Masking release and the contribution of obstruent consonants on speech recognition in noise by cochlear implant users

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

(Received 3 September 2009; revised 18 June 2010; accepted 18 June 2010)

Cochlear implant (CI) users are unable to receive masking release and the reasons are unclear. The present study examines the hypothesis that when listening to speech in fluctuating maskers, CI users cannot fuse the pieces of the message over temporal gaps because they are not able to perceive reliably the information carried by obstruent consonants (e.g., stops). To test this hypothesis, CI users were presented with sentences containing clean obstruent segments, but corrupted sonorant segments (e.g., vowels). Results indicated that CI users received masking release at low signal-to-noise ratio levels. Experiment 2 assessed the contribution of acoustic landmarks alone by presenting to CI users noise-corrupted stimuli which had clearly marked vowel/consonant boundaries, but lacking clean obstruent consonant information. These stimuli were created using noise-corrupted envelopes processed using logarithmic compression during sonorant segments and a weakly-compressive mapping function during obstruent segments. Results indicated that the use of segment-dependent compression yielded significant improvements in intelligibility, but no masking release. The results from these experiments suggest that in order for CI users to receive masking release, it is necessary to perceive reliably not only the presence and location of acoustic landmarks (i.e., vowel/consonant boundaries) but also the information carried by obstruent consonants. © *2010 Acoustical Society of America.* [DOI: 10.1121/1.3466845]

PACS number(s): 43.66.Ts, 43.71.Ky [MSS] Pages: 1262–1271

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). NH listeners have 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. The benefit received when listening to speech in fluctuating maskers compared to steady maskers is often called "release of masking." Unlike normal-hearing listeners who benefit greatly from "listening in the dips," cochlear implant listeners are not able to receive masking release when listening to speech in fluctuating maskers. This was confirmed in studies involving cochlear implant users (Nelson *et al.*, 2003; Fu and Nogaki, 2005; Nelson and Jin, 2004; Stickney *et al.*, 2004; Cullington and Zeng, 2008) and in studies involving NH listeners listening to cochlear implant 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 in cochlear implants 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 are available (Stickney *et al.*, 2007; Qin and Oxenham, 2003, 2005). 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 0–15 semitones. Several other hypotheses were investigated. Nelson *et al.* (2003) hypothesized that the fluctuating maskers may cause modulation interference[1] particularly when the signal spectral representation is poor, as is the case with current cochlear implant systems. The study by Qin and Oxenham (2003) indicated that spectral resolution was not the determining factor for the lack of masking release. Stickney *et al.* (2004) observed greater masking with 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 the maskers 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 provide clear evidence as to why CI users do not receive release from masking. In our previous study (Li and Loizou, 2009), we investigated an alternative hypothesis using normal-hearing listeners. The proposed hypothesis was that the CI user's ability to fuse information across temporal gaps is limited by their ability to perceive information carried by obstruent consonants and associated acoustic landmarks.

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

TABLE I. Biographical data of the CI users.

Subject	Gender	Age (yr)	Duration of deafness prior to implantation (yr)	CI use (yr)	No. of active electrodes	Stimulation rate (pulses/s)	Etiology
S1	Female	60	2	4	15	2841	Medication
S2	Male	42	2	4	15	1420	Hydrops/Menier's syndrome
S3	Female	47	>10	5	16	2841	Unknown
S4	Male	70	3	5	16	2841	Unknown
S5	Female	62	<1	4	16	1420	Medication
S6	Female	53	2	4	16	2841	Unknown
S7	Female	40	5	8	14	1420	Genetic

These landmarks have been posited to play an important role in models of lexical access (Stevens, 2002) as they are hypothesized to be used by NH listeners in the identification of word/syllable boundaries. This hypothesis was tested by providing normal-hearing listeners with vocoded speech that contained clean obstruent consonants (and therefore clean acoustic landmarks) but corrupted sonorant segments (e.g., vowels). Results were consistent with the above hypothesis as listeners performed better in fluctuating maskers than in steady-noise maskers when they were provided with clean obstruent consonant information along with access to the acoustic landmarks in the signal.

In the present study, we will test in Experiment 1 the same hypothesis as in Li and Loizou (2009), but with CI users. Given that the CI users in experiment 1 will have access to both clean obstruent consonant information and clear acoustic landmarks (e.g., stop closures) associated with the presence of obstruent consonants, we investigate in experiment 2 the contribution of acoustic landmarks alone to speech recognition in noise. This was done by presenting to CI users stimuli which had clearly marked vowel/consonant boundaries, but lacking clean obstruent consonant information. At issue is whether having access to clear acoustic landmarks alone is sufficient to receive masking release. Experiment 2 also evaluates a secondary hypothesis that envelope compression might be partially responsible for the lack of masking release in cochlear implants. More precisely, we test the secondary hypothesis that envelope compression, which is commonly used in cochlear implants for mapping the acoustic signal to the limited electrical dynamic range, amplifies the weak consonants along with noise, thereby smearing the acoustic landmarks. The main question probed in experiment 2 is whether having access to clear acoustic landmarks is sufficient to receive masking release. To test this hypothesis, we investigate the use of selective envelope compression wherein a log-compressive mapping is used during the sonorant segments and a weakly compressive mapping is used during the weak consonant segments. The underlying motivation behind selective compression is to suppress the envelopes of the weak consonants and make the vowel/consonant boundaries more evident to the CI user. The detrimental effect of envelope compression in noise was also reported in prior studies (e.g., Fu and Shannon, 1999), but not in the context of examining its influence on masking release. Finally, in experiment 3 we investigate the viability of an approach that detects automatically the presence of sonorant/obstruent segments from corrupted speech stimuli. Such an approach could be used in a realistic setting to first classify each segment as sonorant/obstruent, and subsequently apply selective compression.

