Contribution of Consonant Landmarks to Speech Recognition in Simulated Acoustic-Electric Hearing

Fei Chen and Philipos C. Loizou

Objectives: The purpose of this study is to assess the contribution of information provided by obstruent consonants (e.g., stops and fricatives) to speech intelligibility in simulated acoustic-electric hearing. As a secondary objective, this study examines the performance of an objective measure that can potentially be used for predicting the intelligibility of vocoded speech.

Design: Noise-corrupted sentences are used in experiment 1 in which the noise-corrupted obstruent consonants are replaced with clean obstruent consonants, while leaving the sonorant sounds (vowels, semivowels, and nasals) corrupted. In one condition, listeners have only access to the low-frequency (<600 Hz) acoustic portion of the clean consonant spectra, in other condition, listeners have only access to the higher frequency (>600 Hz) portion (vocoded) of the clean consonant spectra, and in the third condition, they have access to both. In experiment 2, we investigate a speech-coding strategy that selectively attenuates the low-frequency portion of the consonant spectra while leaving the vocoded portion corrupted by noise. Finally, using the data collected from experiments 1 and 2, we evaluate the performance of an objective measure in terms of predicting intelligibility of vocoded speech. This measure was originally designed to predict speech quality and has never been evaluated with vocoded speech.

Results: Significant improvements (about 30 percentage points) in intelligibility were noted in experiment 1 in steady and two-talker masker conditions when the listeners had access to the clean obstruent consonants in both the acoustic and the vocoded portions of the spectrum. The improvement was more evident in the low signal to noise ratio levels (−5 and 0 dB). Further analysis indicated that it was access to the vocoded portion of the consonant spectra, rather than access to the low-frequency acoustic portion of the consonant spectra that contributed the most to the large improvements in performance. In experiment 2, a small (14 percentage points) but statistically significant improvement in performance was obtained at 0 dB signal to noise ratio (steady masker) when the obstruent consonants were selectively attenuated in the low-frequency acoustic portion alone (the vocoded portion was left noise corrupted). The examined objective measure predicted with a relatively high correlation ($r = 0.92$ to 0.94) and the intelligibility of vocoded speech improved in both steady and two-talker masking conditions.

Conclusions: Providing access to the clean obstruent spectra can yield substantial improvements in intelligibility relative to the simulated acoustic-electric condition. Much of this improvement can be attributed to the listeners having access to the clean vocoded portion of the obstruent consonants. The large contribution of obstruent consonants in speech recognition in simulated acoustic-electric hearing stems from the fact that these consonants provide reliable acoustic landmarks which in turn enable listener to integrate effectively pieces of the message glimpsed over temporal gaps into one coherent speech stream. It is argued that these landmarks are smeared in existing cochlear implant systems, including the bimodal systems, owing to envelope compression, and the fact that the obstruent consonants are probably the first to be masked by background noise. Overall, the outcomes from this study suggest that the obstruent consonants need to be treated differently for improved speech recognition in noise.

(Ear & Hearing 2010;31;259–267)

Department of Electrical Engineering, The University of Texas at Dallas, Richardson, Texas.

INTRODUCTION

Background noise is known to mask vowels and consonants differently and to a different extent. For one, the low-energy obstruent consonants (e.g., stops) are masked more easily by noise (Parikh & Loizou 2005; Phatak & Allen 2007) than the high-energy vowels and semivowels. A recent study showed that the information carried by the first two vowel formants is preserved to some degree even at low signal to noise ratio (SNR) levels (Parikh & Loizou 2005). However, both the spectral tilt and burst frequency of stop consonants, which are known to be responsible for conveying place of articulation information (Blumstein & Stevens 1979), were significantly influenced by noise. Furthermore, background noise corrupts acoustic landmarks produced by (abrupt) spectral discontinuities, such as those created by the closing and release of stop consonants. These consonant landmarks are believed to be crucial in lexical-access models (Stevens 2002). In brief, the above findings indicate that, though the acoustic cues present in voiced speech segments (e.g., vowels) may be resistant, to some extent, to corruption by noise (Parikh & Loizou 2005), the acoustic cues in the unvoiced and week energy segments (e.g., consonants) are severely completed and sometimes rendered useless. This raises the question as to whether cochlear implant (CI) listeners' degradation of performance in noise can be attributed, at least partially, to the loss of information carried by obstruent consonants.

In this study, we assess the contribution of information provided by obstruent consonants to speech intelligibility in simulated acoustic-electric hearing. These simulations emulate to a certain degree a recent development in CIs, known as the combined electric and acoustic stimulation (EAS). In EAS patients, an electrode array is implanted only partially into the base region of cochlea so as to preserve the residual acoustic hearing at low frequencies (typically 20 to 60 dB HL up to 750 Hz and severe to profound hearing loss at 1000 Hz and above), which many patients still have (von Ilberg et al. 1999; Kiefer et al. 2005; Gantz et al. 2006). The low-frequency and high-frequency (>1000 Hz) speech information is provided to these patients via a hearing aid and a CI, respectively. Thus, these patients perceive speech via a combined EAS mode. A substantial amount of evidence exists supporting the benefits of EAS in terms of better speech recognition, in noisy environments, in studies involving EAS patients (Gantz & Turner 2003; Turner et al. 2004; Kiefer et al. 2005; Kong et al. 2005; Gantz et al. 2006) and simulation studies with normal-hearing (NH) listeners listening to vocoded speech (Dorman et al. 2005; Chang et al. 2006; Qin & Oxenham 2006; Kong & Carlyon, 2007; Li & Loizou, 2008a).

