Original Article

Trends in Hearing
Volume 23: 1–12
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2331216519825930
journals.sagepub.com/home/tia
SAGE

Effect of Noise Reduction Gain Errors on Simulated Cochlear Implant Speech Intelligibility

Abigail A. Kressner[1], Tobias May[1], and Torsten Dau[1]

Abstract

It has been suggested that the most important factor for obtaining high speech intelligibility in noise with cochlear implant (CI) recipients is to preserve the low-frequency amplitude modulations of speech across time and frequency by, for example, minimizing the amount of noise in the gaps between speech segments. In contrast, it has also been argued that the transient parts of the speech signal, such as speech onsets, provide the most important information for speech intelligibility. The present study investigated the relative impact of these two factors on the potential benefit of noise reduction for CI recipients by systematically introducing noise estimation errors within speech segments, speech gaps, and the transitions between them. The introduction of these noise estimation errors directly induces errors in the noise reduction gains within each of these regions. Speech intelligibility in both stationary and modulated noise was then measured using a CI simulation tested on normal-hearing listeners. The results suggest that minimizing noise in the speech gaps can improve intelligibility, at least in modulated noise. However, significantly larger improvements were obtained when both the noise in the gaps was minimized and the speech transients were preserved. These results imply that the ability to identify the boundaries between speech segments and speech gaps may be one of the most important factors for a noise reduction algorithm because knowing the boundaries makes it possible to minimize the noise in the gaps as well as enhance the low-frequency amplitude modulations of the speech.

Keywords

cochlear implant, noise reduction, sound coding, speech intelligibility

Date received: 6 February 2018; revised: 26 December 2018; accepted: 27 December 2018

Introduction

Listening to speech in the presence of interfering noise is a demanding task. This is especially true for cochlear implant (CI) recipients (e.g., Hochberg, Boothroyd, Weiss, & Hellman, 1992), at least in part because of the fact that CI recipients have limited access to the underlying spectral and temporal information in speech. More recently, there has been extensive research on noise reduction algorithms and sound coding strategies to improve CI recipients' resilience to noise. This research has led to speech intelligibility improvements both with single-microphone noise reduction (e.g., Mauger, Arora, & Dawson, 2012) and with multimicrophone directional noise reduction (e.g., Hersbach, Grayden, Fallon, & McDermott, 2013; Spriet et al., 2007). However, the benefit of these single-microphone algorithms diminishes in the presence of modulated noise types (e.g., Mauger et al., 2012). An exception could be made for algorithms that use machine learning techniques (e.g., Goehring et al., 2017), but these algorithms currently suffer from limited generalization abilities and have not yet been implemented in clinical devices. Furthermore, the benefit of the multimicrophone directional noise reduction algorithms diminishes when the target and interfering signals are not well separated in space. Thus, despite the recent improvements in speech intelligibility outcomes for CI recipients in noisy

[1]Hearing Systems, Department of Health Technology, Technical University of Denmark, Denmark

Corresponding Author:
Abigail A. Kressner, Hearing Systems, Department of Health Technology, Technical University of Denmark, Ørsteds Plads, Building 352, 2800 Kgs. Lyngby, Denmark.
Email: aakress@dtu.dk

environments, there still remains room for improvement, especially in more realistic scenarios, such as a restaurant, where interfering noises typically come from many directions and are almost certainly fluctuating in time.

One potential barrier for improving speech intelligibility in the presence of noise is that relatively little is known about which cues CI recipients rely on most to understand speech in noise. Without knowing which information should be prioritized for encoding, it is difficult to properly design and optimize any sound coding algorithm. In an effort to improve this understanding, Qazi, van Dijk, Moonen, and Wouters (2013) investigated the effects of noise on electrical stimulation sequences and speech intelligibility in CI recipients. They suggested that noise affects stimulation sequences in three primary ways: (a) noise-related stimulation can fill the gaps between speech segments, (b) stimulation levels during speech segments can become distorted, and (c) channels that are dominated by noise can be selected for stimulation instead of channels that are dominated by speech. To measure the effect of each of these factors, Qazi et al. (2013) generated several artificial stimulation sequences, each of which contained different combinations of these errors. They presented these artificial stimulation sequences to CI recipients, as well as to normal-hearing listeners with a vocoder, and measured speech intelligibility in stationary noise. Their results indicated that the most important factor for maintaining good speech intelligibility was the preservation of the low-frequency (i.e., what they called "ON/OFF") amplitude modulations of the clean speech. Furthermore, they argued that one possible method for preserving these cues would be to minimize the noise presented in speech gaps.

