner, W., Schmidt, M., Tillein, J., Ye, Q., and Gstöettner, W. (2005). "Combined electric and acoustic stimulation of the auditory system: Results of a clinical study," *Audiol. Neuro-Otol.* **10**, 134–144.

Kong, Y., and Carlyon, R. (2007). "Improved speech recognition in noise in simulated binaurally combined acoustic and electric-stimulation," *J. Acoust. Soc. Am.* **121**, 3717–3727.

Kong, Y., Stickney, G., and Zeng, F.-G. (2005). "Speech and melody recognition in binaurally combined acoustic and electric hearing," *J. Acoust. Soc. Am.* **117**, 1351–1361.

Li, N., and Loizou, P. (2007). "Factors influencing glimpsing of speech in noise," *J. Acoust. Soc. Am.* **122**, 1165–1172.

Liberman, A., Delattre, P., and Cooper, F. (1958). "Some rules for the distinction between voiced and voiceless stops in initial position," *Lang Speech* **1**, 153–167.

Loizou, P. (2006). "Speech processing in vocoder-centric cochlear implants," in *Cochlear and Brainstem Implants*, edited by A. Møller, *Advances in OtoRhino-Laryngol.*, Vol. **64** (Basel, Karger), pp. 109–143.

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

Loizou, P., Dorman, M., and Tu, Z. (1999). "On the number of channels needed to understand speech," *J. Acoust. Soc. Am.* **106**, 2097–2103.

McKay, C., and Henshall, K. (2002). "Frequency-to-electrode allocation and speech perception with cochlear implants," *J. Acoust. Soc. Am.* **111**, 1036–1044.

Parikh, G., and Loizou, P. (2005). "The influence of noise on vowel and consonant cues," *J. Acoust. Soc. Am.* **118**, 3874–3888.

Qin, M., and Oxenham, A. (2003). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *J. Acoust. Soc. Am.* **114**, 446–454.

Qin, M., and Oxenham, A. (2005). "Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification," *Ear Hear.* **26**, 451–460.

Qin, M., and Oxenham, A. (2006). "Effects of introducing unprocessed low-frequency information on the reception of the envelope-vocoder processed speech," *J. Acoust. Soc. Am.* **119**, 2417–2426.

Stickney, G., Assmann, P., Chang, J., and Zeng, F.-G. (2007). "Effects of implant processing and fundamental frequency on the intelligibility of competing sentences," *J. Acoust. Soc. Am.* **122**, 1069–1078.

Turner, C., Gantz, B., Vidal, C., Behrens, A., and Henry, B. (2004). "Speech recognition in noise for cochlear implant listeners: Benefits of acoustic hearing," *J. Acoust. Soc. Am.* **115**, 1729–1735.

von Ilberg, C., Kiefer, C., Tillein, J., Pfennigdorff, T., Hartman, R., Sturzebecher, E., and Klinke, R. (1999). "Electric-acoustic stimulation of the auditory system," *ORL* **61**, 334–340.

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.

Wang, D., and Brown, G. (2006). *Computational Auditory Scene Analysis* (Wiley, New York).