II. EXPERIMENT 1: MASKING RELEASE BY COCHLEAR IMPLANT USERS

A. Methods

1. Subjects

A total of seven postlingually deafened Clarion CII implant users participated in this experiment. All subjects had at least 3 years of experience with their implant device. The biographical data for each subject are given in Table I.

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., Houston, TX) 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 speech-shaped noise (SSN), which is continuous (steady-state) and had the same long-term spectrum as the test sentences in the IEEE corpus. The second masker was a two-talker (TT) competing speech (female) recorded in our laboratory. Two long sentences, taken from the IEEE database and produced by a female talker, were concatenated and used as the TT masker. 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 the vowels, semivowels and nasals. The segmentation procedure was described in detail in Li and Loizou (2008, 2009).

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 16 channels using sixth-order Butter-

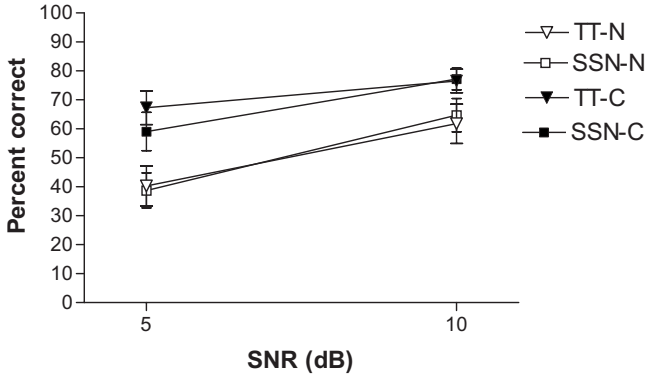

FIG. 1. Mean speech recognition scores as a function of SNR level and type of masker (TT=two-talker and SSN=steady noise). Filled symbols denote scores obtained with stimuli containing clean obstruent consonants (denoted with the suffix–C), and open symbols denote scores obtained with the control stimuli containing corrupted obstruent consonants (denoted with the suffix–N). Error bars indicate standard errors of the mean.

worth filters. Logarithmic filter spacing was used to allocate the channels across a 300–5500 Hz bandwidth. 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 log compressed [see Eq. (1) in experiment 2] to the subject's electrical dynamic range. The same parameters (e.g., stimulation rate, pulse width, etc.) used in the subject's daily strategy were used. The speech stimuli were generated using the above algorithm in two different conditions. In the first condition, which served as the control condition, the corrupted speech stimuli were left unaltered. That is, the obstruent consonants (and sonorants) remained corrupted by the masker. In the second condition, the speech stimuli contained clean (uncorrupted) obstruent segments but corrupted sonorant segments (e.g., vowels). The clean obstruent segments were extracted from the speech stimuli prior to their mixing with the masker stimuli. In the clean obstruent condition, the nominal SNR increases, but only by 1.5 dB (Li and Loizou, 2008).

4. Procedure

The above stimuli were generated off-line in MATLAB and presented directly to CI users via the Clarion Research Interface platform. Prior to the test, subjects listened to some sentences to become familiar with the processed stimuli. The training session lasted for about 20–30 min and involved listening to 40 sentences. During the test, the subjects were asked to write down the words they heard. Subjects participated in a total of 8 conditions (=2 SNR levels $\times$2 algorithms$\times$2 maskers). 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 and discussion

The mean scores for all conditions are shown in Fig. 1. Performance was measured in terms of the percentage of words identified correctly (all words were scored). Two-way ANOVA with repeated measures was used to assess the effect of masker type. The control noisy stimuli (shown in Fig. 1 with open symbols) showed no significant effect of masker type [$F(1,6)=0.034$, $p=0.86$]. No significant interaction [$F(1,6)=2.3$, $p=0.179$] was found between SNR level and masker type. Performance with two-talker masker was not higher to that attained when using steady noise, consistent with findings reported in other cochlear implant studies (e.g., Stickney *et al.*, 2004).

A different pattern in performance emerged in the conditions in which the obstruent consonants were clean and the remaining sonorant sounds were left corrupted (shown in Fig. 1 with filled symbols). Performance obtained with the two-talker masker was higher than performance obtained with the steady noise masker, at least at 5 dB SNR. Two-way ANOVA showed a significant effect [$F(1,6)=19.5$, $p=0.004$] of SNR level, a non-significant effect [$F(1,6)=2.6$, $p=0.157$] of masker type and a significant interaction [$F(1,6)=9.91$, $p=0.02$]. The interaction between SNR level and masker type was due to the fact that a larger improvement was observed for the low SNR level (5 dB) condition compared to the higher SNR (10 dB) condition. More specifically, performance improved (relative to the corrupted stimuli) by roughly 20–30 percentage points at 5 dB SNR, and by 10–15 percentage points at 10 dB SNR. Post-hoc tests revealed that performance in the two-talker masker conditions (with clean obstruent consonants) was significantly higher ($p=0.03$) than the corresponding performance in SSN conditions at 5–dB SNR, but the difference was not statistically significant ($p=0.77$) at 10 dB SNR. This outcome suggests that the CI users received masking release, at least in the low-SNR condition, when they had access to the clean obstruent consonants.

