Factors affecting masking release in cochlear-implant vocoded speech

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

(Received 27 August 2008; revised 3 February 2009; accepted 22 April 2009)

Cochlear-implant (CI) listeners generally perform better when listening to speech in steady-state noise than in fluctuating maskers, and the reasons for that are unclear. The present study presents a new hypothesis for the observed absence of release from masking. When listening to speech in fluctuating maskers (e.g., competing talkers), CI users cannot fuse the pieces of the message over temporal gaps because they are not able to perceive reliably the acoustic landmarks introduced by obstruent consonants (e.g., stops). These landmarks are evident in spectral discontinuities associated with consonant closures and releases and are posited to aid listeners determine word/syllable boundaries. To test this hypothesis, normal-hearing (NH) listeners were presented with vocoded (6–22 channels) sentences containing clean obstruent segments, but corrupted (by steady noise or fluctuating maskers) sonorant segments (e.g., vowels). Results indicated that NH listeners performed better with fluctuating maskers than with steady noise even when speech was vocoded into six channels. This outcome suggests that having access to the acoustic landmarks provided by the obstruent consonants enables listeners to integrate effectively pieces of the message glimpsed over temporal gaps into one coherent speech stream.

© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3133702]

PACS number(s): 43.66.Ts, 43.71.Ky [JCM] Pages: 338–346

I. INTRODUCTION

It is generally accepted that normal-hearing (NH) listeners are able to recognize speech in modulated or fluctuating maskers with higher accuracy than in continuous (steady-state) noise (e.g., Festen and Plomp, 1990). The benefit received when listening to speech in fluctuating maskers compared to steady maskers is often called "release of masking." This benefit can be quite substantial and can range from less than 5 to near 10 dB (e.g., Festen and Plomp, 1990; Peters *et al.*, 1998), depending on the temporal/spectral characteristics of the masker. Several factors contribute to the masking release (see review in Assmann and Summerfield, 2004) including segregation of the target on the basis of F0 differences (between the target and masker) and the ability to glimpse the target during the portions of the mixture in which the signal-to-noise ratio (SNR) is favorable, i.e., during periods in which the temporal envelope of the masker reaches a dip.

Unlike NH listeners who benefit greatly from "listening in the dips," cochlear-implant (CI) listeners are not able to receive masking release when listening to speech in fluctuating maskers. This was confirmed in studies involving CI users (Nelson *et al.*, 2003; Fu and Nogaki, 2004; Nelson and Jin, 2004; Stickney *et al.*, 2004; Cullington and Zeng, 2008) and in studies involving NH listeners listening to CI simulations, i.e., vocoded speech (Qin and Oxenham, 2003, 2005; Stickney *et al.*, 2004). Stickney *et al.* (2004) assessed speech recognition by CI users at SNR levels ranging from 0 to 20 dB using as maskers single talkers (male or female) and steady-state noise. Results showed no release from masking. In fact, performance with single talker maskers was lower than performance with steady-state noise.

The reasons for the lack of masking release are not clear, and several hypotheses have been proposed. One hypothesis suggests that CI users are not able to effectively use F0 cues to segregate the target even when a large number of channels is available (Stickney *et al.*, 2007; Qin and Oxenham, 2003, 2005). Qin and Oxenham (2005) demonstrated that NH 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 processings. A similar outcome was noted by Stickney *et al.* (2007) with CI users listening to target and competing sentences with an F0 separation ranging from 0 to 15 semitones. Others hypothesized (Nelson *et al.*, 2003) that the fluctuating maskers may cause modulation interference particularly when the signal spectral representation is poor, as is the case with current implant systems. Nelson *et al.* (2003) tested CI users with sentences embedded in modulated (gated) maskers with modulation rates varying from 1 to 32 Hz. No release of masking was observed for rates of 2–8 Hz. In fact, lower performance was observed at the syllabic rates (2–4-Hz gating) and that was attributed to the possibility that the modulated maskers were actually a distraction or interference rather than a benefit. Most CI users received benefit with the 1-Hz modulation rate, which assumes unrealistically long (500-ms) silent intervals of opportunity to glimpse the target. As argued in Nelson *et al.* (2003), the lack of masking release could not have been due to lack of audibility in the "dips" since in their study the signal level exceeded the masker level by 8 and 16 dB. Stickney *et al.* (2004) observed greater masking with

[a)]Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu single-talker than noise maskers, and they attributed that to a stronger influence of informational masking compared to energetic masking. They argued that even though the single-talker maskers are spectrally degraded, it is possible that they retain some phonetic properties of natural speech which may be easily confused with those of the target.

Overall, the outcomes from the above studies do not paint a clear, or overly convincing, picture as to why CI users do not receive release from masking. In the present paper, we investigate an alternative, and new, hypothesis that explains prior findings. As argued in most of the above studies, it is very likely that CI users are not able to integrate the pieces of the message which are glimpsed across temporal gaps to a single auditory image. We then ask the following question: Which pieces (or phonetic segments) in the noisy speech stream are difficult to perceive due to noise masking and/or perhaps CI processing? Put differently, what characteristics or features of the speech signal are more susceptible to noise? As shown by Munson and Nelson (2005) not all phonetic features/segments are affected the same way in noise. Sounds, for instance, with rapidly changing spectral patterns were found to be most vulnerable to misperception in noise by CI users (Munson and Nelson, 2005). From the NH literature we know that the obstruent consonants (stops, fricatives, and affricates) are more susceptible to noise masking than the more-intense sonorant sounds (vowels, semivowels, and nasals). Phatak and Allen (2007), for instance, showed that aside from a small subset of consonants, the vowel-to-consonant recognition ratio is well above unity for a large range of SNR levels (−20−0 dB), suggesting that vowels are easier to recognize than consonants in speech-weighted noise. The study by Parikh and Loizou (2005) showed that the information contained in the first two vowel formants is preserved to some degree even at low SNR levels. In contrast, both the spectral tilt and burst frequency of stop consonants, which are known to convey place of articulation information (e.g., Blumstein and Stevens, 1979), were significantly altered by noise.