The reasons for the large benefit of EAS in speech recognition in noise are not very clear. This study aims to ascertain whether having reliable access to the low- and/or high-fre-

0196/0202/10/3102-0259/0 • Ear & Hearing • Copyright © 2010 by Lippincott Williams & Wilkins • Printed in the U.S.A.

quency acoustic information helps listeners determine the location of the consonant landmarks, and subsequently aids them to identify the syllable/word boundaries in the speech stream. We do this, in experiment 1, by using noise-corrupted sentences in which we replace the noise-corrupted obstruent consonants with clean obstruent consonants, while leaving the sonorant sounds (vowels, semivowels, and nasals) corrupted. In one condition, listeners have only access to the low-frequency (<600 Hz) acoustic portion of the clean consonant spectra, whereas in other condition, listeners have only access to the higher frequency (>600 Hz) portion (vocoded) of the clean consonant spectra. In experiment 2, we investigate a speech-coding strategy that selectively attenuates the low-frequency portion of the consonant spectra while leaving the vocoded portion corrupted by noise. The motivation for attenuating selectively the acoustic portion of the consonant spectra is to determine whether listeners can better identify the consonant landmarks present in the speech signal, particularly in situations wherein the target is masked by steady continuous noise. Finally, as a secondary goal of this study, we examine the performance of an existing objective measure, known as the Perceptual Evaluation of Speech Quality (PESQ) measure (ITU-T 2000; Rix et al. 2001), in terms of predicting intelligibility of vocoded speech. The PESQ measure was originally designed to assess speech quality (Rix et al. 2001) and was never evaluated in terms of predicting the intelligibility of vocoded speech. For this evaluation, we make use of all the intelligibility scores collected in experiments 1 and 2.

EXPERIMENT 1: CONTRIBUTION OF CLEAN OBSTRUENT CONSONANTS TO SIMULATED ACOUSTIC-ELECTRIC HEARING

Methods

Subjects • Seven NH listeners participated in the experiment. All subjects were native speakers of American English and were paid for their participations. The subjects' age ranged from 19 to 29 yrs, with majority being undergraduate students from The University of Texas at Dallas.

Stimuli • The speech material consisted of phonetically-balanced sentences taken from the Institute of Electrical and Electronics Engineers (IEEE 1969) database. All the sentences were produced by a male speaker and recorded at a 25-kHz sampling rate in a sound-proof booth (Acoustic Systems) in our lab. Details on the recording setup, copies of recordings, and the phonetic labels indicating the unvoiced/voiced boundaries in the IEEE corpus are available from a connected discourse in Loizou (2007). Two types of maskers were used to corrupt the IEEE sentences. The first was continuous steady-state noise (SSN), which had the same long-term spectrum as the test sentences in the IEEE corpus. The second was two equal-level interfering female talkers (2-talker) based on two of the longest (in duration) sentences in the corpus. The same masker segment was used for all sentences. The test sentences were corrupted by the SSN and 2-talker maskers at −5, 0, and 5 dB SNR.

Signal processing • The stimuli were presented in six different processing conditions. The first processing condition was designed to simulate the effect of eight-channel electrical stimulation and used an eight-channel sinewave-excited vocoder (Loizou et al. 1999). Signals were first processed through a preemphasis (highpass) filter (2000 Hz cutoff) with a 3 dB/octave rolloff and then bandpassed into eight frequency bands between 80 and 6000 Hz using sixth-order Butterworth filters. The equivalent rectangular bandwidth scale (Glasberg & Moore 1990) was used to allocate the eight channels within the specified bandwidth. This filter spacing has also been used by Qin and Oxenham (2006) and is shown in Table 1. The envelope of the signal was extracted by full-wave rectification and low-pass (LP) filtering using a second-order Butterworth filter (400 Hz cutoff). Sinusoids were generated with amplitudes equal to the root-mean-square energy of the envelopes (computed every 4 msecs) and frequencies equal to the center frequencies of the bandpass filters. The sinusoids of each band were finally summed up, and the level of the synthesized speech segment was adjusted to have the same root-mean-square value as the original speech segment.

TABLE 1. Filter cutoff (−3 dB) frequencies used for the V and LP + V processing conditions

	V		LP + V	
Channel	Low (Hz)	High (Hz)	Low (Hz)	High (Hz)
1	80	221	Unprocessed	
2	221	426	(80–600)	
3	426	724		
4	724	1158	724	1158
5	1158	1790	1158	1790
6	1790	2710	1790	2710
7	2710	4050	2710	4050
8	4050	6000	4050	6000

LP, low-pass filtered speech alone; V, vocoded speech alone.

The second processing condition simulated the acoustic stimulation alone. The signal was LP filtered to 600 Hz using a sixth-order Butterworth filter. The 600-Hz cutoff was chosen because it closely mimicked the situation with EAS patients who had residual hearing up to approximately 500 to 750 Hz and precipitous hearing loss thereafter (von Ilberg et al. 1999; Turner et al. 2004; Kiefer et al. 2005; Gantz et al. 2006). The third processing condition simulated the combined EAS. To simulate the effects of EAS with residual hearing <600 Hz, we combined the LP stimulus from condition 2 with the upper five channels of the eight-channel vocoder from condition 1, as shown in Table 1.

The remaining three processing conditions investigated the impact to speech recognition of having access to clean vocoded portion of the obstruent consonants alone, clean acoustic portion alone, or both. The sonorant segments (e.g., vowels) were left corrupted by either the SSN or 2-talker maskers. In the fourth processing condition, the acoustic portion of the obstruent consonants was left corrupted, whereas the vocoded portion was clean. In the fifth processing condition, the (low-frequency) acoustic portion of the obstruent consonants was clean, whereas the vocoded portion was left corrupted. Finally, in the sixth processing condition, both the acoustic and the vocoded portions of the obstruent consonant spectra were clean. We will refer to the above-mentioned six processing conditions as: (1) vocoded speech alone (V), (2) LP filtered speech alone, (3) combined LP and vocoded speech (LP + V), (4) LP + V with corrupted LP portion but clean vocoded portion in weak consonants (LP + Vc), (5) LP + V with clean