Introducing the clean obstruent consonants in the corrupted vocoded stimuli produced a substantial improvement in performance in both SNR conditions (Fig. 1). The magnitude of the improvement obtained by the CI users when provided with access to information carried by the clean obstruent consonants seemed to depend on the input SNR level. At 5 dB SNR, the improvement ranged from a low of 20 percentage points (in the SSN masker condition) to nearly 30 percentage points (in the two-talker masker condition). The improvement was smaller at 10 dB SNR and ranged from 12–15 percentage points. This SNR dependency is probably due to the different set of acoustic cues (and reliability of those cues) available to the listeners when presented with spectrally degraded speech. The fact that masking release was observed only at low SNR levels is consistent with the outcomes of prior studies with normal-hearing listeners (Oxenham and Simonson, 2009; Bernstein and Grant, 2009). In the study by Oxenham and Simonson (2009), for instance, masking release was observed only for negative target-to-mask ratios (TMRs) under conditions in which the low or high frequency regions of the spectrum were removed (no masking release was observed for positive TMR values). In that study, as well as in ours, masking release was assessed in conditions wherein spectral information was degraded and speech redundancy was reduced. In our study, when CI users

were presented with a severely degraded (spectrally) signal at low SNR levels and were provided with access to the obstruent consonants and associated landmarks, they were able to exploit the dips in the fluctuating masker, much like normal-hearing listeners are able to do. At higher SNR levels, however, there is little room for confusion between the target and masker, and CI users likely utilize other cues available to them. Consequently, they rely less on exploiting the masker dips in the envelopes.

The outcomes of this experiment are consistent with those observed in our previous study (Li and Loizou, 2009) wherein NH listeners were presented with vocoded speech processed in 6–22 channels. In that study, we showed that the magnitude of the improvement depended on both the SNR level and number of channels. The largest improvement (50 percentage points at 0 dB SNR) was obtained in the two-talker masker conditions when speech was vocoded using a relatively large number (22) of channels. Smaller improvement was obtained with 6–12 channels and ranged from 20–30 percentage points with 6 channels to 50 percentage points with 12 channels. Compared to the improvement observed in Li and Loizou (2009), the improvement seen in the present experiment falls within the range seen for 6–12 channels of stimulation in the simulation study. This is consistent with the belief that CI users receive only a limited number (6–8) of channels of information (e.g., Friesen *et al.*, 2001).

As mentioned earlier, the nominal SNR of the clean obstruent stimuli was slightly larger (by 1.5 dB) than the control (corrupted) stimuli. The large improvement in performance obtained, however, with the clean obstruent stimuli (but otherwise corrupted sonorant segments) cannot be explained by this small increase in the nominal SNR. There are several possible underlying mechanisms responsible for the above improvement in performance (Li and Loizou, 2009). For one, listeners had access to multiple spectral/temporal cues when the clean obstruent consonants were introduced, although the saliency of those cues was highly dependent on the SNR level. Additionally, CI users had better access to F1/F2 transitions to/from the vowel and sonorant sounds, more accurate voicing information and consequently better access to acoustic landmarks which perhaps aided the listeners in identifying more easily word boundaries in the noisy speech stream. Of the above mechanisms, we believe that two contributed the most to the large improvement in intelligibility reported in Fig. 1: (1) access to clean obstruent consonant information, and (2) access to clear acoustic landmarks signaling the presence of vowel/consonant boundaries. In the present experiment, the CI users had access to both clean obstruent consonant information and to clear acoustic landmarks. Hence, it was not clear which of the two factors contributed the most and whether both were required to receive masking release. The next experiment was designed to answer this question.

III. EXPERIMENT 2: IMPACT OF HAVING ACCESS TO CLEAR ACOUSTIC LANDMARKS ON SPEECH RECOGNITION IN NOISE

It was not clear from experiment 1 whether CI users obtained improvements in intelligibility because they had access to clean obstruent consonant information, had access to clear acoustic landmarks or because they had access to both. The present experiment was designed to assess the contribution of acoustic landmarks alone to speech recognition in noise. This was done by presenting to CI users stimuli which had clearly marked vowel/consonant boundaries, but lacking clean obstruent consonant information. In doing so, we can examine the individual contribution of having clear acoustic landmarks to speech recognition in noise.

The stimuli used in the presented experiment were created by processing noise-corrupted sentences via an algorithm that compressed the envelopes (in the low frequency channels) using a logarithmic acoustic-to-electric mapping function during the sonorant segments (e.g., vowels) and a weakly-compressive mapping function during the obstruent segments (e.g., stops) of the sentence. The underlying motivation for the use of the weakly-compressive mapping function applied during obstruent segments is to maintain a more natural vowel-to-consonant ratio, which in turn would make the acoustic landmarks more evident. The above selective envelope compression was applied only to the low frequency channels (<1 kHz) of the corrupted stimuli. The envelopes of the high frequency channels were left unaltered. In doing so, CI users were provided with clear acoustic landmarks but corrupted obstruent consonant information. This enabled the listeners to identify the location of the weak consonants (i.e., obstruents) in the corrupted speech stream, without necessarily allowing the listeners to perceive reliably their identity. The main question probed in this experiment is whether having access to clear acoustic landmarks alone is sufficient to receive masking release.

A. Methods

1. Subjects and stimuli

The same seven subjects who participated in Experiment 1 also participated in the present experiment. The testing was carried out during the second day of the subjects' visit to our laboratory. The same maskers were used as in Experiment 1 and a different set of sentences taken from the IEEE corpus was used.

