Comparison of two channel selection criteria for noise suppression in cochlear implants

Oldooz Hazrati^{a)} and Philipos C. Loizou

Department of Electrical Engineering, The University of Texas at Dallas, Richardson, Texas 75080

(Received 20 August 2012; revised 2 January 2013; accepted 7 January 2013)

The performance of current channel selection criteria used in cochlear implant (CI) devices (e.g., maximum selection criterion used in ACE) degrades significantly in the presence of noise. In noisy backgrounds, coding strategies that select the “right” channels for stimulation could potentially produce substantial improvements in intelligibility. In this study, the performance of two alternative channel selection criteria is assessed in terms of intelligibility and subjective quality with CI users in noise. The performance is compared against that of the maximum selection scheme employed in the ACE strategy (comparison is also made with the CIS strategy). Sentences were presented to seven CI users in speech weighted noise (−5, 0, and 5 dB SNR). Both channel selection criteria were implemented under ideal conditions where *a priori* knowledge of the target and masker was assumed. This was done to assess the full potential benefit of selecting the “right” channels for stimulation in noisy backgrounds. Substantial intelligibility improvement relative to the CI users’ daily strategy (i.e., ACE or CIS) was achieved with the two different channel selection criteria under all noisy conditions considered. No significant difference in subjective quality of noisy speech processed by the two channel selection criteria was observed.

© 2013 Acoustical Society of America. [http://dx.doi.org/10.1121/1.4788999]

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

Pages: 1615–1624

I. INTRODUCTION

Most contemporary cochlear implant (CI) speech coding strategies, including advanced combination encoder or ACE (Vandali *et al.*, 2000) and continuous interleaved sampling or CIS, are based on channel vocoder principles (Dudley, 1939; Peterson and Cooper, 1957; Loizou, 2006). The signal is decomposed into a small number of bands by either a bank of bandpass filters or the fast Fourier transform (FFT), and the envelope of each band is extracted. The envelopes are used to modulate biphasic pulses, which are in turn sent to the electrodes for stimulation. The number of electrodes selected for stimulation at each cycle differs between ACE and CIS strategies. The ACE strategy selects a subset of envelopes with highest energy for stimulation (8–12 of 22). However, in the CIS strategy a fixed number of envelopes (e.g., 16) are computed and used for stimulation.

Although the ACE strategy works well in quiet as it captures the formant peaks, its performance (along with CIS) degrades significantly when noise or reverberation is present (e.g., Kokkinakis *et al.*, 2011; Hu and Loizou, 2008). Because the ACE strategy is based on the maxima selection criterion, problems arise when additive noise or reverberation is present. In segments where the masker is dominant, for instance, the masker-dominant channels are selected thus extracting potentially the wrong information from the mixture envelopes. This problem often occurs in low-energy segments (e.g., weak consonants) where the signal is heavily masked. Furthermore, as discussed in Hu and Loizou (2008), the maxima selection criterion, when operating in back-

ground noise, might potentially create perceptual confusions between the flat-spectra consonants (e.g., /f/, /θ/, /v/) and the rising-tilt or high-frequency spectra consonants (e.g., /s/, /t/, /d/). When reverberation is present and during the unvoiced segments (e.g., stops) of the utterance, the ACE strategy may mistakenly select the channels containing reverberant energy, since those channels have the highest energy (Kokkinakis *et al.*, 2011). Unlike the ACE strategy, the CIS strategy retains the information in all channels leaving the task of selecting the channels carrying information about the target to the CI user. It is clear from the preceding studies that a better, and perhaps more appropriate, selection criterion is needed when background interference is present.

A channel selection scheme based on a psychoacoustic model (similar to that adopted in MP3, e.g., see ISO/IEC, 1999) was proposed by Noguiera *et al.* (2005). Their rationale was that amplitudes falling below the masking threshold ought to be discarded as inaudible. The proposed channel selection scheme, when compared against the ACE strategy, resulted in large improvement in speech-shaped noise (SSN) at 15 dB SNR when only four channels were retained in each stimulation cycle. No significant improvement in intelligibility was observed, however, when the number of selected channels was increased to eight.

In this study, we evaluate and compare the full potential of two alternative channel selection schemes in terms of intelligibility and subjective quality. The first channel selection scheme was explored and evaluated in our prior study with CI users (Hu and Loizou, 2008) and is used in the present study for comparative purposes. This selection criterion is based on the signal-to-noise ratio (SNR) of individual channels and selects target-dominated (SNR > 0 dB) channels while discarding masker-dominated (SNR < 0 dB)

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

channels. Large improvements in intelligibility were reported with the SNR channel selection criterion in ideal conditions where it was assumed that we had access to the clean and noise signals prior to mixing. This was found to be true for both CI listeners (Hu and Loizou, 2008) and normal-hearing (NH) listeners (Brungart *et al.*, 2006; Li and Loizou, 2008).

The SNR selection criterion is directly applied to the noisy (un-processed) signals and not to the signals processed by noise reduction algorithms. Therefore using it in conjunction with noise-suppression algorithms will not be helpful as it does not make use of any information contained in the noise-suppressed output envelopes, nor does it limit in any way the distortions introduced by noise-reduction algorithms. In the SNR channel selection technique, the background noise is not suppressed explicitly (i.e., no noise detector is used), yet the intelligibility of noisy speech is improved because noise-dominant channels are discarded at each stimulation cycle. Note that noise suppression algorithms attenuate low-SNR channels without discarding them. A different channel selection scheme was recently proposed to account for the fact that the distortions introduced by noise-suppression algorithms have different effects on speech intelligibility (Kim and Loizou, 2011). Very briefly, because the frequency-specific gain functions used in noise-reduction algorithms are far from perfect, amplification distortions are often introduced when the target signal envelope is overestimated, and attenuation distortions are introduced when the target signal envelope is underestimated. Listening tests with NH listeners indicated that attenuation distortions had a minimal effect on speech intelligibility while amplification distortions markedly degraded intelligibility (Kim and Loizou, 2011). Hence it is reasonable to consider retaining noise-suppressed envelopes that are smaller in amplitude than the true target envelopes, while discarding envelopes that are larger in amplitude than the target envelope. Such a criterion was evaluated with NH listeners (Kim and Loizou, 2011) and shown to yield substantial improvements in intelligibility compared to the unprocessed noisy conditions but has not been tested with CI users. It was thus not clear whether such a channel selection criterion would be beneficial to CI users who only receive a small number of channels of information. It is worth mentioning that both aforementioned channel selection criteria assume access to the unmixed target and masker signals and have been found to be effective for most, if not all, types of maskers (e.g., SSN, babble, competing-talker, etc.). In conditions where reverberation is present, a different channel selection rule, based on the signal-to-reverberant ratio (SRR), has been shown to be more appropriate (Kokkinakis *et al.*, 2011).