If we accept that the obstruent consonants are heavily masked by noise, the question arises as to why better perception of the obstruent consonants (occurring roughly 33% of the time, Mines *et al.*, 1978) would help listeners identify more words in the noisy speech stream. For one, the obstruent consonants are characterized by spectral discontinuities, such as those introduced by the closure and release of stop consonants. These discontinuities manifest themselves as acoustic landmarks, which are posited to be crucial in the segmentation stage of lexical-access models (Stevens, 2002). There is evidence (see Li and Loizou, 2008) that suggests that NH listeners can receive substantial improvements in speech recognition in noise when presented with sentences containing clean obstruent consonants but noise-corrupted voiced sounds, e.g., vowels. The study by Li and Loizou (2008) focused on assessing the contribution of acoustic landmarks to the recognition of speech corrupted by steady-state maskers rather than fluctuating maskers. The present study extends the scope of the study by Li and Loizou (2008) and examines the effect of fluctuating maskers on the perception of vocoded speech. In the context of CIs, the proposed study tests the hypothesis that listeners cannot integrate the pieces of the message across temporal gaps because they cannot perceive reliably the obstruent consonants and associated acoustic landmarks. Restoring the obstruent consonants (and associated landmarks) ought to aid listeners identify more words and allow them to receive release from masking. To test this hypothesis, we present to NH listeners vocoded noisy sentences containing clean obstruent consonants but corrupted sonorant sounds. In doing so, we will assess the contribution of information carried by obstruent consonants (and associated landmarks) to masking release. Vocoded speech with varying spectral resolution and NH listeners will be used in the present paper to study masking release in the absence of confounding factors (e.g., electrode insertion depth) associated with CI users.

II. EXPERIMENT: CONTRIBUTION OF OBSTRUENT CONSONANTS TO RECOGNITION OF VOCODED SPEECH IN NOISE

A. Methods

1. Subjects

Seven NH 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 yrs, 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, Inc.) in our laboratory at a 25-kHz sampling rate. Details about the recording setup and copies of the recordings are available in Loizou (2007). Two types of maskers 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 two-talker competing speech (female) recorded in our laboratory. Two long sentences, produced by a female talker, were used from the IEEE database. This was done to ensure that the target signal was always shorter (in duration) than the masker.

The IEEE sentences were manually segmented into two broad phonetic classes: (a) the obstruent consonants which included the stops, fricatives, and affricates and (b) the sonorant sounds which included vowels, semivowels, and nasals. The segmentation was done in two steps. In the first step, a highly accurate F0 detector, taken from the STRAIGHT algorithm (Kawahara *et al.*, 1999), was used to provide the initial classification of voiced and unvoiced segments. The stop closures were classified as belonging to the unvoiced segments. The F0 detection algorithm was applied every 1 ms to the stimuli using a high-resolution fast Fourier transform to provide for accurate temporal resolution of voiced/unvoiced boundaries. Segments with non-zero F0 values are initially classified as voiced and segments with zero F0 value (as determined by the STRAIGHT algorithm) are classified as unvoiced. In the second step, the voiced and

TABLE I. Cutoff frequencies of the bandpass filters used in the vocoder simulations.

	6 channels		12 channels		22 channels	
Channel	Low (kHz)	High (kHz)	Low (kHz)	High (kHz)	Low (kHz)	High (kHz)
1	0.300	0.487	0.300	0.382	0.300	0.390
2	0.487	0.791	0.382	0.487	0.390	0.489
3	0.791	1.284	0.487	0.620	0.489	0.595
4	1.284	2.085	0.620	0.791	0.595	0.711
5	2.085	3.387	0.791	1.008	0.711	0.835
6	3.387	5.500	1.008	1.284	0.835	0.970
7			1.284	1.636	0.970	1.117
8			1.636	2.085	1.117	1.275
9			2.085	2.658	1.275	1.446
10			2.658	3.387	1.446	1.631
11			3.387	4.316	1.631	1.832
12			4.316	5.500	1.832	2.049
13					2.049	2.228
14					2.228	2.539
15					2.539	2.815
16					2.815	3.114
17					3.114	3.437
18					3.437	3.787
19					3.787	4.165
20					4.165	4.575
21					4.575	5.019
22					5.019	5.500

unvoiced decisions are inspected for errors, and the detected errors are manually corrected. Segments belonging to voiced stops with pre-voicing (e.g., /b/) as well as segments belonging to voiced fricatives (e.g., /z/) are classified as obstruent consonants. Waveform and time-aligned spectrograms were used to refine the voiced/unvoiced boundaries. Criteria for identifying a segment belonging to voiced sounds (sonorant sounds) included the presence of voicing, a clear formant pattern, and absence of signs of a vocal-tract constriction. For the special boundary that separates a prevocalic stop from a following semivowel (as in *truck*), we adopted the rule used in the phonetic segmentation of the TIMIT corpus (Seneff and Zue, 1988). More precisely, the unvoiced portion of the following semivowel or vowel was absorbed in the stop release and was thus classified as an obstruent consonant. The two-class segmentation of all IEEE sentences was saved in text files (same format as the TIMIT phn files) and is available from a CD ROM in Loizou (2007).

3. Signal processing

Signals were first processed through a pre-emphasis filter (2000-Hz cutoff), with a 3-dB/octave rolloff, and then bandpass filtered into 6, 12, or 22 channels using sixth-order Butterworth filters. Logarithmic filter spacing was used to allocate the 6 and 12 channels across a 300–5500-Hz bandwidth, and mel filter spacing was used for the 22-channel condition (see 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. The envelopes in each channel were modulated by white noise and re-filtered with the same analysis filters. The filtered waveforms 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 (clean)speech waveform of each band.