2. Signal processing

The signal processing strategy used by the CI users is the same as in Experiment 1. The main difference lies in the use of two different acoustic-to-electric mappings, which are applied to the corrupted envelopes depending on the phonetic segment present in the sentences. For sonorant segments (e.g., vowels) a logarithmic mapping is used (same as used in the CI user's daily strategy), while for obstruent segments a less compressive mapping function is utilized. The acoustic-to-electric mapping is implemented as follows:

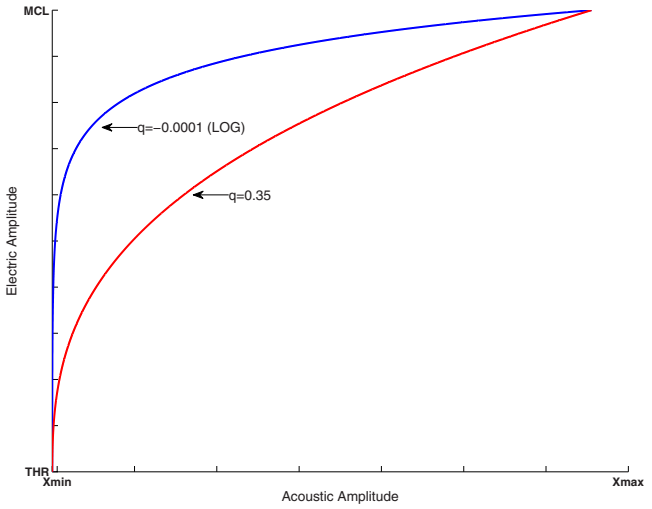

FIG. 2. (Color online) Plot of the two acoustic-to-electric mapping functions used in Experiment 2. The acoustic signal spanning the range of [Xmin, Xmax], is mapped to the electrical output signal spanning the range of [THR, MCL], where THR is the threshold and MCL is the most-comfortable level (expressed in clinical units or microamperes).

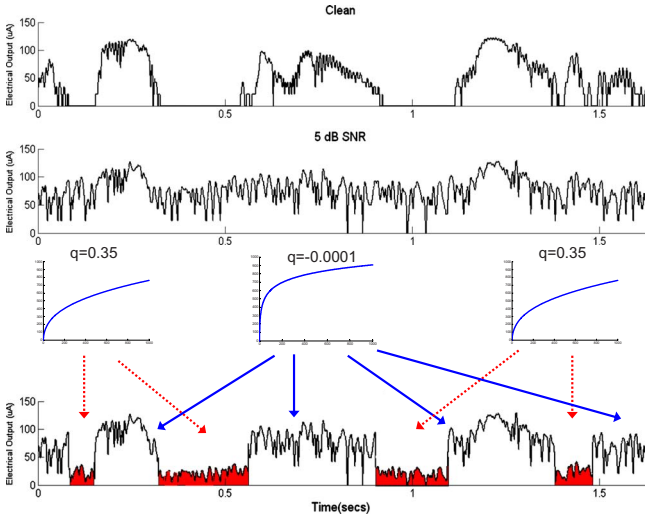

FIG. 3. (Color online) The envelope (4th channel with center frequency of 600 Hz) of the clean signal (taken from a sentence) is shown in the upper panel, and the envelope of the corrupted envelope in 5 dB SSN is shown in the middle panel [log compression was used, i.e., $q=-0.0001$ in Eq. (1)]. The bottom panel illustrates the operation of selective compression in which a log compressive function is used during sonorant segments and a weakly compressive function ($q=0.35$) is used during the obstruent segments. The envelopes were computed for the partial sentence "The birch canoe slid…" taken from the IEEE corpus.

$$Y(n) = A \cdot [X(n)]^q + B, \tag{1}$$

where $Y(n)$ indicates the electrical amplitude output (measured in clinical units or microamperes) at time n, $X(n)$ denotes the acoustic envelope amplitude, and the constants A and B are used to ensure that the acoustic amplitudes are mapped within the electrical dynamic range. The power exponent q is used in the present study to control the steepness of the compression function. In the present study, the value of $q=-0.0001$ was used for log compression and the value of $q=0.35$ was used for weak (less) compression. The two mapping functions used are shown in Fig. 2.

The speech stimuli are processed in two different conditions. In the first control condition, the corrupted speech stimuli are processed using the log compression, as used in their daily strategy. In the second condition, the corrupted speech envelopes are compressed using a logarithmic-shaped function ($q=-0.0001$) during sonorant segments (e.g., vowels) and a less-compressive mapping function ($q=0.35$) during obstruent segments (e.g., stops). The weakly compressive function is only applied to the low frequency channels, and more precisely, the seven most apical channels spanning the bandwidth of 200–1000 Hz. The remaining nine higher-frequency channels are processed using the log-mapping function. For subjects with only 14–15 active electrodes (see Table I), the remaining 7–8 higher-frequency channels are processed using the log mapping function. The motivation for applying a different mapping function in the low frequencies is to make the low-frequency phonetic boundaries more evident without suppressing the high-frequency cues commonly present in most obstruent consonants (e.g., /t/). It should be pointed out that the selective compression is applied to the *corrupted* speech envelopes and only during the obstruent segments of the sentences. That is, unlike the conditions in Experiment 1, in the present experiment both sonorant and obstruent segments in the sentences remained noise corrupted. Similar to experiment 1, it is assumed that we have access to the true sonorant/obstruent consonant acoustic boundaries. The detection of these boundaries can alternatively be done using an automatic algorithm, and this is investigated in experiment 3.

Figure 3 shows as an example a noise-corrupted envelope (with center frequency=600 Hz) processed using selective compression. The sentence was corrupted in SSN at 5 dB SNR. By comparing the bottom two panels, it is clear that the use of selective compression renders the (low-frequency) consonant boundaries more evident and perhaps perceptually more salient. The effect of applying a weakly compressive mapping function to the low-frequency region of the obstruent segments is evident in the bottom panel. As can be seen from this panel, the envelopes are attenuated relative to the envelopes in the sonorant segments, thereby rendering the vowel/consonant boundaries (present in the low frequencies, i.e., <1000 Hz) more clear.

3. Procedure

The above stimuli were generated off-line in MATLAB and presented directly to CI users via the Clarion Research Interface platform. Subjects participated in a total of 8 conditions (=2 SNR levels×2 algorithms×2 maskers). Two lists of IEEE sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions or experiments. In order to eliminate possible learning effects, we repeated the control conditions involving the corrupted stimuli. None of the sentence lists chosen for Exp. 2 were used in Exp. 1. 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.