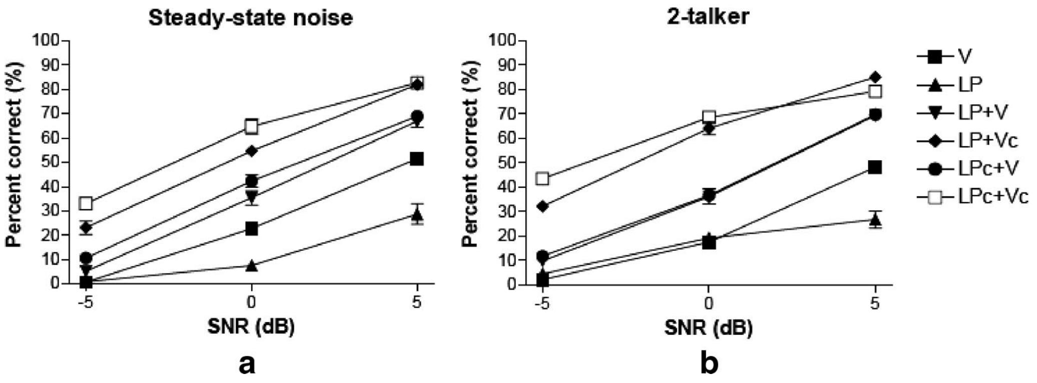

Fig. 1. Mean speech recognition scores (percent of words identified correctly) as a function of signal to noise ratio level for two maskers: (a) steady-state noise and (b) 2-talker. The error bars denote ± 1 SE of the mean.

LP portion but corrupted vocoded portion in weak consonants (LPc + V), and (6) LP + V with clean LP portion and clean vocoded portion in weak consonants (LPc + Vc).

Procedure • The experiment was performed in a sound-proof room (Acoustic Systems) using a PC connected to a Tucker-Davis system 3. Stimuli were played to listeners monaurally through a Sennheiser HD 250 Linear II circumaural headphone at a comfortable listening level. Before the test, each subject participated in a 10-min training session to listen to a set of V and LP + V stimuli and familiarize themselves with the testing procedure. At the training session, the subject selected the ear (left or right) they were comfortable when listening to the stimuli, and then used the selected ear to complete the whole test. During the test, the subjects were asked to write down all the words they heard. Each subject participated in a total of 36 conditions (= two maskers × three SNR levels × six algorithms). Two lists of IEEE sentences (20 sentences) were used per condition, and none of the lists were repeated across the conditions. The order of the test conditions was randomized across subjects. Subjects were given a 5-min break every 30 mins during the test.

Results

The mean scores for all conditions are shown in Figure 1. Performance was measured in terms of percent of words identified correctly (all words were scored). For the steady-state masker (SSN) conditions, two-way analysis of variance (ANOVA) (with repeated measures) indicated significant effect ($F[2,12] = 134.0$, $p < 0.0005$) of SNR level, significant effect of processing ($F[5,30] = 81.9$, $p < 0.0005$) of obstruent consonants, and significant interaction ($F[10,60] = 2.8$, $p = 0.006$). For the two-talker masker conditions, two-way ANOVA (with repeated measures) indicated significant effect ($F[2,12] = 146.5$, $p < 0.0005$) of SNR level, significant effect of processing ($F[5,30] = 84.5$, $p < 0.0005$) of obstruent consonants, and significant interaction ($F[10,60] = 5.7$, $p < 0.0005$).

Significant improvements in intelligibility were noted in steady and two-talker masker conditions when the listeners had access to the clean obstruent consonants in both the acoustic and vocoded portions of the spectrum (LPc + Vc stimuli). In the steady-state masker conditions, improvement with the LPc + Vc stimuli relative to the LP + V stimuli ranged from 15 percentage points at 5 dB SNR to 30 percentage points at −5 and 0 dB SNR levels. Post hoc tests, according to Fisher's Least Significant Difference (LSD) test, revealed that the scores with LPc + Vc stimuli were significantly ($p < 0.005$) higher than the scores obtained with LP + V stimuli at −5 and 0 dB SNR levels, but not at the 5 dB SNR level ($p = 0.101$). A similar pattern was observed in the two-talker conditions. Scores obtained with LPc + Vc stimuli were significantly ($p < 0.005$) higher (by about 32 to 34 percentage points) than scores obtained with LP + V stimuli at −5 and 0 dB SNR levels, but not at 5 dB SNR level ($p = 0.199$).

At extremely low SNR levels (−5 dB), scores obtained with LPc + Vc in two-talker conditions were higher (by 10 percentage points) than corresponding scores in SSN conditions. This outcome suggests a trend for masking release, at least for extremely low SNR levels, and extends the results reported in Li and Loizou (2009) with vocoded stimuli. Results from the study by Li and Loizou indicated that NH listeners performed better with fluctuating maskers than with steady noise when the listeners had access to the clean obstruent consonants, even when speech was vocoded into six channels. This outcome was interpreted to suggest 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. The finding that masking release was largest at low SNR levels is consistent with other studies (Bernstein & Grant 2009; Oxenham & Simonson 2009).

There was a clear advantage and benefit when the clean obstruent consonants were introduced amid otherwise corrupted (noise masked) voiced segments. It was not clear, however, whether it was access to the low-frequency (<600 Hz) acoustic portion of the consonant spectrum or access to the vocoded portion (>600 Hz) of the consonant spectrum that contributed the most. The data in Figure 1 indicate that it was access to the vocoded portion of the consonant spectra, rather than access to the low-frequency acoustic portion of the consonant spectra that contributed the most to the large improvements in performance obtained with the LPc + Vc stimuli. Post hoc tests confirmed that performance with the LP + Vc stimuli was significantly higher ($p < 0.05$) than performance with the LP + V stimuli at −5 and 0 dB SNR levels in the SSN masker conditions, and significantly higher at all SNR levels in the two-talker masker conditions. This finding is not surprising because most of the energy of