Koning and Wouters (2012), however, argued that it is the information encoded in the transient parts of the speech signal that contributes most to speech intelligibility. They demonstrated that enhancing speech onset cues alone improves speech intelligibility in CI recipients (Koning & Wouters, 2016). By comparison, Qazi et al. (2013) also inherently enhanced onset and offset cues in the conditions where they removed noise in the gaps between speech segments because they always identified these segments via onset and offset detection with a priori information. Thus, by removing noise in the speech gaps in their experiment, they simultaneously enhanced the saliency of the onsets and offsets. Qazi et al. (2013) did not, however, investigate the effect of reducing noise in the gaps when the boundaries between the speech segments and speech gaps were not perfectly aligned. Therefore, it is unclear how advantageous the minimization of the noise in speech gaps is when it does not co-occur with accurate onset and offset cues. Furthermore, the importance of the separation of these two factors becomes clear when considering that realistic algorithms will not always be able to perfectly identify the boundaries between speech segments and speech gaps.

The main purpose of the present study was to systematically quantify the relative impact of errors in the noise reduction gains that are applied within speech segments, speech gaps, and the transitions between them to determine which errors contribute most to reducing the benefit of noise reduction for CI recipients, especially in nonstationary noise where a clinically significant benefit has yet to be shown with existing single-channel noise reduction algorithms. Specifically, noise reduction gain matrices (i.e., sets of gains across time and frequency) were synthesized for noisy sentences by combining the sets of gains calculated in each of the three temporal regions from either a priori signal-to-noise ratios (SNRs) or SNRs computed by noise power density estimation (Cohen, 2003). Speech intelligibility was then measured in denoised sentences using a basic CI vocoder simulation with normal-hearing listeners. This protocol provides insight into the impact of the spectrotemporal degradation in isolation from an impaired auditory system.

Methods

Whereas Qazi et al. (2013) primarily manipulated channel selection and current levels within each temporal region to investigate the impact of noise-induced errors in stimulation strategies, the present study manipulated the gains that were applied in a preceding noise reduction stage to investigate the impact of noise-induced errors on noise reduction algorithms rather than on channel selection. Therefore, an underlying assumption in this study was that a maxima selection strategy, such as the Advanced Combination Encoded (ACE^{TM}, Cochlear Ltd., New South Wales, Australia), would stimulate the correct set of channels if it chooses channels from a representation that has been sufficiently denoised.

Stimuli

A CI with an N-of-M strategy encodes sound by first separating the input signal into M channels and subsequently stimulating a subset of at most N channels at each frame l. In this study, speech was divided into 128-sample overlapping frames (8 ms with a sampling rate of 16 kHz), and then a Hann window and the short-time discrete Fourier Transform was applied with $K = 128$ points to obtain the time-frequency representation of speech, $X(k, l)$, where k represents the discrete Fourier Transform bin index. The short-time discrete Fourier Transform magnitudes were then combined into $M = 22$ channels using nonoverlapping rectangular weights with spacing that matches Cochlear Ltd.'s (New South Wales, Australia) sound processor to obtain the time-frequency representation $X(m, l)$, where m represents the channel index, and l represents the frame index. The channel center frequencies ranged from 187.5 Hz to 7937.5 Hz, with bandwidths of 125 Hz for the lowest 9 bands, 250 Hz for bands 10 to 13, 375 Hz for bands 14 and 15, 500 Hz for bands 16 and 17, 625 Hz for bands 18 and 19, 750 Hz for band 20, 875 Hz for band 21, and 1000 Hz for band 22. A new frame was calculated every 1 ms.