The performance of the preceding two channel selection criteria in noisy conditions is compared against the performance of the CI users' daily coding strategies (i.e., ACE or CIS) in terms of subjective quality and intelligibility. The aim of the present study is twofold: (1) to determine whether the use of the two selection criteria offers significant benefits in intelligibility in noise relative to the default coding strategies (ACE or CIS) and (2) to determine if either of the two selection criteria is more advantageous compared to the

other, perhaps in terms of providing better subjective quality, and examine whether the selection criteria produce any unwanted distortions compared to the stimuli processed by ACE (or CIS) strategies.

II. CHANNEL SELECTION SCHEMES

The first channel selection criterion evaluated in this study is the SNR-rule that retains envelopes corresponding to channels with local SNRs exceeding a certain threshold and discards envelopes corresponding to channels with local SNRs below the threshold. The local (instantaneous) SNR is defined as

$$\text{SNR}(f, t) = 10 \log_{10} \frac{|X(f, t)|^2}{|D(f, t)|^2}, \quad (1)$$

where t and f indicate the time frame and channel index respectively, and $X(f, t)$ and $D(f, t)$ denote the complex FFT spectra of the clean and masker signals, respectively, available prior to mixing. Envelopes (computed using the FFT magnitude spectrum of each 20-ms segment with 10-ms overlap) corresponding to channels with $\text{SNR} > T$ are selected, while envelopes corresponding to channels with $\text{SNR} \leq T$ are discarded, where T denotes a pre-set threshold value. The channel selection is implemented in the FFT domain by applying a binary gating function to the spectrum of the noisy signal as follows:

$$\hat{X}(f, t) = Y(f, t) \cdot B(f, t), \quad (2)$$

where $Y(f, t)$ and $\hat{X}(f, t)$ denote the complex FFT spectra of the noisy (mixtures) and binary-gated noisy signals, respectively. The binary gating function is defined as:

$$B(f, t) = \begin{cases} 1, & \text{SNR}(f, t) > T \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where T represents the threshold value expressed in dB. To reconstruct the enhanced signal in the time-domain, the inverse FFT of $\hat{X}(f, t)$ is computed and the signal is finally synthesized using the overlap-add (OLA) method. In our study, the threshold in Eq. (3) was set to $T = -5$ dB. Pilot data showed no difference in performance with other threshold values, ranging from -8 to 0 dB. This outcome is consistent with that first observed in the study by (Brungart *et al.*, 2006) with NH listeners. For this reason, only one threshold value ($T = -5$ dB) will be used to evaluate the SNR selection scheme.

The preceding channel-selection criterion cannot be easily incorporated in noise suppression algorithms as it does not make use of any information contained in the noise-suppressed output envelopes. For that reason, a different channel selection criterion was developed (Loizou and Kim, 2011; Kim and Loizou, 2011) and found to yield substantial improvements in speech intelligibility by NH listeners. The new channel selection criterion is based on imposing constraints on the noise-suppressed envelopes, and we thus refer to it as the constrained channel-selection rule and denote it

as the C-rule. The constraints are imposed to limit the type of distortion introduced by noise-suppression algorithms. Amplification distortions, for instance, occur when the envelope amplitude of the target signal is overestimated (by the noise-suppression algorithm) and these distortions have been found to be more detrimental to speech intelligibility than the attenuation distortions introduced when the target envelope amplitude is underestimated (Kim and Loizou, 2011). Hence taking this finding into account, it is reasonable to consider retaining noise-suppressed envelopes that are smaller in amplitude than the true target envelopes, while discarding envelopes that are larger in amplitude than the target envelope. Assuming that $\hat{X}(f, t)$ is the noise-suppressed envelope in band f obtained after Wiener filtering (see the Appendix), the channel selection is implemented in the FFT domain as follows:

$$\hat{X}_M(f, t) = \begin{cases} 1, & \frac{|\hat{X}(f, t)|}{|X(f, t)|} \leq T' \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where T' denotes the threshold value which was set to either 2 or 1 and $\hat{X}_M(f, t)$ denotes the modified spectrum containing only the channels satisfying the preceding criterion. To reconstruct the enhanced signal in the time domain, the inverse FFT of $\hat{X}_M(f, t)$ is computed, and the signal is finally synthesized using the OLA method. For comparative purposes, the Wiener filtering noise-reduction algorithm is applied to the mixtures but without applying channel selection. Note that the Wiener filtering algorithm retains all channels but attenuates the envelopes in proportion to the estimated SNR, which is accomplished according to the rule given in Eq. (A2) in the Appendix. The Wiener filtering algorithm was chosen here because it has been used before in both hearing aids (Chung, 2004) and CIs (Hu *et al.*, 2007; Dawson *et al.*, 2011) although with a different implementation.

The thresholds $T' = 1$ and 2 in Eq. (4) were chosen for the following reasons (Kim and Loizou, 2011). When $T' = 1$, only attenuation distortions are allowed because $|\hat{X}_M(f, t)| < |X(f, t)|$. Whereas, when $T' = 2$, attenuation distortions are allowed along with limited (<6.02 dB) amplification distortions (Kim and Loizou, 2011). Data from NH listeners (Kim and Loizou, 2011) indicated that performance with $T' = 2$ was in most conditions significantly better than performance with $T' = 1$, and this seemed to be dependent on the type of masker used (single-talker masker vs steady noise). Given the difference in performance with the two thresholds, we decided to include both of them in the present study.