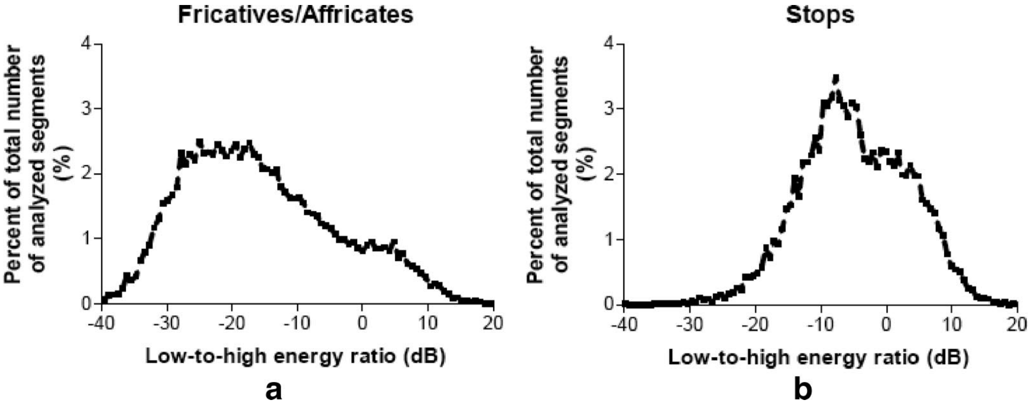

Fig. 2. Histograms of low-to-high energy ratios (dB) for (a) fricatives/affricates and (b) stops.

obstruent consonant spectra lies in the high frequencies (>1.5 kHz). Fricatives such as /s/, for instance, have much of their energy concentrated above 4 kHz (Manrique & Masone 1981).

The outcomes from this experiment imply that much of the degradation in performance of CI users in noisy conditions can be attributed to the fact that the weak consonants (e.g., /f/) are masked by noise, making it extremely difficult for them to identify these consonants (and fuse them with other phonetic segments) in the noisy speech stream. The high-frequency portion of the weak consonants' spectra is likely to be masked to a larger degree than the low-frequency portion of the spectrum. This was supported by the data in Figure 1 showing that larger improvements in intelligibility were obtained when listeners had access to the clean vocoded portion (>600 Hz) of the consonant spectra. Access to the low-frequency portion of the consonant spectra is also important because it provides important acoustic landmarks, necessary for detecting syllable/word onsets (Stevens 2002; Li & Loizou 2008b). The contribution of the low-frequency (<600 Hz) acoustic portion of the consonant spectra alone is investigated in the next experiment.

EXPERIMENT 2: EFFECT OF SELECTIVE ATTENUATION OF OBSTRUENT CONSONANTS

The significant contribution of information carried by obstruent consonants on speech recognition in simulated acoustic-electric hearing was demonstrated in experiment 1. The subjects in experiment 1 had access to the clean obstruent consonants, hence they were provided with both clean acoustic landmark information (e.g., stops' closures) and clean consonant (sonorant segments were left corrupted) information. In the present experiment, we assess the contribution of having access to acoustic landmark information alone, while limiting the access to the clean consonant information in the vocoded portion of the spectrum. This experiment is thus designed to assess the contribution of acoustic landmarks alone, such as those evident in spectral discontinuities associated with consonant closures and releases.

Methods

Subjects and stimuli • Six new NH listeners participated in this experiment. All subjects were native speakers of American English and were paid for their participations. The subjects' age ranged from 18 to 39 yrs, with majority being undergraduate students from The University of Texas at Dallas. The same test sentences (IEEE 1969) were used as in experiment 1 and were corrupted by the SSN and 2-talker maskers at 0 and 5 dB SNR levels.

Signal processing • The stimuli were presented in six different processing conditions. The first and second processing conditions were vocoded speech alone (V) and combined LP and vocoded speech (LP + V). The third was LP + V with clean LP portion and clean vocoded portion in weak consonants (LPc + Vc). These three conditions were the same as those used in experiment 1 and are used here as control conditions for comparison.

To create stimuli with relatively accurate acoustic landmark information, we selectively attenuated the low-frequency portion of the consonant stimuli while leaving the vocoded portion corrupted. The assumption made here is that for the majority of the obstruent consonants there exists little energy in the low-frequency portion of the spectrum, with most of the energy concentrated in the higher frequencies. Hence, by attenuating the low-frequency portion, we provide to the listeners acoustic landmark information without providing them with information about the consonant spectra in the high-frequency (vocoded) portion of the spectra. The vocoded consonant spectra (residing in region above 600 Hz) are left noise corrupted. Having access to the hand-labeled voiced/unvoiced boundaries (Loizou 2007), we further identified the weak consonants as stops or fricatives and applied different attenuation scaling factors for stops and fricatives in the LP portion. The motivation for using different attenuation is given in Figure 2 showing the histograms of low-to-high energy ratios* for stops and fricatives/affricates. As can be seen from Figure 2, the low-to-high energy ratio of fricatives is smaller (by approximately 10 dB) than that of

*The low-to-high energy ratio was computed as the ratio of the consonant's energy in the low-frequency portion (80 Hz<f<600 Hz) to the high frequency portion (600 Hz<f<6000 Hz) of the spectrum. For the analysis, we used consonant segments extracted from the TIMIT corpus (Garofolo et al. 1993) based on the phonetic transcriptions available in the TIMIT database. The seven selected stops included /t, p, b, d, g, k, ʔ/, and the 10 fricatives and affricates included /s, v, ʃ, z, ʒ, f, θ, ð, dʒ , tʃ/. At least 100 segments were extracted for each consonant from a total of 678 TIMIT sentences. Based on the sampling frequency (16 kHz) of the TIMIT sentences, the consonants' spectra were computed using a 256-point FFT and a 16-ms window. The low-to-high energy ratios of all stops and fricative/affricate segments were accumulated and plotted as a histogram of ratios expressed in dB. Figure 2 shows the histogram of low-to-high energy ratios of stops and fricatives/affricates.