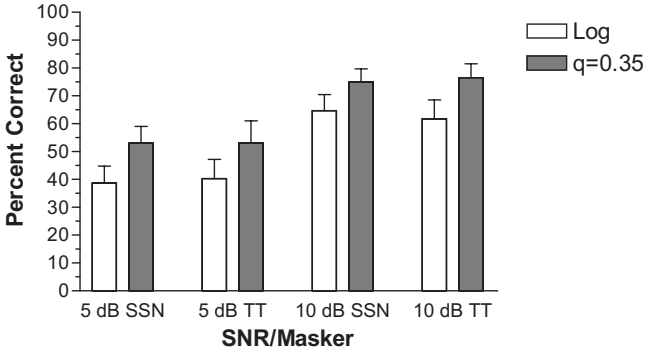

FIG. 4. Mean speech recognition scores as a function of SNR level (SSN masker) of the control stimuli in which all segments were processed using logarithmic compression (denoted as Log) and the stimuli in which corrupted sonorant segments were processed using log compression while the corrupted obstruent segments were processed using a weakly compressive function ($q=0.35$). Error bars indicate standard errors of the mean.

B. Results and discussion

The mean scores for all conditions are shown in Fig. 4. Performance was measured in terms of the percentage of words identified correctly (all words were scored). Two-way ANOVA, with repeated measures, was used to assess the effect of masker type. ANOVA indicated no significant [$F(1,6)=0.102$, $p=0.762$] effect of masker type and no significant interaction [$F(1,6)=1.23$, $p=0.317$] between SNR level and masker type. This suggests that the selective compression did not provide masking release.

Additional analysis was conducted to assess whether the use of selective compression improved performance relative to that obtained with the un-processed stimuli, which were processed (for all phonetic segments) using the log-mapping function. Two-way ANOVA, run on the scores obtained in the SSN conditions, indicated significant effect [$F(1,6)=161.7$, $p<0.0005$] of SNR level, significant effect [$F(1,6)=36.8$, $p=0.001$] of the method of compression and non-significant interaction [$F(1,6)=1.2$, $p=0.307$]. Similarly, two-way ANOVA, run on the scores obtained in the two-talker masker conditions, indicated significant effect [$F(1,6)=10.24$, $p=0.024$] of SNR level, significant effect [$F(1,6)=23.4$, $p=0.005$] of the method of compression and non-significant interaction [$F(1,6)=1.8$, $p=0.235$]. The above analysis clearly indicates that the use of selective compression can improve significantly speech intelligibility at both SNR levels, relative to the baseline condition.

As shown in Fig. 4, selective compression improved performance by nearly 15 percentage points at 5 dB SNR and by approximately 10–15 percentage points at 10 dB SNR. The improvement in performance was found to be consistent for both types of maskers. As mentioned earlier, both sonorant and obstruent segments in the sentences were left noise corrupted. Yet, a consistent, and statistically significant, improvement was noted when selective compression was applied to the low-frequency channels. The above outcome is consistent with that observed in our previous study (Li and Loizou, 2008) with normal-hearing listeners. In that study, listeners were presented with corrupted stimuli (in babble) which contained clean obstruent spectral information in the 0–1000 Hz region, and noise corrupted information in the higher frequencies (the sonorant segments remained noise corrupted). Results indicated that access to the low-frequency (0–1000 Hz) region of the clean obstruent-consonant spectra was sufficient to realize significant improvements (about 10–15 percentage points) in performance and that was attributed to improvement in transmission of voicing information. Hence, we deduce that by applying selective compression in the low-frequencies (as done in the present study) we can improve significantly the transmission of voicing information.

The improvement obtained with selective envelope compression was not as large as that obtained in Experiment 1 (about 20–30 percentage points) when the listeners had access to the clean obstruent consonant spectra, and subsequently clear acoustic landmarks. No masking release was found in the present experiment. We attribute this to the following reasons. First, unlike the conditions tested in Experiment 1, the obstruent consonants in Experiment 2 were either left corrupted or were suppressed (at least in the low frequencies). Consequently, listeners did not have access to clean obstruent consonant information. That, in turn, impaired their ability to fuse (or "glimpse") pieces of the target signal across temporal gaps (Li and Loizou, 2008, 2009). Second, the low-frequency envelope suppression was done without taking into account the spectral content or spectral energy distribution of the obstruent consonants at hand. The labial stop consonants (e.g., /p/, /b/), for instance, are characterized by low-frequency energy concentration; hence suppressing the low-frequency region might introduce conflicting (burst) cues to the listeners. The presence of conflicting burst cues will in turn force listeners to rely on formant transitions (Dorman *et al.*, 1977; Dorman and Loizou, 1996), which we know that CI listeners cannot perceive reliably (Munson and Nelson, 2005). On the other hand, the alveolar stop consonants and fricatives (e.g., /t/, /s/) have high-frequency energy concentration and the applied envelope suppression would be more appropriate and more likely to be beneficial. In all, selective envelope compression cannot reach its full potential (as demonstrated in Experiment 1) given that it is applied to *all* obstruent consonants.

The impact of envelope compression on speech recognition (in quiet and in noise) was also examined in other studies (Fu and Shannon, 1998, 1999). The data in Fu and Shannon (1999), for instance, indicated that the nonlinear acoustic-to-electric mapping had only a minor effect on phoneme recognition in quiet, consistent with the previous findings with normal-hearing listeners (Fu and Shannon, 1998). However, as the SNR level decreased, the effect of nonlinear mapping became dramatic and asymmetric: performance with weakly compressive mappings declined mildly in noise, but performance declined dramatically in noise with a strongly compressive amplitude mapping. This outcome is partially consistent with the findings of the present experiment. Performance with the strongly compressive mapping was significantly worse (see Fig. 4) than performance with the (selective) weak mapping ($q=0.35$). Hence, in agreement with prior studies (Fu and Shannon, 1999), we can conclude that the use of a strongly compressive mapping function that

is applied to all phonetic segments is not appropriate, or beneficial, for CI users when listening to speech in noisy environments.