The speech stimuli were vocoded using the above algorithm in two different conditions. In the first control condition, the corrupted speech stimuli were left unaltered. That is, the obstruent consonants remained corrupted by the maskers. In the second condition, the speech stimuli contained clean obstruent segments but corrupted sonorant segments. The same level normalization factor was applied to the synthesized waveforms in both conditions. Figure 1 shows an example sentence embedded in steady noise (5-dB SNR) and processed in the two conditions. The top panel shows the spectrogram of the corrupted sentence vocoded into six channels, and the bottom panel shows the same sentence containing clean (vocoded) obstruent segments but corrupted (vocoded) sonorant segments. As shown in Fig. 1 (top panel), two of the fricative segments (at $t=0.6$ s and $t=1.9$ s) are vaguely visible in the corrupted sentence; however, the majority of the stop consonant segments are not easily discernible. The closure and release of the stop /p/, for instance (see bottom panel at $t=2.3-2.5$ s) is completely masked by noise.

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,

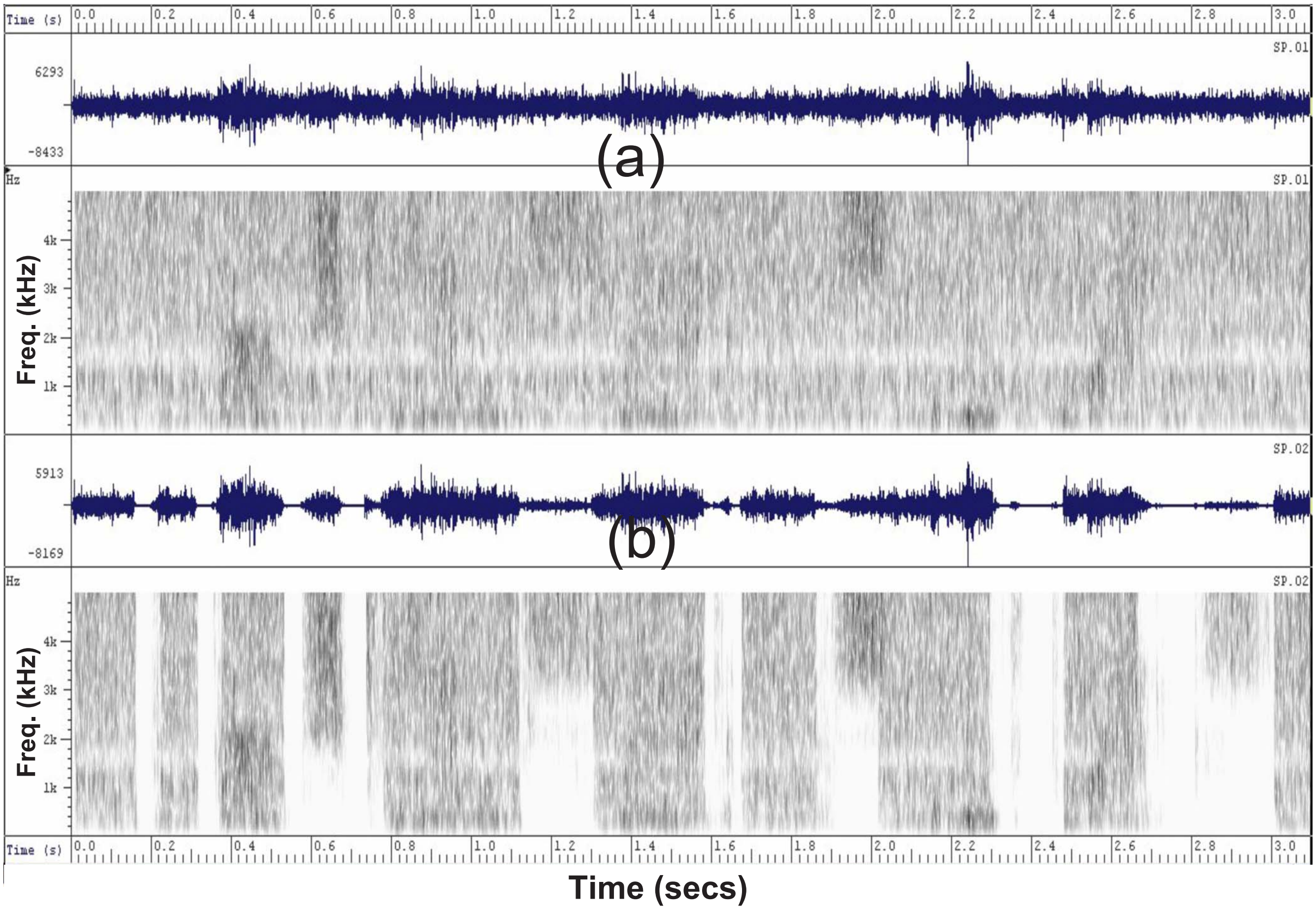

FIG. 1. (Color online) (a) Top panel shows time waveform and wide-band spectrogram of a sentence in 5-dB steady noise vocoded into six channels. (b) Bottom panel shows the same sentence containing clean (vocoded) obstruent segments but corrupted sonorant segments.

subjects listened to vocoded (6, 12, and 22 channels) sentences to become familiar with the processed stimuli. Sentences taken from the H.I.N.T. corpus (Nilsson *et al.*, 1994) were used for the training session. The training session lasted for about 20–30 min. During the test, the subjects were asked to write down the words they heard. Subjects participated in a total of 36 conditions (=3 SNR levels × 2 algorithms × 2 maskers × 3 channels). Two lists of IEEE 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 for 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. 2. Performance was measured in terms of the percentage of words identified correctly (all words were scored). Results are divided into three panels according to the number of channels used, and the individual panels are plotted as a function of SNR level. The performance obtained in quiet (denoted as Q) is also shown for comparative purposes. Two-way analysis of variance (ANOVA) with repeated measures was used to assess effects of masker type. The control noisy stimuli (shown in Fig. 2 with open symbols) processed via six channels showed no significant effect of masker type $[F(1,6)=3.8, p=0.098]$. No significant interaction $[F(2,12)=3.3, p=0.07]$ was found between SNR level and masker type. Similarly, vocoded speech processed via 12 channels showed no significant effect of masker type $[F(1,6)=0.9, p=0.379]$ and non-significant interaction $[F(2,12)=0.03, p=0.96]$ between SNR level and masker type. Finally, vocoded speech processed via 22 channels showed significant effect $[F(1,6)=40.1, p=0.001]$ of masker type and non-significant interaction $[F(2,12)=3.24, p=0.075]$ between SNR level and masker type. Performance with steady noise was significantly better than performance with the two-talker masker, consistent with findings reported in CI studies (e.g., Stickney *et al.*, 2004). The data obtained in the 6- and 12-channel conditions are partially consistent with that obtained in the studies by Stickney *et al.* (2004) and Nelson and Jin (2004). Performance obtained in the present study with 6 and 12 channels was limited to some degree by flooring effects, at least in the low SNR levels (−5 and 0 dB), and therefore failed to show a masker effect.