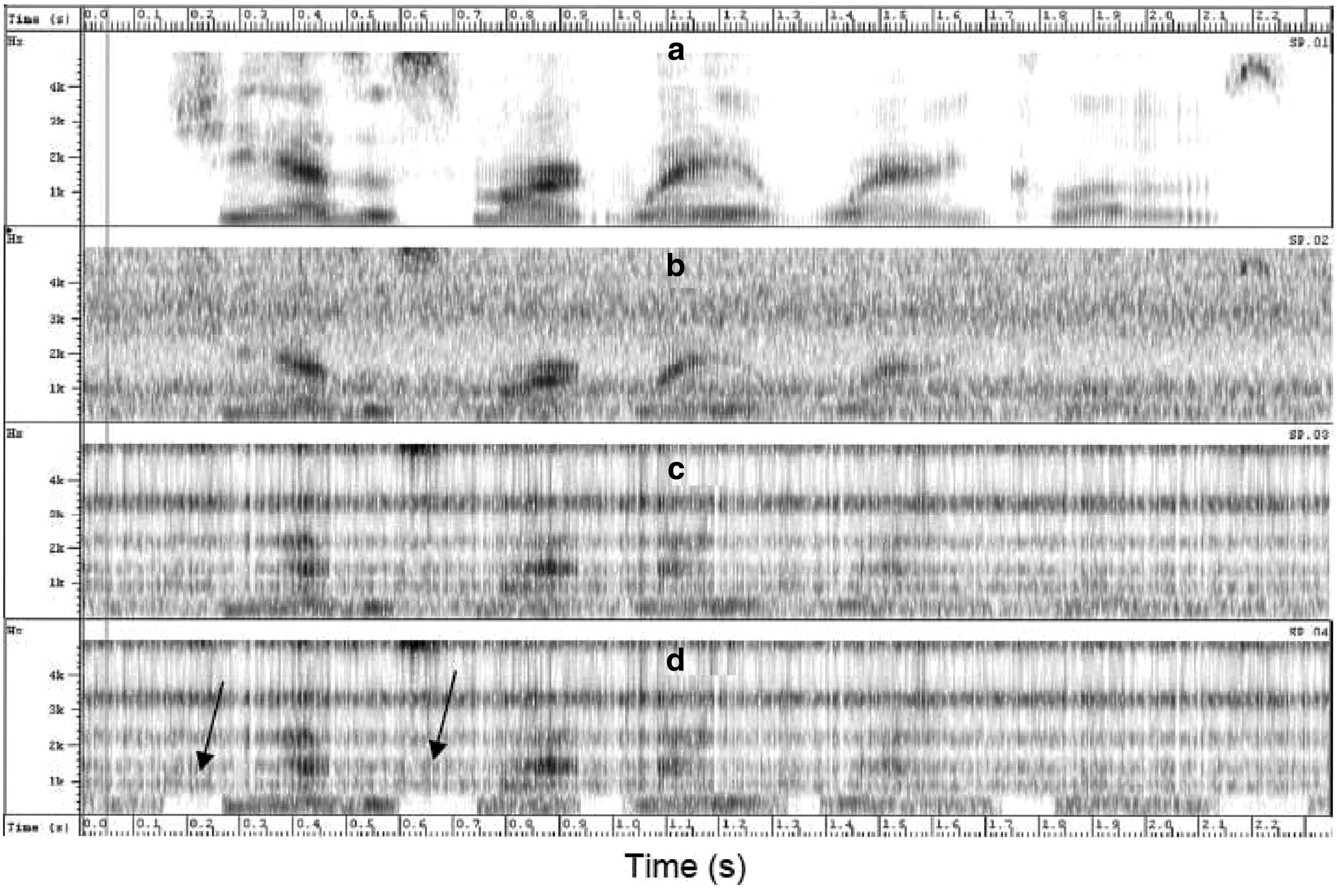

Fig. 3. Spectrograms of a sentence corrupted in 0 dB SSN and processed in the LP + V (S: 0.01, F: 0.1) condition. Panels a, b, and c show spectrograms of the clean sentence, the corrupted sentence in 0 dB SSN, and the LP + V stimulus, respectively. Panel d shows the spectrogram of the LP + V (S: 0.01, F: 0.1) stimulus along with two example acoustic landmarks (indicated by the arrows) signifying the onset/offset of weak consonants.

stops, suggesting a more aggressive attenuation of the low-frequency (<600 Hz) portion of fricatives/affricates. Based on the data from the histogram in Figure 2, the fourth processing condition multiplied the envelopes of the stops and fricatives in the LP portion by 0.1 and 0.01, respectively, corresponding to −20 and −40 dB attenuation of the envelopes. The envelopes of the stops and fricatives in the vocoded portion of the spectrum (>600 Hz) remained unchanged. Similarly, the fifth processing condition used the value of 0.01 to attenuate the stops' envelopes in the LP portion and the value of 0.1 for the fricatives. The sixth processing condition applied the same value, 0.01, for both stops and fricatives in the LP portion. The above-mentioned three processing conditions will be referred as: LP + V (S: 0.1, F: 0.01), LP + V (S: 0.01, F: 0.1), and LP + V (S, F: 0.01), where S indicates stops and F indicates fricatives/affricates. Figure 3 shows example spectrograms of a sentence corrupted in 0 dB SSN and processed in the LP + V (S: 0.01, F: 0.1) condition. As shown, the acoustic landmarks signifying the onset/offset of weak consonants are more evident in the LP + V (S: 0.01, F: 0.1) stimulus (panel d) than in the LP + V stimulus (panel c).

Procedure • The procedure was identical to that used in experiment 1. Each subject participated in a total of 24 conditions (= two maskers × two SNR levels × six algorithms). Two lists of sentences (or 20 sentences) were used per condition, and none of the lists were repeated across the conditions. The order of the test conditions was randomized across subjects.