A selective compression function was proposed in this experiment for enhancing access to the acoustic landmarks in noisy conditions. 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 input-output 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 and energy compared to that of sonorant segments. The main advantage of using s-shaped functions for mapping the acoustic signal to electrical output 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 cochlear implant users (Kasturi and Loizou, 2007). From a practical point of view, the s-shaped approach (Kasturi and Loizou, 2007) is more attractive as it does not require explicit detection of acoustic landmarks. However, it requires reliable estimation of the noise floor level for accurate determination of the knee point.

The use of different degrees of compression in the high and low frequency channels has also been explored in hearing aids (Killion *et al.*, 1990; Dillon, 2001). When a higher compression is applied in the high-frequency channels than the low frequency channels, a higher frequency emphasis is produced particularly at low input levels. This effect has been referred to as a treble increase at low levels (TILL) in the hearing-aid literature (Killion *et al.*, 1990). The main difference between the TILL approach used in hearing aids and our approach is that the latter is phonetic-segment dependent, in that it is selectively applied only to the obstruent segments of the utterance.

Combining the findings for Experiments 1 and 2, we can reach the conclusion that in order for CI users to receive masking release it is necessary for them to perceive reliably not only the acoustic landmarks (associated with obstruent consonants) but also the information carried by the low-energy weak consonants (e.g., stops). The perception of weak consonants can be facilitated or mediated, at least to some extent, by reliable detection of the vowel/consonant boundaries. As demonstrated in the present experiment, providing access to the vowel/consonant boundaries to CI users produced significant improvement in performance, but that was not sufficient to observe masking release owing to the fact that the weak consonants were left noise corrupted. Put differently, the use of selective compression enabled the listeners to identify the location (the *where*) of the weak consonants in the corrupted speech stream, but did not allow the listeners to perceive reliably their identity (the *what*). Further (selective) enhancement, perhaps by a noise-reduction algorithm, of the weak consonants might be needed to obtain both (location and identity of weak consonants), and subsequently observe masking release.

IV. EXPERIMENT 3: AUTOMATIC DETECTION OF VOWEL/CONSONANT BOUNDARIES

In the previous experiment, the true boundaries for sonorant/obstruent segments were assumed to be known. In a realistic scenario, one can envision an algorithm that first detects the location of those boundaries and then removes or suppresses the noise from the corrupted weak consonants. Such an algorithm would require automatic detection of sonorant/obstruent boundaries in noise. The performance of an algorithm that automatically detects sonorant/obstruent boundaries is assessed in the present experiment. Following the classification into sonorant or obstruent segments, the speech envelopes are selectively compressed (as in Exp. 2) and presented to CI users for identification.

A. Methods

1. Subjects and stimuli

The same seven subjects who participated in previous experiments also participated in the present experiment. The testing was carried out during the third day of the subjects' visit to our laboratory. Sentences taken from the IEEE corpus (same as in Experiment 1) were used. Due to the limited number of sentence lists available in the IEEE corpus, only a single masker, namely speech-shaped noise, was used in this Experiment.

2. Signal processing

A two-class Bayesian classier (Duda *et al.*, 2001) was used for detecting and classifying sonorant and obstruent segments from the corrupted sentences. Sixteen band energy values, extracted from 16 channels, were used as features. The filter spacing was the same as used in the CI user's strategy. The probability distribution of the feature vectors of each class was represented with a Gaussian Mixture Model (GMM) (Paalanen *et al.*, 2006). The features were computed as follows. The corrupted speech stimuli were first segmented into 10-ms frames, with 5 ms overlap, and then Hann windowed. A 256-point Fast Fourier Transform (FFT) was computed of each frame and the FFT bins were distributed across 16 channels, based on the filter spacing used in the CI user's strategy. The total energy of each band was computed by summing up the energy of all bins falling within each band. The band energy was log compressed and then smoothed (across time) using a first-order recursion with the smoothing constant set to 0.5. Finally, the smoothed band energy values were linearly mapped to the range of [0, 1]. A total of 16 normalized band-energy values (extracted from the noise-corrupted signals) were used as features for the Bayesian classifier.

We utilized 64-mixture Gaussian models for modeling the distributions of the feature vectors in each class. The initial Gaussian model parameters (mixture weights, mean vectors and covariance matrices) were obtained by running several iterations of the k-means clustering algorithm (Duda *et al.*, 2001). The GMM parameters were obtained using the expectation-maximization training algorithm (Dempster *et*

TABLE II. Classification performance, in terms of hit and false alarm rates, for the Bayesian classifier of sonorant/obstruent segments.

Noise type	Hit (%)	False alarm (%)
5 dB SSN	94.71	10.7
10 dB SSN	93.37	8.91

al., 1977). A total of 180 sentences from the IEEE database were used to train the Bayesian classifier. A different set of sentences, not used in training, was used to test the performance of the Bayesian classifier.

Following the training, the trained GMM model parameters were stored in memory, and then used during the testing stage to classify each 10-ms frame into sonorant and obstruent sounds. Hence, unlike experiment 1 wherein the true sonorant/obstruent boundaries were used, in the present experiment we identify the sonorant/obstruent boundaries automatically using the Bayesian classifier. Following the classification of each corrupted frame of speech as sonorant or obstruent sounds, the speech stimuli were processed using the selective compression method presented in experiment 2. More specifically, the corrupted speech envelopes were compressed using a logarithmic-shaped function ($q=-0.0001$) during sonorant segments (e.g., vowels) and a less-compressive mapping function ($q=0.35$) during obstruent segments (e.g., stops). As before, the weakly compressive function was applied only to the low frequency channels, and more precisely, the seven most apical channels spanning the bandwidth of 200–1000 Hz. In the control condition, the corrupted speech stimuli were log compressed, as done in the CI user's daily strategy.