A different pattern in performance emerged with the vocoded stimuli in which the obstruent consonants were clean and the remaining sonorant sounds were left corrupted (shown in Fig. 2 with filled symbols). Vocoded speech processed through six channels showed a significant effect $[F(1,6)=10.9, p=0.016]$ of masker type. Performance obtained with the two-talker masker was significantly higher than performance obtained with the steady noise masker. That is, subjects benefited from the masker fluctuation. Vo-

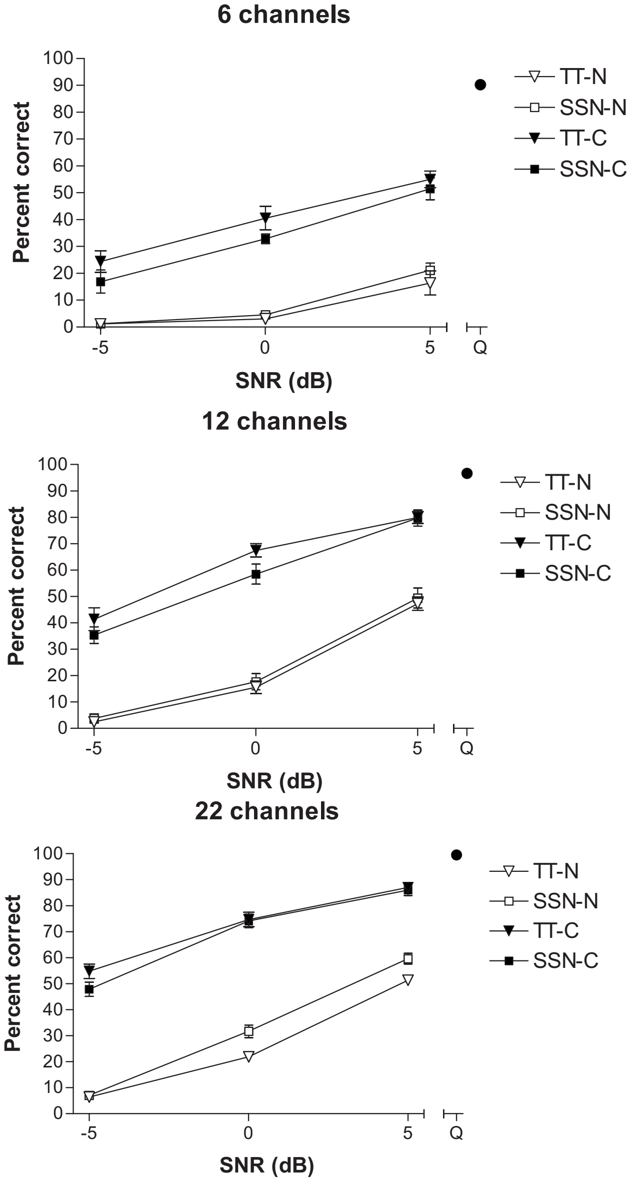

FIG. 2. Mean speech recognition scores as a function of SNR level for the various masker (TT=two-talker and SSN=steady noise) and channel conditions. Filled symbols denote scores obtained with stimuli containing clean obstruent consonants, and open symbols denote scores obtained with the control stimuli containing corrupted obstruent consonants. The performance obtained in quiet (Q) is also shown for comparative purposes. Error bars denote standard errors of the mean.

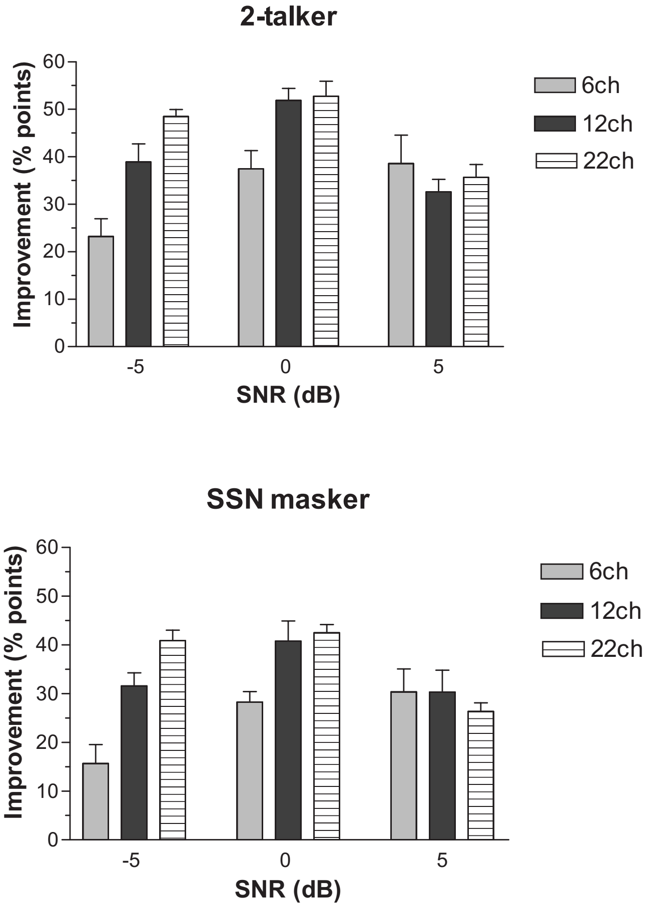

FIG. 3. Improvement in performance (in terms of overall percentage points) obtained when listeners had access to clean obstruent consonants in the various channel and SNR conditions. Error bars denote standard errors of the mean.