Results

The mean scores for all conditions are shown in Figure 4. Performance was measured in terms of percent of words identified correctly (all words were scored). For comparative purposes, the scores obtained in the V and LPc + Vc conditions are also included. For the steady-state masker (SSN) conditions, two-way ANOVA (with repeated measures) indicated significant effect ($F[1,5] = 80.1$, $p < 0.0005$) of SNR level, significant effect of processing ($F[4,20] = 31.2$, $p < 0.0005$) the LP portion of the obstruent consonant spectra, and significant interaction ($F[4,20] = 4.7$, $p = 0.008$). For the two-talker masker conditions, two-way ANOVA (with repeated measures) indicated significant effect ($F[1,5] = 320.3$, $p < 0.0005$) of SNR level, significant effect of processing ($F[4,20] = 20.9$, $p < 0.0005$) of obstruent consonants, and nonsignificant interaction ($F[4,20] = 2.5$, $p = 0.077$).

As shown in Figure 4, small improvement (14 percentage points) in performance was obtained at 0 dB SNR (SSN masker) with stimuli processed in the LP + V (S: 0.01, F: 0.1) condition relative to the LP + V processed stimuli. Post hoc tests, according to Fisher's LSD, confirmed that the difference was statistically significant ($p = 0.015$). The difference be-

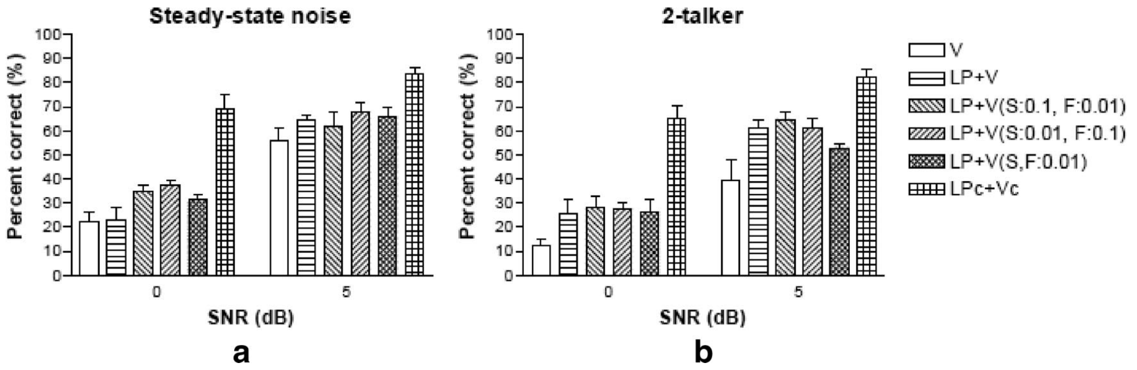

Fig. 4. Mean speech recognition scores (percent of words identified correctly) as a function of signal to noise ratio level for (a) a steady-state noise masker and (b) 2-talker masker. The error bars denote ± 1 SE of the mean.

tween the scores obtained with LP + V processed stimuli and LP + V (S: 0.1, F: 0.01) stimuli was also found to be statistically significant ($p = 0.043$). The scores obtained with the LP + V (S, F: 0.01) stimuli did not differ significantly ($p = 0.136$) from the LP + V scores. This outcome suggests that the use of different attenuations for stops and fricatives is necessary to receive significant improvements in intelligibility. The differential attenuation was found to be necessary given the difference between the stops and fricatives in the values of low-to-high energy ratios (see Fig. 2).

Post hoc tests (Fisher's LSD) assessing the difference in scores between the LP + V stimuli in the 2-talker masker conditions, and other processing conditions involving different forms of attenuation of the LP portion of the stimuli, indicated nonsignificant ($p > 0.05$) differences. We take the absence of improvement in the 2-talker conditions with the introduction of acoustic landmarks to suggest that the acoustic landmarks were already evident. Hence, selectively attenuating the LP portion of the obstruent consonants did not provide a clearer picture of the already existing landmarks. In contrast, as shown in the spectrograms in Figure 3, the consonant acoustic landmarks are absent in SSN conditions.

The finding that the introduction of low-frequency acoustic landmarks can provide a small (about 10 percentage points), but statistically significant improvements in intelligibility is consistent with the outcome of our previous study with nonvocoded stimuli (Li & Loizou 2008b). This study assessed the importance of providing partial information, within a frequency region, of the obstruent-consonant spectra while leaving the remaining spectral region unaltered (i.e., noise corrupted). Access to the low-frequency (0–1000 Hz) region of the clean obstruent consonant spectra was found to be sufficient to realize significant improvements (20 percentage points) in performance and that was attributed to improvement in transmission of voicing information. Taken together, the outcomes from the present and the study by Li and Loizou (2008b) suggest that much of the improvement in performance must be due to the enhanced access to acoustic landmarks. These landmarks, often blurred in noisy conditions, are critically important for understanding speech in noise (particularly at low SNR levels) for better determination of the syllable structure and word boundaries (Stevens 2002).

In the context of CIs, this study suggests that EAS patients can potentially benefit from signal processing techniques that can make the obstruent-consonant landmarks more evident. We postulate that the landmarks are not evident, or perceptible, to CI patients owing to the fact that these landmarks are smeared by envelope compression and also because they are easily masked by background noise. As argued in the study by Li and Loizou (2009), envelope compression smears the acoustic landmarks a great deal (more so in background noise) making it extremely difficult for CI users to identify word boundaries. Methods suggested for addressing the envelope compression effects are presented in the study by Li and Loizou. Poor spectral resolution, as afforded by current CI devices, further exacerbates the situation because it reduces 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.