The C-rule channel selection implementation is depicted in Fig. 1 in the context of noise suppression (Wiener filtering algorithm). As shown in this figure, the C-rule channel selection scheme takes as input the noise-suppressed output from Wiener filtering and makes decisions regarding retaining or discarding channels based on the criterion in Eq. (4). The decisions also make use of information in the input target spectrum prior to mixing. Nonetheless, in terms of what is presented to the listeners, the main difference between the

FIG. 1. Diagram of the Wiener filtering and C-rule channel selection criterion.

C-rule output and the Wiener-filtering output is the number of channels retained at each cycle. In Wiener filtering, no channel is discarded (all channels are presented), whereas the C-rule discards channels not satisfying the above-noted constraint in Eq. (4).

The channel-selection rule in Eq. (4) was found to result in significant speech intelligibility improvement by NH listeners for both steady-noise and competing-talker background conditions (Kim and Loizou, 2011) but has not been tested with CI listeners. The Wiener filtering implementation reported by Scalart and Filho (1996) is used as it is simple and amenable to real-time processing. Wiener filtering was used in the present study for noise suppression, but the choice of noise-suppression algorithm was found to have no influence on the outcome (Kim and Loizou, 2011). The Wiener-filtering algorithm took as input the mixture envelopes and produced the output envelopes $|\hat{X}_M(f, t)|$, as shown in Eq. (4) and Fig. 1. A brief description of the Wiener-filter noise suppression algorithm is provided in the Appendix (MATLAB code is available from our website¹).

Our implementation of the preceding two channel-selection rules was done in the context of the ACE/CIS coding strategy. More specifically, the channel selection (with either the SNR- or C-rule) was implemented in the FFT domain and not at the stage following power summation of each band and log-compression of the resulting envelope (referred to as the log-envelope stage) for two reasons. First, the spectral resolution at the FFT stage is far better than that at the log-envelope stage (64 channels vs 22 channels for Nucleus users or 16 channels for CIS users). Second, the band power-summation operation (as in ACE) smoothes out abrupt (on/off) across-frequency discontinuities that might be introduced by the binary gating applied to the mixture envelopes. The across-frequency smoothing subsequently eases or minimizes some temporal discontinuities that might be introduced by binary gating. This is important for avoiding any distortions that might cause degradation in speech quality. Note that the maxima selection criterion is used at the log-envelope stage, and at first glance, it might seem that two different criteria are used. In reality, however, the channel selection criterion used at the FFT stage will bias the channel selection at the envelope stage. For instance, if 22 target-dominant channels are selected (with either the SNR- or C-rule) following the power summation in each band, the eight (or 10) channels with the highest SNR will be selected at the log-envelope stage. Similarly, if some channels are noise dominated and have high level amplitude, they would have been selected by the maximum rule (ACE) but would

have been rejected by the SNR- or C-rules. Hence, the channel selection done at the FFT stage influences (biases) the channel selection at the log-envelope stage.

III. EXPERIMENT: COMPARISON OF CHANNEL SELECTION SCHEMES

A. Methods

1. Subjects

Seven adult postlingually deafened CI users (mostly bilateral) participated in this study. All participants were native speakers of American English and were paid for their participation. Their age ranged from 52 to 77 yr (with average of 65 yr). Five CI users were using Nucleus (Cochlear, Ltd) devices, and two subjects were using Advanced Bionics Corporation (ABC) devices. All participants used their devices routinely and had a minimum of 1 yr experience with their device. Detailed biographical data for the subjects are given in Table I.

2. Research processor

Three of the subjects tested were using the Nucleus Freedom device, two were using the Nucleus 5 speech processor (N5), and the other two were using the Harmony processor (ABC) on a daily basis. The five Nucleus users were temporarily fitted with the SPEAR3² device programmed with the ACE speech coding strategy, and the two ABC users were tested using their own devices (fitted with the CIS speech coding strategy). The Seed-Speak GUI application was used to program the SPEAR3 wearable research processor for the Nucleus CI users.

3. Stimuli

The IEEE sentence corpus (IEEE, 1969), available from Loizou (2007), was used. There are 720 sentences (10 sentences/list) in the IEEE corpus produced by a male speaker, and each sentence is composed of approximately 7–12 words. The root mean square (RMS) value of all sentences was equalized. The sentences were corrupted using steady-state speech-shaped noise (SSN) at –5, 0, and 5 dB SNR levels. The desired SNR was achieved by keeping the speech power constant and varying the noise intensity.

4. Implementation

Both channel-selection strategies were implemented in the offline mode meaning that the noisy signals were first

processed with the channel-selection strategies, and based on the selected channels, the speech sentences were reconstructed. Then the noise-suppressed (through channel-selection) signals were presented to the Nucleus CI users using the SPEAR3, which was programmed with the ACE strategy. The stimuli were presented to the ABC CI users in a double-wall sound proof booth where the listeners wore their own devices. All listeners (Nucleus and ABC) were tested using their every day mapping settings. No special mappings (used for noisy environments) were adopted here.

5. Procedure

Prior to testing, each subject participated in a short practice session to gain familiarity with the listening task. Four sentences per condition (excluded from the actual test) were presented to the CI users in the practice session. The stimuli were presented to the five Nucleus users unilaterally through the auxiliary input jack of the SPEAR3 processor in a double-wall sound proof booth (Acoustic Systems, Inc.). For the bilateral users, the ear with the highest sentence score in quiet was chosen for testing. During the practice session, the subjects adjusted the volume level to a comfortable level, and the volume level was fixed throughout the tests. The stimuli were presented to the two ABC users through speakers at a level of 68 dB SPL in a double-wall sound-proof booth (Acoustic Systems, Inc.). Subjects were given a 15 min break every 60 min during the test session to avoid listener fatigue. The overall test lasted for almost 4 h.

For the subjective intelligibility evaluations, the CI listeners participated in a total of 16 conditions, which included: Clean stimuli, noisy stimuli (UN) presented at three SNR levels (–5, 0, and 5 dB), Wiener processed (Wiener) stimuli at three SNR levels, stimuli processed via the SNR-rule at three SNR levels, and stimuli processed via the C-rule at three SNR levels (two threshold values). Twenty IEEE sentences (two lists) were used per condition. None of the lists used was repeated across conditions. To minimize any order effects, the order of the test conditions was randomized across subjects. During testing, the participants were instructed to repeat as many words as they could identify. The responses were scored off-line based on the number of words correctly identified. All words were scored.