The performance of the binary Bayesian classifier was assessed quantitatively in terms of detection rate (correctly classifying obstruent segments) and false-alarm rate (incorrectly classifying sonorant segments as obstruent segments). The detected sonorant/obstruent segments were compared (within 5 ms) with the corresponding manual sonorant/obstruent transcriptions of the IEEE database. A total of 540 IEEE sentences (not used in the training) in 5 and 10 dB SNR were used for testing. The results are tabulated in Table II. As can be seen, high detection rate (>93%), within 5 ms, was obtained with relatively low false alarm (<11%) rate for both SNR levels tested.

3. Procedure

The processed stimuli were generated off-line in MATLAB and presented directly to CI users via the Clarion Research Interface platform. Subjects participated in four conditions (=2 SNR levels×2 algorithms). Two lists of IEEE sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions. In order to eliminate possible learning effects, we repeated the control conditions involving the corrupted stimuli. None of the sentence lists chosen for Exp. 3 were used in either Exp. 1 or 2. Sentences were presented to the listeners in blocks, with 20 sentences/block for each condition. The presentation order of the different conditions was randomized across the subjects.

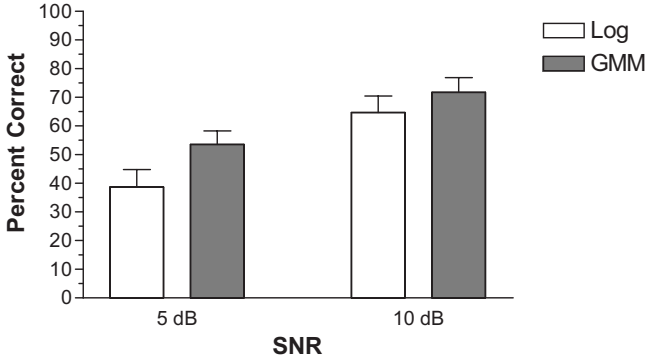

FIG. 5. Mean speech recognition scores as a function of SNR level (SSN masker) of the control condition in which all segments were processed using logarithmic compression (labeled as Log) and the condition (labeled as GMM) in which the sonorant and obstruent (corrupted) segments were detected using a GMM Bayesian classifier, and subsequently compressed selectively using the log (during sonorant segments) and weakly compressive functions (during obstruent segments). Error bars indicate standard errors of the mean.

B. Results

The mean intelligibility scores obtained with selective compression when the GMM sonorant/obstruent sound classifier was used for detection of the sonorant/obstruent boundaries are shown in Fig. 5. Listener performance was measured in terms of the percentage of words identified correctly (all words were scored). Statistical analysis was conducted to assess whether the use of GMM classifier improved performance relative to that obtained with the un-processed stimuli. Two-way ANOVA indicated significant effect [$F(1,6)=51.7$, $p<0.0005$] of SNR level, significant effect [$F(1,6)=13.83$, $p=0.01$] of processing with the GMM classifier and nonsignificant interaction [$F(1,6)=2.92$, $p=0.138$].

The above analysis suggests that the GMM classifier yielded reliable detection of sonorant/obstruent boundaries, and when used in conjunction with selective envelope compression (Exp. 2) produced significant improvements in intelligibility at both SNR levels. Comparing the performance obtained using the true sonorant/obstruent boundaries in Experiment 2 with the performance obtained with the GMM classifier (Exp. 3) we note that it is nearly identical. In brief, the classification accuracy (see Table II) of the GMM classifier seems to be sufficiently high to observe statistically significant improvements in intelligibility. This makes the proposed selective envelope compression approach (Exp. 2) a viable approach for use in realistic scenarios. The GMM classifier was trained in the present experiment in SSN noise, but can be easily trained and extended to other noisy environments. That is, in a realistic scenario different GMM classifiers can be trained for different listening environments encountered by the CI users. The parameters of the trained classifier could be stored in the processor and activated when the user visits a particular listening environment. Further research is warranted to investigate that.

V. SUMMARY AND CONCLUSIONS

The present study examined the longstanding question as to why CI users are not able to receive masking release. The hypothesis posed and tested was that CI users are not able to receive masking release because they are not able to fuse the pieces of the message "glimpsed" over temporal gaps owing to the fact that they are unable to perceive reliably the information carried by the severely masked obstruent consonants (e.g., stops). That is, while it seems easy for CI users to perceive reliably the sonorant segments (e.g., vowels) of the utterance in noise, it is considerably more difficult to perceive the obstruent segments, as those are easily masked by noise. Hence, by providing to CI users access to clean obstruent consonant information, we would expect the CI users to be able to better integrate pieces of the message "glimpsed" across the noise-corrupted utterance. This hypothesis was tested in Experiment 1 by presenting to listeners stimuli containing corrupted sonorant segments (e.g., vowels) but clean obstruent consonants (e.g., stops). Results indicated substantial improvement (20–30 percentage points) in intelligibility, particularly at low SNR levels (5 dB). Performance in the 2-talker masker conditions (5 dB SNR) was found to be significantly higher than performance in the SSN conditions, thus demonstrating that CI users can receive masking release.