coded speech processed through 12 channels also showed a significant effect $[F(1,6)=14.3,\ p=0.009]$ of masker type. The interaction between SNR level and masker type was not significant $(p>0.2)$ in either 6-channel or 12-channel conditions. The 22-channel condition showed a non-significant effect $[F(1,6)=5.43,\ p=0.059]$ of masker type and a significant $[F(2,12)=4.0,\ p=0.047]$ interaction. *Post-hoc* tests indicated no significant $(p>0.05)$ difference in performance between the two masker types at 0- and 5-dB SNR, but a significant $(p=0.008)$ difference at −5-dB SNR. Performance obtained with 22 channels at −5-dB SNR with the two-talker masker was significantly higher than performance obtained with steady noise.

Introducing the clean obstruent consonants in the corrupted vocoded stimuli produced a substantial improvement in performance in all SNR and channel conditions (Fig. 2). For better clarity, Fig. 3 depicts the improvement in performance (in terms of overall percentage points) in the various channel and SNR conditions. The improvement ranged from a low of 15 percentage points (with the noise masker and with 6 channels) to a high of 50 percentage points (with the two-talker masker and with 12 or 22 channels). The estimated speech reception threshold improvement in performance (as assessed by interpolating the scores in Fig. 2) for the 12- and 22-channel conditions was near 10 dB. A steady improvement of 30–40 percentage points was observed in the 5-dB SNR conditions, independent of the number of channels used. A mild dependency on the number of channels was noted in the −5 and 0-dB SNR conditions, with 12 and 22 channels producing larger improvement than 6 channels. ANOVA analysis confirmed that the improvement in performance was highly significant in all channel and SNR conditions. Performance, for instance, with 6-channel vocoded speech containing clean obstruent consonants was significantly higher in both the two-talker masker $[F(1,6)=160.6,\ p<0.0005]$ and steady masker $[F(1,6)=444.9,\ p<0.0005]$ conditions, compared to the corresponding 6-channel control vocoded conditions with the noisy obstruent consonants. There was no significant interaction $(p>0.05)$ between SNR level and processing (clean vs. corrupted obstruent consonants). In summary, the magnitude of the improvement in performance obtained by listeners when they had access to information provided by the clean obstruent consonants seems to depend on both the SNR level and number of channels. As it will be discussed next, this dependency is probably due to the different sets of acoustic cues (and reliability of those cues) available to the listeners when presented with speech vocoded into a small number (e.g., 6) vs. a large number (e.g., 22) of channels.

III. DISCUSSION

Performance obtained by NH subjects when listening to spectrally degraded speech containing clean obstruent sounds but noisy sonorant sounds was significantly higher in conditions in which speech was corrupted by a two-talker masker than by a steady-noise masker (Fig. 2). That is, subjects benefited from release of masking when they had access to information carried by the clean obstruent consonants. We contend that the obstruent consonants carry information about the location of acoustic landmarks that are present in the signal. Knowing the location of these landmarks is crucial as it enables listeners to identify word boundaries and fuse the pieces of the underlying message across temporal gaps. The listeners were able to do this even in conditions wherein speech was degraded to six channels. The importance and contribution of acoustic landmarks to speech recognition in noise are cast in a lexical-access framework and are discussed next along with the implications of the present findings in CIs.

A. Contribution of obstruent consonants and acoustic landmarks to masking release

Many speech recognition models (e.g., Stevens, 2002; Pisoni and Sawusch, 1975; Cutler and Norris, 1988) assume that speech is first segmented at reliable points ("islands of reliability") of the acoustic signal followed by a classification of the segmented units into a sequence of phonetic units (e.g., syllables and words). The identified phonetic units are then matched against the items in the lexicon to select the best word sequence intended by the speaker. Stevens (2002) proposed a lexical-access model based on distinct acoustic landmarks partitioning vowels and consonants. The prelexical stage of this model consists of three steps. In the first step, the signal is segmented into acoustic landmarks based on detection of peaks and spectral discontinuities in the signal. These landmarks define the boundaries of the vowels, consonants, and glide segments. The second step involves extraction of acoustic cues from the vicinity of the landmarks signifying which articulators are active when the vowel, glide, or consonant landmarks are created and how these articulators are shaped or positioned. The third step consolidates, taking context into account, all the cues collected in step 2 to derive a sequence of features for each of the landmarks detected in step 1. In this speech perception model, each word in the lexicon is represented by a set of features that specify the manner of articulation, the position and shape of the articulators, as well as information about the syllable structure.

It is clear that the first step in Stevens' (2002) model (i.e., segmentation into acoustic landmarks) is crucial to the lexical-access model. If the acoustic landmarks are not detected accurately by the listeners or if the landmarks are perceptually not clear or distinct owing to corruption of the signal by external noise, this will affect the subsequent stages of the model in the following ways. First, errors will be made in step 2 in terms of identifying the shape and position of the articulators used in each landmark. Second, the absence of reliable landmarks can disrupt the syllable structure, which is known to be important for determining word boundaries in fluent speech. This is so because the onset of a word is always the onset of a syllable. Therefore, not knowing when the syllable starts makes word boundary determination very difficult (e.g., Gow *et al.*, 1996). In summary, external noise can degrade the salient cues present in syllable-initial consonants. In the context of CIs, envelope compression can also degrade these cues (see discussion in Sec. III B). These cues are present in the vicinity of the acoustic landmarks, hence identifying or somehow enhancing access to these landmarks ought to aid in identifying word boundaries and consequently improving word recognition.