For the subjective quality comparison, the AB paired-comparison test was used. Listeners were presented with pairs of the same input stimuli processed via two algorithms (A and B), and asked to indicate which, if any, of the two stimuli they prefer (see review in Loizou, 2011). The results

TABLE I. Biographical data of the CI users tested.

Subjects	Age	Gender	Years implanted (L/R)	Number of active electrodes	CI processor	Etiology of hearing loss
S1	65	F	15/8	22	Freedom	Unknown
S2	60	F	6/9	22	Freedom	Hereditary
S3	70	F	3/1	20	N5	Unknown
S4	74	M	10/3	11	Harmony	Fall
S5	57	F	-/9	16	Harmony	Unknown
S6	52	F	2/1	22	N5	Unknown
S7	77	M	6/6	22	Freedom	Hereditary

were reported in terms of percentage preference of algorithm A over algorithm B. The degree of preference was rated on a four-point scale with the rating of 0 indicating no difference (same), 1 indicating a small difference, 2 indicating a large difference, and 3 indicating a very large difference. This preference scoring was based on the comparison category rating (CCR) (ITU-T, 1996) and has been used extensively in the literature to evaluate the quality of the output of speech compression algorithms. In our study, the scores were categorized into three groups: (1) percentage preference of algorithm A (without considering the weight of preference), (2) percentage preference of algorithm B (again without considering the weight of preference), and (3) the same category (equal preference). Rating of preference on the four-point scale was not analyzed.

In the present study, CCR preference tests were used to: (1) compare the quality of stimuli processed via the SNR- and rule and (2) compare the quality of stimuli processed via either the SNR- or C-rule against the stimuli in quiet. The same sentence was processed by the two criteria and presented to the listeners twice. The listeners were asked to rate the quality of the second sentence compared to the first sentence using the previously described rating scale. The following instructions were given to the listeners prior to the test:

“You are asked to judge the quality of the sentences as to whether the sentences are pleasant, clear and free of sound artifacts or distortions. Focus on the quality of the sentences rather than on the intelligibility, i.e., the number of words you can understand.”

The listeners participated in a total of nine listening conditions, which included the AB comparison with the clean stimuli, and the comparison between the SNR- and C-rule ($T' = 2$) processed stimuli at three different SNR levels. Twenty IEEE sentences (two lists) were used per condition. None of the lists used in intelligibility tests were used in the quality rating experiment and none of the lists used was repeated across conditions. The order of the test conditions was randomized across listeners.

B. Results and discussion

1. Intelligibility comparison

The mean intelligibility scores obtained by the CI listeners in the various conditions are shown in Fig. 2. For comparative purposes, the average score obtained in the unprocessed (UN) noisy conditions is also shown. As an additional control condition, we show the mean scores obtained with the Wiener noise-reduction algorithm. Note that unlike the SNR- and C-rule implementations which required *a priori* information about the target signals, no such information was used in the implementation of the Wiener algorithm. A two-way analysis of variance (ANOVA) with repeated measures was run to assess the effects of SNR and channel selection rule (SNR-rule, C-rule with $T' = 1$, and C-rule with $T' = 2$, ACE/CIS) on speech recognition. The ANOVA indicated a significant effect [$F(3,15) = 313.8$, $p < 0.005$] of selection rule, a significant effect [$F(2,10) = 173.3$, $p < 0.005$] of SNR, and a significant interaction [$F(6,30) = 24.3$, $p < 0.005$]. The interaction was

FIG. 2. Mean percent correct scores by CI users ($n = 7$) for the three SNR conditions tested. Error bars indicate standard deviations.

caused by the fact that performance with the maximum selection criterion used in ACE dropped significantly (CIS performance dropped as well) as SNR decreased. In contrast, performance with the other channel selection rules remained high, even at low SNR levels (-5 dB). As shown in Fig. 2, performance with either the SNR- or C-rule ($T' = 2$) reached over 80% correct for all SNR conditions, including SNR $= -5$ dB.

As is evident from Fig. 2, substantial gains in speech intelligibility were obtained by the CI users in all noisy conditions with both the SNR- and C-rule channel-selection criteria compared to the subjects' daily strategies (ACE/CIS). Use of the channel-selection rules improved, for instance, the intelligibility at -5 dB SNR from near 0% correct obtained with unprocessed (UN) stimuli (ACE/CIS) to 79% correct (SNR-rule) and 86% correct (C-rule, $T' = 2$). This difference was found to be statistically significant

($p < 0.0005$, paired samples t -tests, Bonferroni protected) for all three SNR conditions. Further analysis of the data obtained with the five Nucleus subjects tested indicated that performance obtained with the maximum-selection criterion (ACE) and CIS was significantly ($p < 0.003$) lower than performance obtained using the SNR- and C-rule ($T' = 1$, $T' = 2$) channel selection criteria for all input SNRs. These outcomes are consistent with previous findings with CI users (Hu and Loizou, 2008) and NH listeners (Brungart *et al.*, 2006; Li and Loizou, 2008; Kim and Loizou, 2011).

Performance with the SNR channel-selection remained high (79% correct) even at the -5 dB SNR condition but relative to the higher SNR conditions a small decrement in performance was noted. Paired samples t -tests (with Bonferroni correction) indicated that performance (79% correct) at -5 dB SNR was significantly lower ($p < 0.005$) than performance (91% correct) at 5 dB SNR. In contrast, performance obtained using the C-rule ($T' = 2$) remained high (86%–90%) at all three SNR levels. No statistically significant difference ($p > 0.017$, Bonferroni corrected) was found between the performances obtained at the three SNR levels with the C-rule ($T' = 2$ or $T' = 1$). Of the two selection rules examined, the C-rule ($T' = 2$) yielded significantly higher ($p = 0.03$, paired-samples t -test) performance at -5 dB SNR compared to the SNR rule (no difference was noted at higher SNR levels). Hence, the C-rule proved to be a robust channel-selection rule capable of yielding consistently high performance even at the lowest SNR tested (-5 dB).