Sentences were divided temporally into three regions: speech segments, speech gaps, and speech transitions. In comparison, Qazi et al. (2013) divided sentences into only two temporal regions (i.e., speech segments and speech gaps). Having three rather than two temporal regions facilitated manipulation of the noise in the gaps independently from manipulation of the transitions. More specifically, this protocol made it possible to measure the impact of minimizing noise in the gaps when the transitions are not perfectly encoded. To do this segmentation, broadband channel activity, $A(l)$, was defined for each frame as the number of channels containing speech above a threshold:

$$A(l) = \sum_{m=1}^{M} T_\lambda(X(m, l)) \tag{1}$$

where the function $T_\lambda(\cdot)$ performs elementwise thresholding and returns a value of one for elements that have a sound pressure level of more than 25 dB (i.e., the default threshold level in ACE). As in Qazi et al.'s (2013) study, speech segment onsets were then identified as frames in which $A(l) = 0$ and $A(l+1) > 0$, and speech segment offsets were defined as frames in which $A(l) > 0$ and $A(l+1) = 0$. Speech segments with $A(l) \leq 1$ for the duration of the segment were dropped, and speech segments shorter in duration than 20 ms that were within 20 ms of another speech segment were combined. This merging prevented rapid switches between speech and nonspeech labels. Subsequently, a transition region was defined at each onset and offset as the 10 ms before and the 10 ms after, such that a transition region of 20 ms in duration was created at the start and end of each speech segment. Finally, the remaining frames were labeled as speech gaps. An example stimulation sequence for the Danish sentence, "*Stuen skal nok blive hyggelig*," is shown in Figure 1(a), wherein the gap regions are indicated with the underlying dark gray shading, the transition regions with the light gray shading, and the speech regions with white shading. The 20-ms duration for the transition region was heuristically chosen to ensure the transition regions were long enough to be perceptible but short enough to maintain a segmentation that was still comparable with the segmentation in Qazi et al. (2013).

The following general signal model was thereby considered: $Y(k, l) = X(k, l) + D(k, l)$, with $X(k, l)$