In terms of practical implementation, one can devise a speech-coding strategy that first identifies the presence of consonant landmarks, and then either applies the necessary attenuation to the low-frequency portion of the spectrum or enhances ("cleans") the high-frequency portion of the spectrum. Enhancing or "cleaning," however, the high-frequency portion of the spectrum is relatively more challenging given that the high-frequency region is masked more easily by background noise than the low-frequency region (Parikh & Loizou 2005; Li & Loizou 2008a). Hence, attenuating the low-frequency portion of the spectrum of detected weak consonants would be a more reasonable and more reliable speech-coding strategy. Landmark-detection algorithms do exist in literatures and are capable of identifying the boundaries of voiced/unvoiced segments with a satisfactory high accuracy (Liu 1996; Salomon et al. 2005; Jayan & Pandey 2008; Junega & Espy-Wilson 2008), at least in quiet. Further research is warranted to extend and perhaps redesign some of the existing landmark-detection algorithms to perform well in noisy conditions.

AN OBJECTIVE MEASURE FOR PREDICTING INTELLIGIBILITY OF VOCODED SPEECH

As a secondary goal of this study, we investigated the performance of an existing objective measure, originally designed to assess speech quality, in terms of predicting the intelligibility of vocoded speech. Although a number of speech intelligibility indices exist, such as the articulation index (Kryter 1962; ANSI 1997) and speech transmission index (STI) (Steeneken & Houtgast 1980), none of these indices have been evaluated with vocoded speech. Goldsworthy and Greenberg (2004) suggested that the STI computation can potentially be customized to match a particular CI speech processor by matching the frequency bands and method of envelope calculation. They further suggested that an alternate mapping from STI to percent correct scores may be required for vocoded speech. No correlations, however, were reported in their study with vocoded speech. This study takes the first step in examining the correlation of an objective measure with vocoded speech. This measure is currently an established standard by the International Telecommunication Union (ITU) for assessment of speech quality (ITU-T 2000).

Description of PESQ Measure

The PESQ measure was originally designed to predict speech quality (ITU-T 2000; Rix et al. 2001). Nonetheless, a moderately high correlation was found by Ma et al. (2009) between PESQ values and speech intelligibility scores. Briefly, the PESQ measure is computed as follows. The original (clean) and degraded signals are first level equalized to a standard listening level, and filtered by a filter with response similar to a standard telephone handset. The signals are aligned in time to correct for any time delays, and then processed through an auditory transform to obtain the loudness spectra. The absolute difference between the degraded and original loudness spectra is used as a measure of audible error in the next stage of PESQ computation. Note that unlike most objective measures that treat positive and negative loudness differences the same (by squaring the difference), the PESQ measure treats these differences differently. This is because positive and negative loudness differences affect the perceived quality differently. A positive difference would indicate that a component, such as noise, has been added to the spectrum, whereas a negative difference would indicate that a spectral component has been omitted or heavily attenuated. Compared with additive components, the omitted components are not as easily perceived due to masking effects, leading to a less objectionable form of distortion. Consequently, different weights are applied to the positive and negative differences. The differences, termed the disturbances, between the loudness spectra are computed and averaged over time and frequency to produce the prediction of speech quality. The final PESQ score is computed as a linear combination of the average disturbance value (d_{sym}) and the average asymmetrical disturbance value (d_{sym}) as follows:

$$\text{PESQ} = a_0 + a_1 \cdot d_{\text{sym}} + a_2 \cdot d_{\text{asym}} \qquad (1)$$

where $a_0 = 4.5$, $a_1 = -0.1$, and $a_2 = -0.0309$. The range of the PESQ score is -0.5 to 4.5, although for most cases, the output range will be a score between 1.0 and 4.5. High correlations ($r > 0.92$) with subjective listening tests were reported by Rix et al. (2001) using the above PESQ measure for a large number of testing conditions taken from mobile, fixed, and voice over internet protocol applications. More details regarding the PESQ computation, along with its MATLAB implementation, can be found in Loizou (2007) and ITU-T (2000).

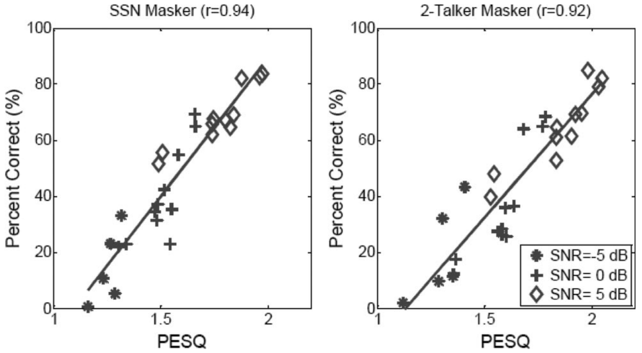

Fig. 5. Scatter plots of intelligibility scores versus predicted Perceptual Evaluation of Speech Quality (PESQ) scores for the steady-state noise and 2-talker maskers. The data were collected from experiments 1 and 2.

Results

The speech recognition scores collected from all conditions in experiments 1 and 2 were used to assess the predictive power of the PESQ measure. More specifically, the average speech recognition scores (or intelligibility) for each condition were used in conjunction with the corresponding average PESQ scores in the correlation analysis. In all, there were a total of 27 pairs (15 from experiment 1 [three SNR levels × five conditions] and 12 from experiment 2 [two SNR levels × six conditions]) of intelligibility and PESQ scores for each masker tested. Figure 5 shows separately the scatter plots of intelligibility scores versus predicted PESQ scores for the SSN and 2-talker maskers. High correlations ($r = 0.92$ to 0.94) were obtained in both masker conditions for vocoded and LP + V processed speech. The resulting standard error of the prediction (SEP) was 8.37% and 9.58%, respectively, for SSN and 2-talker masker conditions. The correlation of PESQ with vocoded speech alone (i.e., excluding LP + V processed speech) was also high, $r = 0.99$ (SEP = 2.49%) for the SSN conditions and $r = 0.94$ (SEP = 6.72%) for the 2-talker conditions. The clustering of the PESQ scores in the low range (1 to 2) was expected, given the poor quality of vocoded speech and the low-SNR levels examined. Higher PESQ scores are expected for speech vocoded at higher SNR levels.