Further analysis was conducted to examine the differences between the SNR- and C-rule ($T' = 2$) channel-selection criteria. For one, it is of interest to examine the channel-selection agreement in as far as whether the two criteria select or discard the same channels. Percent agreement was computed for 20 sentences taken from the IEEE corpus. Agreement in channel selection was counted if a specific channel was either selected by both rules or discarded by both rules. For the -5 , 0, and 5 dB SNR conditions, the percent agreement were 62%, 60%, and 60%, respectively. Hence, on average, roughly two of three channels selected/discarded were identical. As a secondary measure of similarity/comparison between the two selection criteria, we computed the percentage of channels selected in each frame. This measure is similar to percentage of hits. The results are tabulated in Table II for the three SNR conditions tested. As shown in Table II, the C-rule selects on the average more channels than the SNR-rule, particularly in the lowest SNR condition tested (-5 dB). Hence the SNR-rule processed signal is sparser (i.e., fewer number of ON channels), at least in the FFT domain, than the signal produced by the C-rule

TABLE II. Mean percentage of selected channels using the C-rule criterion ($T' = 1$ and $T' = 2$) and SNR-rule in different noisy conditions. The numbers in parenthesis indicate standard deviations.

SNR level (dB)	C-rule ($T' = 1$)	C-rule ($T' = 2$)	SNR-rule
-5	38.21 (16.31)	50.91 (20.08)	13.14 (10.42)
0	44.52 (17.03)	58.94 (20.95)	19.30 (13.99)
5	50.62 (17.39)	67.23 (21.08)	27.92 (18.23)

($T' = 2$). Analysis of the electrodiagram pattern of the signal (Fig. 3) produced by the C-rule ($T' = 2$) indicated that its pattern looked more similar to that of the clean stimulus [e.g., compare (a) and (b) in Fig. 3] compared to the pattern

FIG. 3. Stimulus output patterns (electrodiagrams) of the IEEE sentence “Open your book to the first page” uttered by a male speaker. (a) Electrodiagram of clean (uncorrupted) sentence processed by the ACE strategy, (b) electrodiagram of the same sentence corrupted by -5 dB SNR SSN processed by C-rule, and (c) same sentence corrupted by -5 dB SNR SSN processed by SNR-rule. In each electrodiagram, time is shown along the abscissa and the electrode number is shown along the ordinate.

generated by the SNR-rule [compare (a) and (c)]. This is evident in segments before $t < 1$ s and for electrodes 1-6 (Fig. 3). This might have contributed to the better performance of the C-rule ($T' = 2$) at -5 dB SNR.

Performance obtained with the Wiener noise-reduction algorithm was found to be significantly higher ($p < 0.01$, paired-samples t -tests with Bonferroni correction) than the baseline performance (UN) in all three SNR conditions. An improvement of 18 percentage points was noted at 5 dB SNR. Similar improvement was reported in our prior study (Hu *et al.*, 2007) using a Wiener-like gain function and also by Dawson *et al.* (2011) using a parametric Wiener gain function. In the latter study, use of a parametric Wiener function yielded a 2.14 dB benefit in speech-reception threshold (SRT), relative to ACE, in conditions involving SSN. The results from the present and preceding studies indicate that the Wiener algorithm is an effective noise-suppression algorithm for CIs at least in stationary (steady) types of background environments.

Given the rather similar performance in terms of intelligibility between the two channel selection rules, the question arises as to whether any of the two rules (SNR- and C-rule) offers any advantage in terms of subjective quality. Furthermore, it is not clear if there is any noticeable difference between the quality of the processed files using the two criteria and the same files processed with ACE or CIS in quiet. This question is addressed next.

2. Quality comparison

The average CCR preference scores resulting from the quality comparison of clean, SNR-rule processed and C-rule processed ($T' = 2$) stimuli are given in Fig. 4 for the three SNR conditions (-5 , 0 , and 5 dB) tested. A two-way analysis of variance (ANOVA) with repeated measures was run to assess the effects of SNR level and channel selection rule (SNR-rule vs C-rule) on subjective quality. ANOVA indicated a significant effect [$F(2,12) = 1.6$, $p < 0.001$] of SNR level, a significant effect [$F(2,12) = 8.1$, $p = 0.006$] of selection rule, and a non-significant interaction [$F(4,24) = 1.4$, $p = 0.25$]. Three levels were incorporated in the ANOVA: Preference for SNR-rule, preference for C-rule, or same preference. *Post hoc* tests (Scheffe) indicated no preference ($p = 0.96$) in quality for either the C- or SNR-rule at SNR $= -5$ dB. For the remaining SNR conditions, the “same quality” rating was selected significantly more often ($p < 0.05$) than either rule. Overall, the quality comparison between the SNR-rule and C-rule processed speech revealed no distinct preference of quality for stimuli processed with either rule [Fig. 4(a)]. The lack of preference for either rule might be attributed to the fact that the two rules select or discard on the average the same channels; they do so with 60% agreement (as discussed in previous section).

Similar ANOVAs were run to assess the effects of SNR level and processing (clean stimuli vs C-rule or SNR-rule processed stimuli) on subjective quality. For the clean vs SNR-rule comparison, two-way ANOVA (with repeated measures) indicated a significant effect [$F(2,8) = 1.6$, $p < 0.001$] of SNR level, a significant effect [$F(2,8) = 4.3$, $p = 0.053$] of processing,

FIG. 4. Average percent preference of (a) C-rule processed noisy stimuli and SNR-rule processed stimuli, (b) clean stimuli and C-rule processed stimuli, and (c) clean stimuli and SNR-rule processed stimuli. Bars labeled “same” represent the same quality judgment for the two methods. Error bars indicate standard errors of the mean.

and a significant interaction [$F(4,16) = 1.4$, $p < 0.005$]. *Post hoc* tests (according to Scheffe), indicated a significantly ($p = 0.001$) higher preference for the clean stimuli over the SNR-rule processed stimuli at SNR $= -5$ dB. No significantly higher ($p > 0.05$) preference was found, however, for the clean stimuli at either SNR $= 0$ dB or SNR $= 5$ dB conditions. That is, the clean and SNR-rule processed stimuli were judged to have no significant difference in quality. Similarly, for the clean vs C-rule processed comparison, subjects preferred ($p < 0.005$) the clean stimuli at SNR $= -5$ dB but not in the other SNR conditions.