When comparing the two types of maskers used in the present study, it is reasonable to expect that the two-talker masker provides more visible and perceptually more reliable acoustic landmarks than the noise masker. These landmarks include not only the ones associated with the obstruent consonants occupying the low/high frequency regions of the spectrum but also the vowel-to-glide landmarks occupying the mid-frequency region of the spectrum. Consequently, we expect higher performance with the two-talker masker than the steady noise masker. Indeed, subjects received significant release from masking (Fig. 2) when speech was processed through 6 and 12 channels and had access to acoustic landmarks present in the clean obstruent consonants. No significant release of masking was noted with 22 channels, at least for SNR ≥ 0 dB, and we believe that was because of a trading relationship between spectral resolution and importance of acoustic landmarks. When the spectral resolution is poor (e.g., six channels), then speech redundancy is greatly reduced and listeners have to rely on an alternative set of cues, such as those provided by acoustic landmarks to identify word boundaries. Hence, the importance of acoustic landmarks is greatly amplified when the spectral resolution is poor. However, when the spectral resolution is fine (e.g., 22 channels) and the SNR level is relatively high listeners can use other cues in addition to those introduced by acoustic landmarks. For one, listeners could use F1/F2 transition information which is adequately represented with 22 channels (e.g., Qin and Oxenham, 2003). In contrast, F1/F2 information is poorly represented in speech vocoded into a small number of channels (e.g., six channels) as the formant transitions might fall in the same band. The study by Munson and Nelson (2005), for instance, demonstrated that CI users have difficulty discriminating synthetic speech stimuli on the basis of formant transitions. Hence, when the spectral resolution is fine and the SNR is sufficiently high (≥ 0 dB), then the acoustic landmarks play a comparatively minor role on speech recognition as the listeners have access to other cues (e.g., F2 transitions). It is for this reason that we believe that no masking release was noted with 22 channels in most con-

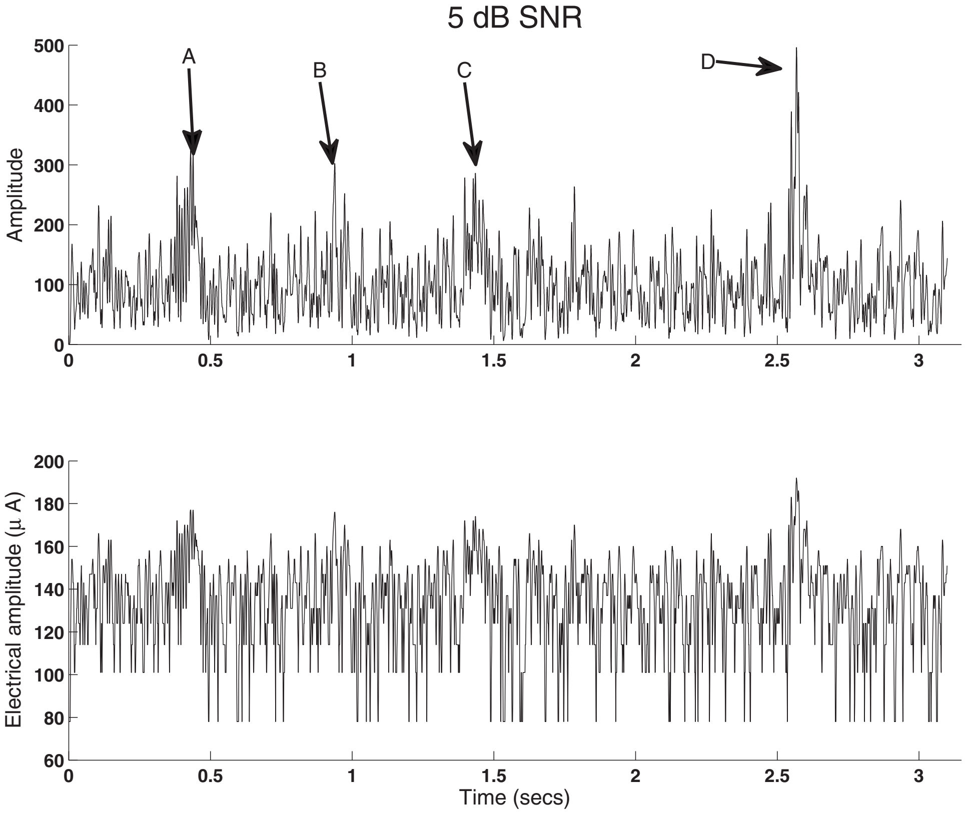

FIG. 4. Envelope (fourth channel with center frequency of 600 Hz) shown before (upper panel) and after (bottom panel) applying log-type compression to a sentence embedded in speech-shaped noise at 5-dB SNR. Arrows show some of the dominant vowel peaks present.

ditions, except in the SNR=−5-dB condition, wherein the acoustic landmarks were severely smeared by external noise. It is also for this reason that we believe that a larger improvement in performance (Fig. 3) was obtained with 12–22 channels compared to 6 channels (at least in the −5- and 0-dB SNR conditions) since the listeners had access to more cues.

The present study focused on the contribution of acoustic landmarks introduced by obstruent sounds on speech recognition in steady and fluctuating noise conditions. The sonorants (e.g., vowels and glides) also introduce landmarks in the signal (Stevens, 2002), but were not studied in this paper. We cannot exclude the possibility that the masking release observed with clean obstruent segments would have been just as large if clean sonorant segments were introduced or if equally long clean segments were placed randomly across the sentence. Introducing clean sonorant segments, however, does not reflect realistic noisy conditions since the acoustic noise does not degrade the sonorant segments to the same degree as the obstruent segments. As mentioned in the Introduction, we restricted our attention to the landmarks introduced by obstruent sounds for the main reason that these sounds are more susceptible to noise (i.e., easily masked) than the sonorant sounds (Phatak and Allen, 2007; Parikh and Loizou, 2005).