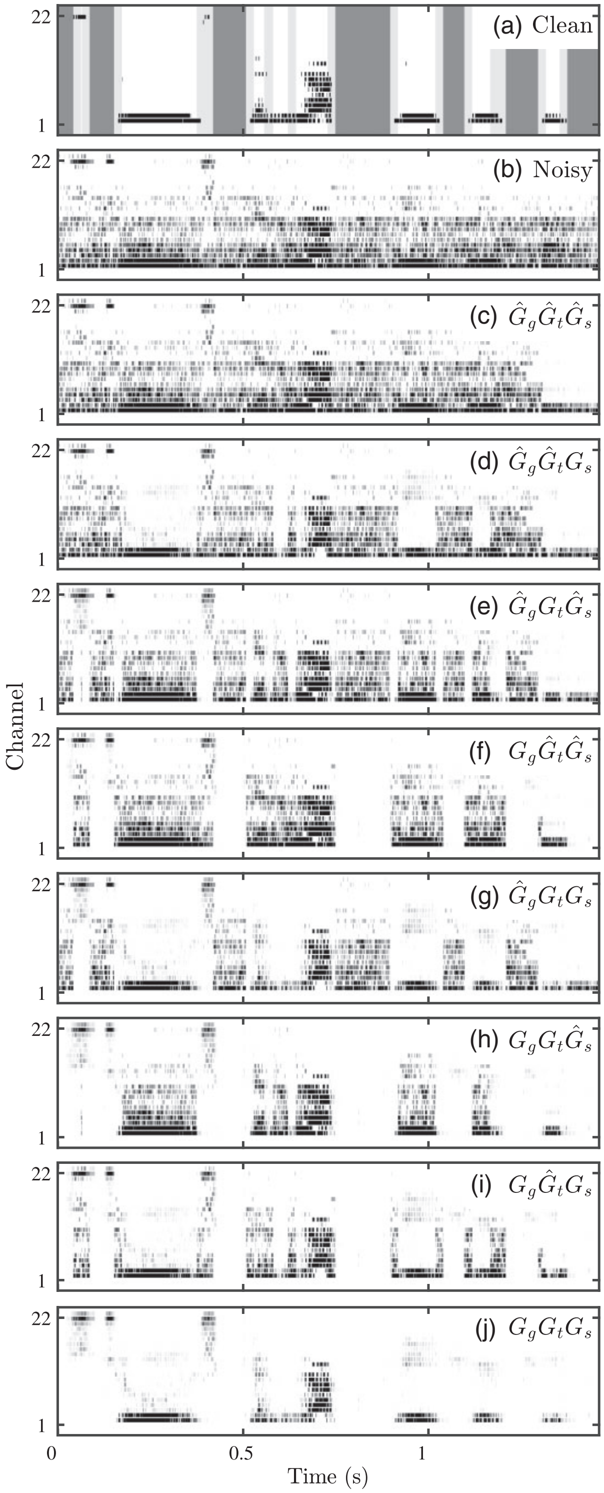

Figure 1. (a) Electrodogram showing stimulation levels above threshold for a clean sentence. Speech segments, transitions, and gaps are identified by the white, light gray, and dark gray shading, respectively. (b to j) Electrodograms showing unthresholded levels for the same sentence mixed with speech-shaped noise at 0 dB and then denoised using the indicated gain matrix, where $\hat{G}_g\hat{G}_t\hat{G}_s$ indicates the use of nonideal gains in the gap, transition, and speech regions, respectively; $\hat{G}_g\hat{G}_tG_s$ indicates the use of nonideal gains in the gap and transition regions but ideal gains in the speech regions, and so forth.

representing the clean speech, $D(k, l)$ representing the noise signal, and $Y(k, l)$ representing the noisy speech signal. An estimate of the noise spectrum $\hat{D}(k,l)$ was computed from the noisy signal $Y(k, l)$ using the improved minima controlled recursive averaging algorithm (Cohen, 2003) with parameters as specified therein and an initial noise power estimate calculated from the first 50 ms of the mixture.

$\hat{D}(m,l)$ was then computed from $\hat{D}(k,l)$ using the same rectangular weights as were used for computing $X(m, l)$ from $X(k, l)$, and a smoothed SNR estimate $\hat{\xi}(m,l)$ was obtained using a CI-optimized smoothing technique (Mauger et al., 2012), which consists of recursive averaging with $\alpha = .984$. From $\hat{\xi}(m,l)$, the gains $\hat{G}(m,l)$ were obtained using the CI-optimized gain function (Mauger et al., 2012):

$$\hat{G}(m,l) = \left(\frac{\hat{\xi}(m,l)}{\hat{\xi}(m,l) + 2.92}\right)^{1.2} \tag{2}$$

In addition, the ideal gains $G(m, l)$ were computed using $\xi(m, l)$, which in turn was based on $D(m, l)$.

Artificial gain matrices were synthesized by concatenating segments from either $\hat{G}(m,l)$ or $G(m, l)$ for each of the three temporal regions. For example, to create stimuli without gain errors in the speech gaps, gains from $G(m, l)$ were applied to the noisy signal $Y(m, l)$ in all of the speech gaps, whereas gains from $\hat{G}(m,l)$ were applied in all of the speech transitions and speech segments. This condition was named $G_g\hat{G}_t\hat{G}_s$ to indicate that the estimated gains were corrected in the speech gaps, but not in the transitions and the speech segments. Accordingly, the condition $\hat{G}_gG_t\hat{G}_s$ indicates that the estimated gains were corrected in the speech transitions, and it follows that the condition $G_gG_tG_s$ signifies that the estimated gains were corrected in all of the temporal regions, which is equivalent to ideal Wiener processing with a CI-optimized gain function.

The final stimulation sequence was computed by selecting the $N = 8$ channels with the largest remaining energy. Figure 1(b) shows the sequences for a noisy version of the sentence in Figure 1(a), and Figure 1(b to j) shows the sequences after denoising with each type of gain matrix. A visual comparison between Figure 1(c) and (j) highlights the extent of the estimation errors in $\hat{G}_g\hat{G}_t\hat{G}_s$. Subsequently, the remaining figures contain the stimulation patterns for the conditions where just one or two of the temporal regions of the gain matrix have been corrected.