IV. DISCUSSION

The impact of two channel selection schemes on intelligibility and subjective quality of noisy speech has been

evaluated in this study with CI users. This evaluation was performed in optimal conditions where knowledge was assumed of the clean and noise signals prior to mixing. These criteria have been found to significantly improve the intelligibility of speech presented to CI users in noisy conditions (Fig. 2). Both criteria have been shown to yield significantly higher performance than the conventional maxima selection criterion used in the ACE strategy (five of the seven subjects tested in the present study were using the ACE strategy in their daily processor). The quality of noisy speech processed by the two channel selection schemes was judged not to be significantly different from each other. Furthermore, the SNR- and C-rule processed stimuli were judged not to have significantly different quality from the clean stimuli with only a weak preference for the clean stimuli at the lowest SNR condition tested (-5 dB). The SNR selection criterion offers a number of advantages over the maxima criteria when applied in noisy environments (Hu and Loizou, 2008). For one, the SNR criterion is more robust to the spectral distribution of the target and the underlying SNR in each channel. Assuming that the SNR criterion is implemented at the log-envelope stage (as done by Hu and Loizou, 2008), in cases where the target (and mixture) spectrum is flat (e.g., fricative /f/) and the channel SNRs are positive, the SNR-rule will select all channels, while the ACE strategy (maxima rule) will only select the largest eight (or 10) in amplitude. In cases where the target is completely masked by background interference, as it often occurs for instance during stop closures or weak speech segments, the SNR criterion will not select any channels (i.e., no electrical stimulation will be provided) due to the negative SNRs of all channels. In contrast, the ACE strategy will select a subset (the largest) of the channels independent of the underlying SNR. Providing no stimulation during stop closures or during low-energy segments in which the masker dominates, is important for two reasons. First, it can reduce masker-target confusions, particularly when the masker(s) is a competing voice(s) and happens to be present during speech-absent regions. Second, it can enhance access to voicing cues and reduce voicing and/or manner errors. The preceding advantages seem to be carried out with the current implementation of the SNR selection criterion.

For the majority of the conditions tested, performance with the C-rule ($T' = 2$) was high (90% correct) and equal to the performance with the SNR selection rule. At the lowest SNR tested (-5 dB), the C-rule ($T' = 2$) was found to yield statistically higher performance than the SNR-rule (see Fig. 2). We cannot exclude the possibility that higher performance might have been attained by the SNR-rule had we used a smaller value (e.g., -10 dB) for the SNR threshold. Previous studies with NH listeners indicated strong dependency of the optimal SNR threshold value on the input global SNR level (Kjems *et al.*, 2009). More specifically, use of lower threshold values was found in the study by Kjems *et al.* (2009) to be more beneficial at the lowest SNR levels (e.g., SNR = -40 dB) tested. In contrast, in the present study, the use of threshold $T' = 2$ in the C-rule was found to be robust to all three SNR conditions tested in as far as maintaining high performance in all conditions.

The quality of noisy speech processed by the two channel selection schemes was judged not to be significantly different (Fig. 4). Most importantly, the SNR- and C-rule processed stimuli were judged not to have significantly different quality as the clean stimuli with only a weak preference for the clean stimuli at the lowest SNR condition tested (-5 dB). This outcome is important as it suggests that the SNR- or C-rule channel selection processing does not result in unacceptable signal quality judgments. The most striking finding from the subjective quality comparison was the subjects' lack of strong preference in quality for the stimulus in quiet or the stimulus that has been processed via either the SNR- or C-rule selection criteria. This was surprising for two reasons. First, the SNR- and C-rule stimuli were obtained by gating the mixture envelopes [see Eq. (2)], which in some conditions were corrupted at the lowest SNR levels tested (-5 dB). It is safe to say that at -5 dB SNR, the majority of the channels were dominated by the masker. Yet gating largely masker-dominated envelopes produced speech that was judged to be of nearly the same quality as that of clean speech. Second, channel selection could cause distortions that might be present due to abrupt and unpredictable on/off switching of channels from frame to frame. The subjective quality results, however, indicated that it was not the case. We attribute this to the manner in which the channel-selection rule was implemented in our study. Ordinarily and assuming that envelopes are computed in the FFT domain (as opposed to using a filter bank), channels are selected following the summation of power of FFT spectral components falling within each of the 22 bands (Nucleus device). Alternatively, channels can be selected at the FFT domain and prior to the stage of summing the power of each band. As mentioned earlier, the advantage of applying the channel selection in the FFT domain is twofold. First, the spectral resolution is much finer in the FFT stage (64 channels for the Nucleus device) than the resolution (22 channels) after the power-summation stage. Second, the power-summation operation smoothes out some discontinuities present in the spectrum, such as those introduced by the binary gating of mixture envelopes. Such smoothing subsequently eases or minimizes some discontinuities occurring over time that might be introduced by binary gating. Hence, by performing the channel selection in the FFT domain, we reduce the possibility of frequent on/off switching distortions.

Optimal (ideal) conditions were assumed in the current study, thus enabling us to have access to the true instantaneous SNR needed for the implementation of the SNR selection rule (same for the C-rule). In such ideal scenarios, however, the channel selection process will most likely eliminate the background noise, and the listeners will lose information about their surrounding environment, i.e., environmental awareness will be affected. Although this can be viewed as a disadvantage of the approach, in practice it will never happen as we will never be able to estimate with perfect precision the true instantaneous SNR from the mixture envelopes. Errors will be introduced when estimating the SNR from the mixtures, and these errors will produce residual background noise containing information about the surrounding environment.