B. Implications for CIs

As mentioned in the Introduction, CI users do not receive release of masking when listening to speech in modulated interference (e.g., Nelson *et al.*, 2003; Fu and Nogaki, 2004; Stickney *et al.*, 2004; Cullington and Zeng, 2008). Based on the findings from the present study, we believe that two interrelated factors contribute to that. The first factor is envelope compression and reduced dynamic range. In current CI systems, the envelopes extracted from each band are compressed with a logarithmic function in order to map the wide acoustic dynamic range to the small (5−15-dB) electrical dynamic range. This envelope compression smears the acoustic landmarks a great deal (more so in noise) making it extremely difficult for CI users to identify word boundaries. Figure 4 shows an example envelope (fourth channel) obtained before and after applying log-type compression to a sentence embedded in speech-shaped noise at 5-dB SNR. Figure 5 shows the same sentence in quiet. Arrows shown and labeled as A–D (Fig. 4) point to some of the vowel landmarks and arrows labeled as E–G (Fig. 5) point to some of the obstruent landmarks. It is clear from Fig. 4 (bottom panel) that the obstruent landmarks are not easily discernible and probably not easily perceptible. As an indirect measure of assessing the presence of obstruent landmarks, one can compute the peak-to-trough ratio for the peaks labeled as A–D in Fig. 4. For instance, for peak A and trough E (Fig. 5), the peak-to-trough ratio for the linearly processed (i.e., no compression) envelope is 10 dB, where the trough level is set to the mean noise floor level of that channel. The corresponding peak-to-trough ratio for the compressed envelope is 2.4 dB. As shown in a previous study in our laboratory (Loizou and Poroy, 2001), it is unlikely that patients will be able to perceive such a small (∼2-dB) peak-to-trough ratio and consequently perceive the acoustic landmarks introduced by the obstruent consonants.

The second factor, which is a direct consequence of the first, is poor access to the location of the acoustic landmarks needed to determine word or syllable boundaries. Poor spectral resolution further exacerbates the situation as it reduces

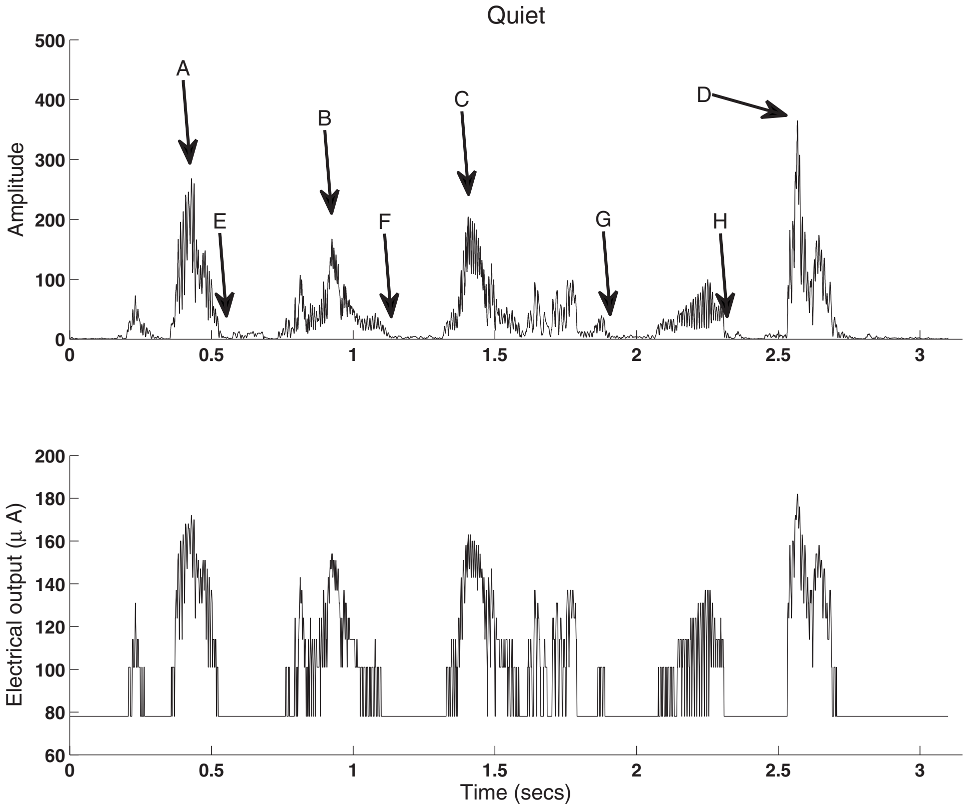

FIG. 5. Envelope (fourth channel with center frequency of 600 Hz) shown before (upper panel) and after (bottom panel) applying log-type compression to a sentence (same as in Fig. 4) in quiet. Arrows labeled E–G show some of the obstruent landmarks present, and arrows labeled A–D show some of the vowel landmarks.

speech redundancy and forces listeners to rely more on information carried by acoustic landmarks to identify word or syllable boundaries. Without good and accurate knowledge of the location of the acoustic landmarks, it becomes extremely difficult for users to first identify the pieces (based perhaps on their delineating boundaries) of the underlying message and then integrate those pieces together. Hence, the second factor is quite detrimental as it limits the CI user's ability to integrate information across temporal gaps into one coherent speech stream.