The high correlations obtained with the PESQ measure were surprising at first, given that this measure assesses overall loudness differences between the input (clean) and processed speech signals, and as such it is more appropriate for predicting subjective quality ratings (Bladon & Lindblom 1981) than intelligibility. The PESQ measure has been shown in Hu and Loizou (2008) to correlate well ($r = 0.81$) with subjective ratings of speech distortion introduced by noise suppression algorithms. Hence, in this regard, it is reasonable to expect that a measure that predicts reliably speech distortion (and overall quality) should also be suitable for assessing speech intelligibility. This is based on the premise (and expectation) that the distortion often introduced by noise suppression algorithms (e.g., spectral attenuation near formant regions) and imparted on the speech signal should degrade speech intelligibility. Indeed, the intelligibility study by Hu and Loizou (2007) showed that some noise-suppression algorithms can degrade speech intelligibility in noisy conditions. The high correlations obtained with the PESQ measure with the intelligibility of vocoded speech suggest that this measure also captures speech distortions present in vocoded speech.

CONCLUSIONS

This study assessed the contribution of consonant landmarks to speech intelligibility in simulated acoustic-electric hearing. On the basis of the data collected from experiments 1 and 2, we can draw the following conclusions:

1. Providing access to the clean obstruent spectra yielded substantial improvements in intelligibility relative to the LP + V condition. Performance improved by as many as 30 percentage points in both masker conditions, particularly at low SNR levels (−5 and 0 dB).
2. Masking release was observed for low SNR levels (−5 dB) when listeners had access to the clean obstruent spectra. This was consistent with our previous study on vocoded speech (Li & Loizou 2009). 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.
3. Data analysis revealed that the improvement obtained with LPc + Vc stimuli can be attributed mostly to having access to clean consonants in the vocoded portion of the spectrum (>600 Hz) rather than to having access to the acoustic portion (<600 Hz). This is explained by the fact that most of the energy of obstruent consonants resides at high frequencies (>1.5 kHz).
4. Analyses in experiment 2 revealed that having access to low-frequency (<600 Hz) consonant landmark information alone can provide significant improvements in intelligibility, at least at low SNR levels and in the SSN conditions. This improvement was small, yet statistically significant, and is consistent with our previous study with nonvocoded speech (Li & Loizou 2008b). The lack of improvement at higher SNR levels can be attributed to the fact that at higher SNR levels, listeners make use of other, perhaps more salient, cues including clearer consonant landmarks.
5. The PESQ measure was found to predict well ($r = 0.92$ to 0.94) the intelligibility of vocoded and LP + V processed speech.

ACKNOWLEDGMENTS

The authors thank the Associate Editor, Dr. Gail Donaldson, and the anonymous reviewers for their valuable comments.

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

Address for correspondence: Philipos C. Loizou, Department of Electrical Engineering University of Texas at Dallas, PO Box 830688, EC 33, Richardson, TX 75083-0688. E-mail: loizou@utdallas.edu.

Received July 9, 2009; accepted October 10, 2009.

REFERENCES

ANSI (1997). *S3.5–1997 Methods for Calculation of the Speech Intelligibility Index*. New York: American National Standards Institute.

Bernstein, J., & Grant, K. (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.

Bladon, R., & Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *J Acoust Soc Am*, *69*, 1414–1422.

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

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

Dorman, M., Spahr, A., Loizou, P., et al. (2005). Acoustic simulations of combined electric and acoustic hearing (EAS). *Ear Hear*, *26*, 371–380.

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

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

Garofolo, J., Lamel, L., Fisher, W., et al. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia: Linguistic Data Consortium.

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

Goldsworthy, R. L., & Greenberg, J. E. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *J Acoust Soc Am*, *116*, 3679–3689.

Hu, Y., & Loizou, P. (2007). A comparative intelligibility study of single-microphone noise reduction algorithms. *J Acoust Soc Am*, *122*, 1777–1786.

Hu, Y., & Loizou, P. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Trans Speech Audio Process*, *16*, 229–238.

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

ITU-T (2000). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codes. ITU-T Recommendation P. 862.

Jayan, A. R., & Pandey, P. C. (2008). Automated detection of speech landmarks using Gaussian mixture modeling. In Proceedings of the International Symposium on Frontiers of Research on Speech and Music (FRSM). Kolkata, India, (pp 323–327).

Junega, A., & 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.

Kiefer, J., Pok, M., Adunka, O., et al. (2005). Combined electric and acoustic stimulation of the auditory system: Results of a clinical study. *Audiol Neurootol*, *10*, 134–144.

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

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

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *J Acoust Soc Am*, *34*, 1689–1697.

Li, N., & Loizou, P. C. (2008a). A glimpsing account for the benefit of simulated combined acoustic and electric hearing. *J Acoust Soc Am*, *123*, 2287–2294.

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

Li, N., & Loizou, P. C. (2009). Factors affecting masking release in cochlear implant vocoded speech. *J Acoust Soc Am*, *126*, 338–348.

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*. Boca Raton, FL: CRC Press.

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

Ma, J., Hu, Y., & Loizou, P. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J Acoust Soc Am*, *125*, 3387–3405.

Manrique, A., & Massone, M. (1981). Acoustic analysis and perception of Spanish fricative consonants. *J Acoust Soc Am*, *69*, 1145–1153.

Oxenham, A. J., & 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.

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

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

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

Rix, A., Beerends, J., Hollier, M., et al. (2001). Perceptual evaluation of speech quality (PESQ)–A new method for speech quality assessment of telephone networks and codecs. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA, *2*, 749–752.

Salomon, A., Espy-Wilson, C. Y., & Deshmukh, O. (2005). Detection of speech landmarks: Use of temporal information. *J Acoust Soc Am*, *115*, 1296–1305.

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech transmission quality. *J Acoust Soc Am*, *67*, 318–326.

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

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

von Ilberg, C., Kiefer, C., Tillein, J., et al. (1999). Electric-acoustic stimulation of the auditory system. *ORL J Otorhinolaryngol Relat Spec*, *61*, 334–340.