Both the C-rule and SNR selection rules provided (in optimal conditions) substantial benefit to CI users with regard to speech intelligibility in noise. This was done by retaining the favorable or right channels and discarding the rest. The definition of favorable was different for the two selection rules. For the SNR-rule, favorable channels were the target-dominated channels. For the C-rule, favorable channels were the ones containing only attenuation distortions, which is one type of distortion introduced by noise-reduction algorithms. This type of distortion, in particular, was found (Kim and Loizou, 2011) not to impair intelligibility, compared to the amplification distortions that severely degrade intelligibility. In the context of noise reduction, speech distortions are unavoidable given that inaccurate gain suppression functions (caused by our lack of knowledge about the noise statistics, among other things) are applied to the mixture envelopes. The C-rule improves speech intelligibility by controlling the type of distortions that are allowed to pass and discarding the unfavorable distortions (amplification distortions).

In the present study, we assumed prior knowledge of the SNR (for the SNR-rule) or the target signal (for the C-rule). This was found necessary for accurate channel selection [see Eqs. (3) and (4)] and was done to assess the full potential of the proposed channel selection criteria. But how is it possible to estimate the target and the masker spectra in realistic conditions with a single microphone, particularly at low SNR levels, to implement the preceding channel selection schemes? For the implementation of the SNR rule, for instance, it is necessary to estimate and track the noise spectrum accurately at every frame [see Eq. (1)]. As it turns out, it is not necessary to estimate the target and/or masker spectra accurately (or even estimate them at all) to implement the preceding channel selection schemes. As we demonstrated in our prior studies with NH listeners (Kim *et al.*, 2009; Kim and Loizou, 2010) and CI users (Hu and Loizou, 2010) it is feasible to implement the channel-selection process using only information available from the mixture envelopes and without requiring access to noise statistics. This can be done by exploiting the fact that it is not necessary to estimate the actual (instantaneous) SNR value of each channel accurately but rather the SNR region ($\text{SNR} > T$ vs $\text{SNR} \leq T$) of the mixture envelopes, i.e., whether the underlying SNR of the channel mixture envelope is greater or smaller than a prescribed threshold T . Similarly, for the C-rule it is not necessary to estimate the actual value of the ratio $|\hat{X}_M(f, t)|/|X(f, t)|$ but rather whether $|\hat{X}_M(f, t)|/|X(f, t)|$ is larger or smaller than 2. Hence given that the classification and retention of channels is binary, one can train a binary classifier to classify the mixture envelopes into two classes, namely, those with $\text{SNR} > T$ and those with $\text{SNR} \leq T$ (or $|\hat{X}_M(f, t)|/|X(f, t)|$ greater or smaller than 2, for the C-rule). True knowledge of the target signal is only required during the training stage of the binary classifier, which can be done off-line using a large inventory of noise sources. In the testing stage, the binary classifier will have access only to the observed noisy signals and will make classification decisions using features extracted from noisy observations. Such an approach was taken by Kim *et al.* (2009), who demonstrated large gains in intelligibility with NH listeners. A binary classifier trained for the C-rule

was implemented and evaluated with NH listeners in the study by Kim and Loizou (2010). Large gains in intelligibility were obtained at SNR levels as low as -5 dB. The binary classifier approach based on the SNR selection rule was extended and applied to CI users in the study by Hu and Loizou (2010). Results with CI listeners indicated substantial improvements (by nearly 44 percentage points) in intelligibility with the proposed algorithm when tested with sentences embedded in three real-world maskers at 5 dB SNR (Hu and Loizou, 2010). In brief, the data from the preceding studies demonstrated that the implementation of C-rule or SNR-rule is challenging yet feasible (for both stationary and non-stationary noises).

V. SUMMARY AND CONCLUSIONS

In summary, the present study evaluated the full potential, in terms of intelligibility benefit and quality, of two different channel selection criteria for CIs. Both criteria used prior knowledge of the input signal and restored intelligibility to the level attained in quiet with the C-rule offering a slight intelligibility advantage at $\text{SNR} = -5$ dB (see Fig. 2). A number of conclusions can be drawn from the present study. First, channel selection provides an effective means of improving speech intelligibility in noise for CI users in the absence of binaural (spatial) cues. While not tested here, large intelligibility benefits can be obtained using channel-selection based coding strategies in other masker conditions including competing talker and multi-talker babble. Prior tests with NH listeners, for instance, indicated no dependency of intelligibility benefit on the type of masker used (Li and Loizou, 2008; Kim and Loizou, 2011) even when the noise spectrum was not known *a priori* and was estimated from the noisy signal (Kim and Loizou, 2010). Second, it is not necessary to increase the number of effective channels of information transmitted by CIs to improve speech intelligibility in noise (e.g., Friesen *et al.*, 2001). This is so because we have shown in the present study that speech intelligibility in noise can be improved by modifying the channel selection criterion rather than increasing the number of effective channels. This is encouraging given the great challenge in reducing channel interaction, contributing to the limited number of channels transmitted with current electrode arrays and electrode configurations.

ACKNOWLEDGMENTS

This work was supported by Grant No. R01 DC010494 awarded from the National Institute of Deafness and other Communication Disorders (NIDCD) of the National Institutes of Health (NIH). The authors would like to thank the CI patients for their time.

APPENDIX

In this appendix, we describe the implementation of the Wiener noise reduction algorithm (Scalart and Filho, 1996). The corrupted sentences were first segmented into 20-ms frames with 50% overlap between adjacent frames. Each speech frame was Hamming windowed and a 320-point

discrete Fourier transform (DFT) was computed. Let $Y(f, t)$ denote the complex noisy spectrum at time frame t and frequency band f . Then the estimate of the target signal spectrum, $\hat{X}(f, t)$, is obtained by multiplying $Y(f, t)$ with the square-root Wiener gain function $G(f, t)$ as follows:

$$\hat{X}(f, t) = G(f, t) \cdot Y(f, t), \quad (\text{A1})$$

where

$$G(f, t) = \sqrt{\frac{\text{SNR}_{\text{prio}}(f, t)}{1 + \text{SNR}_{\text{prio}}(f, t)}}, \quad (\text{A2})$$

and SNR_{prio} is estimated using the “decision-directed” method (Ephraim and Malah, 1984) as