Experiment 2 focused on answering the question: What contributed the most to the large improvement in intelligibility observed in experiment 1? Was it access to clean obstruent consonant information, access to clear acoustic landmarks or access to both? The answers to these questions will help us understand the absence of masking release observed in cochlear implant users. We hypothesized that the envelope compression, which is commonly implemented in CI devices, is partially responsible for that, as it smears the acoustic landmarks, particularly at low SNR levels. The smearing is caused by the fact that the use of log envelope compression tends to amplify the low-energy weak consonants (e.g., fricatives and stops), thus distorting the inherent vowel-to-consonant energy ratio in natural speech (see examples in Figs. 4 and 5 in Li and Loizou, 2009). Hence, by making the vowel/consonant landmarks more distinct, we would expect the CI users to better integrate "glimpsed" segments over the utterance. This hypothesis was tested by using selective envelope compression wherein a weakly compressive function was applied during the obstruent consonants (in the low frequencies, within 1 kHz) and a relatively strong (log) compressive function was applied during sonorant segments. This had the effect of suppressing the envelopes of the obstruent consonants in the low frequencies, thereby making the vowel/consonant boundaries more evident. Results revealed a significant improvement in intelligibility when selective envelope compression was used, but no evidence of masking release. This was attributed to the fact that the CI users had a clear access to the vowel/consonant boundaries, but perhaps perceived conflicting spectral information since the high-frequency region (>1 kHz) of the obstruent consonants was left corrupted. The significant improvement in performance obtained with selective compression applied to the low frequencies, was attributed to the better transmission of voicing information. Considering together the outcomes from Experiments 1 and 2, we can conclude that in order for CI users to receive masking release, it is necessary for them to perceive reliably not only the presence and location of the vowel/consonant boundaries (as tested in Exp. 2) but also the information contained in the low-energy obstruent consonants (as tested in Exp. 1). Both types of information were available to the CI users in experiment 1. Put simply, CI users are not able to receive masking release because they do not perceive reliably the obstruent consonants in noise. These consonants have low energy and they are easily masked in noise, more so than vowels (Li and Loizou, 2008; Parikh and Loizou, 2005; Phatak and Allen, 2007). The situation is further exacerbated by the fact that a rather strongly compressive mapping is typically used in cochlear implants, which in turn smears the vowel/consonant boundaries (thus making it difficult to detect the presence/absence of consonants) and amplifies the already noise-masked weak consonants. The use of selective envelope compression (as done in Exp. 2) seems to be more appropriate for processing speech in noise. Finally, Experiment 3 demonstrated a viable approach for the automatic detection of vowel/consonant boundaries. In this approach, which was designed to address a realistic setting, selective envelope compression was applied to the noisy envelopes based on decisions made by a binary classifier that detected automatically the vowel/consonant boundaries. Significant improvements in intelligibility were noted, consistent with the findings reported in Experiment 2.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC007527 from the National Institute of Deafness and Other Communication Disorders, NIH. The authors would like to thank the Associate Editor, Dr. Mitchell Sommers, and the anonymous reviewers for all their useful and constructive comments.

[1]Modulation interference refers to the phenomenon whereby speech recognition could be degraded by a modulated masker due to interference with auditory processing of the speech envelope (Kwon and Turner, 2001).

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

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.

Dempster, A., Laird, N., and Rubin, D. (**1977**). "Maximum-likelihood from incomplete data via the EM algorithm (with discussion)," J. R. Stat. Soc. Ser. B (Methodol.) **39**, 1–38.

Dillon, H. (**2001**). *Hearing Aids* (Thieme, New York), pp. 169–174.

Dorman, M., and Loizou, P. (**1996**). "Relative spectral change and formant transitions as cues to labial and alveolar place of articulation," J. Acoust. Soc. Am. **100**, 3825–3830.

Dorman, M., Studdert-Kennedy, M., and Raphael, L. (**1977**). "Stop consonant recognition: Release bursts and formant transitions as functionally equivalent context-dependent cues," Percept. Psychophys. **22**, 109–122.

Duda, H. P., Hart, R. O., and Stork, D. (**2001**). *Pattern Classification* (Wiley, New York), Chap. 10, pp. 20–30.

Festen, J., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfer-

ing speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
Friesen, L. M., Shannon, R. V., 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.-J., and Shannon, R. V. (**1999**). "Phoneme recognition by cochlear implant users as a function of signal-to-noise ratio and nonlinear amplitude mapping," J. Acoust. Soc. Am. **106**, L18–L23.
Fu, Q., and Nogaki, G. (**2005**). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," J. Assoc. Res. Otolaryngol. **6**, 19–27.
Fu, Q.-J., and Shannon, R. V. (**1998**). "Effects of amplitude nonlinearity on speech recognition by cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **104**, 2570–2577.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Kasturi, K., and Loizou, P. (**2007**). "Use of s-shaped input-output functions for noise suppression in cochlear implants," Ear Hear. **28**, 402–411.
Killion, M., Staab, W., and Preves, D. (**1990**). "Classifying automatic signal processors," Hearing Instruments **41**, 24–26.
Kwon, B., and Turner, C. (**2001**). "Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference?," J. Acoust. Soc. Am. **110**, 1130–1140.
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.
Li, N., and Loizou, P. (**2009**). "Factors affecting masking release in cochlear implant vocoded speech," J. Acoust. Soc. Am. **126**, 338–348.
Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (Taylor & Francis, Boca Raton, FL), Appendix C, pp. 589–599.
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.
Oxenham, A. J., and Simonson, A. M. (**2009**). "Masking release for low- and high-pass filtered speech in the presence of noise and single-talker interference," J. Acoust. Soc. Am. **125**, 457–468.
Paalanen, P., Kämäräinen, J. K., Ilonen, J., and Kälviäinen, H. (**2006**). "Feature representation and discrimination based on Gaussian mixture model probability densities-practices and algorithms," Pattern Recogn. **39**, 1346–1358.
Parikh, G., and Loizou, P. (**2005**). "The influence of noise on vowel and consonant cues," J. Acoust. Soc. Am. **118**, 3874–3888.
Phatak, S., and Allen, J. (**2007**). "Consonants and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.
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.
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.