In terms of (indirectly) addressing the second factor, it is extremely challenging to improve spectral resolution, at least with existing technology and knowledge. Increasing the number of electrode contacts, for instance, would not necessarily increase the number of independent channels of information (e.g., Friesen *et al.*, 2001). Fortunately, there are ways to address the first factor. One can use a dynamically changing compression function that is less compressive during the obstruent consonant segments and more compressive during the sonorant sound segments (current CIs use the same shape compression function for all phonetic segments). Such a strategy would require the use of automatic landmark detection algorithms that would identify the locations of acoustic landmarks in the signal. Fortunately, several such algorithms exist in the literature and have been found to work quite well, at least in quiet (e.g., Liu, 1996; Junega and Espy-Wilson, 2008). An alternative approach was proposed by Kasturi and Loizou (2007) based on the use of s-shaped input-output functions which are expansive for low input levels, up to a knee point level, and compressive thereafter. The knee points of the s-shaped functions changed dynamically and were set proportional to the estimated noise floor level. For the most part, the expansive (i.e., less compressive) part of the s-shaped functions operated on obstruent segments, which generally have lower intensity compared to that of sonorant segments. The main advantage of using s-shaped functions to map the acoustic signal to electrical output values is that these functions do not require landmark detection algorithms as they are applied to all phonetic segments. Replacing the conventional log mapping functions with the s-shaped functions yielded significant improvements in speech intelligibility in noise by nine CI users (Kasturi and Loizou, 2007). In summary, one research direction that warrants further investigation is the development of dynamically changing compression functions that would maintain or enhance the acoustic landmarks present in the signal. In doing so, and according to the present data (Fig. 2) as well as the data from Kasturi and Loizou (2007), it is reasonable to expect that CI users would obtain large gains in intelligibility in noisy backgrounds and receive release of masking.

Finally, the data from Fig. 2 suggest that the obstruent consonants contribute significantly to speech recognition in noise. Large improvements in performance were observed when listeners had access to the clean obstruent spectra even when the spectral resolution was poor (see Figs. 2 and 3). The improvement was quite substantial and amounted to 30–50 percentage points for speech vocoded in 12–22 channels and to 15–30 percentage points for speech vocoded in 6 channels (Fig. 3). In the context of CI processing, the data in Figs. 2 and 3 also suggest that the obstruent consonants need to be processed, or at least treated, differently than the sonorant sounds. One possibility is to apply, as mentioned above, different shape compression functions to the obstruent segments, and another possibility is to enhance and/or clean selectively the noisy obstruent spectra. Such techniques have the potential of preserving the acoustic landmarks in the signal and consequently improving speech recognition in noise by CI users.

IV. CONCLUSIONS

It is well established that CI users do not receive release of masking when listening to speech in fluctuating maskers (e.g., Nelson *et al.*, 2003; Stickney *et al.*, 2004). The present study tested a new hypothesis for the observed absence of release of masking using vocoded speech and NH listeners as subjects. The proposed hypothesis is that the CI user's ability to fuse information across temporal gaps is limited by their ability to perceive acoustic landmarks such as those introduced by obstruent consonants. These landmarks play an important role in models of lexical access (Stevens, 2002) and are needed to identify word/syllable boundaries. Results indicated that when listeners were presented with vocoded speech (6–22 channels) with corrupted sonorant sounds (e.g., vowels) but clean obstruent sounds (e.g., stops), they were able to receive release of masking, even with 6 channels of stimulation. That is, listeners performed better in fluctuating maskers than in steady-noise maskers. Dramatic improvements in performance were observed when listeners had access to the clean obstruent spectra even when the spectral resolution was poor. The improvement amounted to 30–50 percentage points for speech vocoded in 12–22 channels and to 15–30 percentage points for speech vocoded in 6 channels. The present data suggest that the obstruent consonants need to be treated differently in noisy conditions so as to preserve the acoustic landmarks present in the signal. One possibility suggested is to apply a different shape compression function (e.g., less compressive) during the obstruent segments since the log-compressive function tends to smear and, in some cases, abolish critical obstruent landmarks (see example in Fig. 4).

ACKNOWLEDGMENTS

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

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.

Blumstein, S., and Stevens, K. (**1979**). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017

Cullington, H., and Zeng, F.-G. (**2008**). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant and implant simulation subjects," J. Acoust. Soc. Am. **123**, 450–461.

Cutler, A., and Norris, D. (**1988**). "The role of strong syllables in segmentation for lexical access," J. Exp. Psychol. Hum. Percept. Perform. **14**, 113–121.

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.

Friesen, L. M., Shannon, R. Y., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Fu, Q., and Nogaki, G. (**2004**). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," J. Assoc. Res. Otolaryngol. **6**, 19–27.

Gow, D. W., Melvold, J., and Manuel, S. (**1996**). "How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology and processing," Proc. Int. Conf. Spoken Lang. Proc., Philadelphia, PA, pp. 66–69.

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

Junega, A., and Espy-Wilson, C. (**2008**). "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," J. Acoust. Soc. Am. **123**, 1154–1168.

Kasturi, K., and Loizou, P. (**2007**). "Use of s-shaped input-output functions for noise suppression in cochlear implants," Ear Hear. **28**, 402–411.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.

Li, N., and Loizou, P. (**2008**). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," J. Acoust. Soc. Am. **124**, 3947–3958.

Liu, S. (**1996**). "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am. **100**, 3417–3430.

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

Loizou, P., and Poroy, O. (**2001**). "Minimum spectral contrast needed for vowel identification by normal-hearing and cochlear implant listeners," J. Acoust. Soc. Am. **110**, 1619–1627.

Mines, M., Hanson, B., and Shoup, J. (**1978**). "Frequency of occurrence of phonemes in conversational English," Lang Speech **21**, 221–241.

Munson, B., and Nelson, P. (**2005**). "Phonetic identification in quiet and in noise by listeners with cochlear implants," J. Acoust. Soc. Am. **118**, 2607–2617.

Nelson, P., and Jin, S. (**2004**). "Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **115**, 2286–2294.

Nelson, P., Jin, S., Carney, A., and Nelson, D. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 961–968.

Nilsson, M., Soli, S., and Sullivan, J. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

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

Peters, R., Moore, B., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.

Phatak, S., and Allen, J. (**2007**). "Consonants and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.

Pisoni, D., and Sawusch, J. (**1975**). "Some stages of processing in speech perception," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. Nooteboom (Springer-Verlag, Berlin), pp. 16–34.

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.

Seneff, S., and Zue, V. (**1988**). "Transcription and alignment of the TIMIT database," in *Proceedings of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language*, Oahu, HI (20–22 November 1988).

Stevens, K. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.

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.

Stickney, G., Zeng, F.-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.