$$\text{SNR}_{\text{prio}}(f, t) = \alpha \cdot \frac{|\hat{X}(f, t-1)|^2}{\hat{\lambda}_D(f, t-1)} + (1 - \alpha) \cdot \max \left[\frac{|Y(f, t)|^2}{\hat{\lambda}_D(f, t)} - 1, 0 \right], \quad (\text{A3})$$

where $\alpha = 0.98$ is a smoothing constant and $\hat{\lambda}_D$ is the estimate of the background noise spectral variance obtained from the first 120 ms of the noisy signal and is updated during silence periods:

$$\hat{\lambda}_D(f, t) = \mu \cdot \hat{\lambda}_D(f, t-1) + (1 - \mu) \cdot |Y(f, t)|^2, \quad (\text{A4})$$

where μ is a constant (set to 0.98). The estimate of speech magnitude spectrum and the estimate of the noise spectral variance were initialized as $|\hat{X}(f, 0)| = 0$ and $\hat{\lambda}_D(f, 0) = \hat{\lambda}_D(f, 1) = |Y(f, 1)|^2$.

¹<http://www.utdallas.edu/loizou/cimplants/> (Last viewed 08/29/2012).

²The ACE strategy implemented in the SPEAR3 processor is very similar to that implemented in the Nucleus 24, Nucleus 5, and Freedom systems, and most coding parameters of the SPEAR3 ACE strategy matched those used in the commercial systems. In addition it was ensured that all parameters programmed (e.g., stimulation rate, number of maxima, frequency allocation table) matched the clinical settings of the individual CI users. Use of SPEAR3 provides us with the advantage of testing all subjects with a basic ACE implementation without having to be concerned with special settings (e.g., ADRO, etc.) used by individual patients. That is, it provides a more consistent evaluation without the confounds introduced by advanced features (e.g., ADRO, BEAM, etc.) available (and enabled) in daily processors by some patients. Moreover, the SPEAR3 has the capability of direct connection to the implant.

- Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (2006). “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” *J. Acoust. Soc. Am.* **120**, 4007–4018.
- Chung, K. (2004). “Challenges and recent developments in hearing aids: Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms,” *Trends Amplif.* **8**, 83–124.
- Dawson, P. W., Mauger, S. J., and Hersbach, A. A. (2011). “Clinical evaluation of signal-to-noise ratio-based noise reduction in nucleus cochlear implant recipients,” *Ear Hear.* **32**, 382–390.

- Dudley, H. (1939). “Remarking speech,” *J. Acoust. Soc. Am.* **11**, 169–177.
- Ephraim, Y., and Malah, D. (1984). “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust. Speech Signal Process.* **32**, 1109–1121.
- 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.
- Hu, Y., and Loizou, P. C. (2008). “A new sound coding strategy for suppressing noise in cochlear implants,” *J. Acoust. Soc. Am.* **124**, 498–509.
- Hu, Y., and Loizou, P. C. (2010). “Environment-specific noise suppression for improved speech intelligibility by cochlear implant users,” *J. Acoust. Soc. Am.* **127**, 3689–3695.
- Hu, Y., Loizou, P. C., Li, N., and Kasturi, K. (2007). “Use of a sigmoidal-shaped function for noise attenuation in cochlear implants,” *J. Acoust. Soc. Am.* **122**, EL128–EL134.
- IEEE (1969). “IEEE recommended practice for speech quality measurements,” *IEEE Trans. Audio Electroacoust.* **AU-17**, 225–246.
- ISO/IEC (1999). *JTC1/SC29/WG11 MPEG, Information Technology-Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s-Part 3: Audio, IS11172-3, (MPEG-1)* (International Organization for Standardization, Geneva).
- ITU-T (1996). “Subjective performance assessment of telephone band and wideband digital codecs,” ITU-T Recommendation (International Telecommunications Union Telecommunication Standardization Sector, Geneva), p. 830.
- Kim, G., and Loizou, P. C. (2010). “Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints,” *IEEE Signal Process. Lett.* **17**, 1010–1013.
- Kim, G., and Loizou, P. C. (2011). “Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms,” *J. Acoust. Soc. Am.* **130**, 1581–1596.
- Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (2009). “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” *J. Acoust. Soc. Am.* **126**, 1486–1494.
- Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (2009). “Role of mask pattern in intelligibility of ideal binary-masked noisy speech,” *J. Acoust. Soc. Am.* **126**, 1415–1426.
- Kokkinakis, K., Hazrati, O., and Loizou, P. C. (2011). “A channel-selection criterion for suppressing reverberation in cochlear implants,” *J. Acoust. Soc. Am.* **129**, 3221–3232.
- Li, N., and Loizou, P. C. (2008). “Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction,” *J. Acoust. Soc. Am.* **123**, 1673–1682.
- Loizou, P. C. (2006). “Speech processing in vocoder-centric cochlear implants,” *Adv. Otorhinolaryngol.* **64**, 109–143.
- Loizou, P. C. (2007). *Speech Enhancement: Theory and Practice* (Taylor and Francis, Boca Raton, FL), pp. 1–608.
- Loizou, P. C. (2011). “Speech quality assessment,” in *Multimedia Analysis, Processing and Communications*, edited by W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, and H. Wang (Springer Verlag, New York), Vol. 346, pp. 623–654.
- Loizou, P. C., and Kim, G. (2011). “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” *IEEE Trans. Audio Speech Lang. Process.* **19**, 47–56.
- Noguiera, W., Buchner, A., Lenarz, T., and Edler, B. (2005). “A psychoacoustic ‘NoFM’-type speech coding strategy for cochlear implants,” *EURASIP J. Appl. Signal Process.* **18**, 3044–3059.
- Peterson, E., and Cooper, F. S. (1957). “Peak picker: A bandwidth compression device,” *J. Acoust. Soc. Am.* **29**, 777.
- Scalart, P., and Filho, J. V. (1996). “Speech enhancement based on a priori signal to noise estimation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Atlanta, GA), pp. 629–632.
- Vandali, A. E., Whitford, L. A., Plant, K. L., and Clark, G. M. (2000). “Speech perception as a function of electrical stimulation rate: Using the Nucleus 24 cochlear implant system,” *Ear Hear.* **21**, 608–624.