Procedure

Acoustic signals were constructed from each of the synthesized stimulation sequences using a 22-channel noise vocoder. Speech intelligibility was then evaluated by measuring speech reception thresholds (SRTs) with normal-hearing listeners using the Danish hearing in noise test (HINT; Nielsen & Dau, 2011). Through an adaptive procedure, HINT determines the SNR at which the participants were able to understand 50% of the sentence material. Each HINT sentence was padded with 1 s of zeros before the start of the sentence and with 600 ms of zeros after the end of the sentence. The sentences were then combined with a randomly selected segment of either stationary speech-shaped noise (Nielsen & Dau, 2011) or the International Speech Test Signal (Holube, Fredelake, Vlaming, & Kollmeier, 2010). While the stationary noise is shaped to have the same long-term average spectrum as the HINT sentences, the International Speech Test Signal has the same temporal modulations as speech but is not intelligible and is not shaped to specifically match the target sentences of the Danish HINT corpus. This mixing procedure resulted in the noise being played for 1 s before and 600 ms after the target sentence.

As in the standard Danish HINT, the overall amplitude of each mixture was gradually increased over the first 400 ms and, likewise, gradually decreased over the last 400 ms. Because the noise estimate was initialized during this ramp-up segment, the noise was always underestimated at the start. This setup guaranteed the presence of pronounced, but realistic, noise reduction errors at the start of the target sentence, even in the case of the stationary noise. The resulting mixtures were normalized so that the sound pressure level over the duration of the target sentence was always 65 dB.

At the start of the session, participants first heard vocoded sentences in quiet and then in noise to become familiar with the task. Testing subsequently commenced with either the stationary or the modulated noise. There were eight noise reduction conditions, and together with the reference condition using unprocessed noisy speech (i.e., processed with unity gains), there were nine test conditions. One SRT was collected per condition. The order of the presentation of test lists and conditions was randomized. The testing was carried out in a double-walled booth, using equalized Sennheiser HD-650 circumaural headphones and a computer running a MATLAB graphical user interface.

Participants

Thirty normal-hearing listeners participated in this study. The participants were randomly assigned to one of two groups, each of which heard either the stationary or the modulated noise. Thus, each group consisted of 15 participants. Participants were at least 18 years of age, had audiometric thresholds of less than or equal to 20 dB hearing level in both ears (125 Hz to 8 kHz), and were native Danish speakers. All participants provided informed consent, and the experiment was approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). The participants were paid for their participation.

The first six participants in this study took part in an extended version of the protocol, wherein two SRTs were collected for each condition, and each listener heard both stationary and modulated noise. The results for these six listeners were reported in Kressner et al. (2017). However, because of the limited size of the HINT corpus, this extended protocol required that the participants heard each sentence multiple times. To limit the influence of the training effects that are inevitable with this kind of repetition (Yund & Woods, 2010), the protocol for the remaining 24 participants removed repetitions altogether by collecting only one SRT for one type of noise. The scores for the repetitions (i.e., the second through fourth presentations of each list) from the first six participants were discarded.

Analysis

Statistical inference was performed by fitting a linear mixed-effects model to the SRT improvement scores, which were calculated for each individual relative to the individual's score in the reference unprocessed condition. The fixed effects terms of the mixed model were the noise type, the gains in the gap regions, the gains in the transition regions, and the gains in the speech regions. The model also included a subject-specific intercept (i.e., the participants were treated as a random factor, as is standard in a repeated-measures design). The model was implemented in the R software environment using the *lme4* library (Bates, Mächler, Bolker, & Walker, 2015). Further, model selection was carried out with the *lmerTest* library (Kuznetsova, Brockhoff, & Christensen, 2017), which uses stepwise deletion of model terms with high *p* values to perform backward elimination of random-effect terms and then backward elimination of fixed-effect terms (Kuznetsova, Christensen, Bavay, & Brockhoff, 2015). The *p* values for the fixed effects were calculated from *F* tests based on Satterthwaite's approximation of denominator degrees of freedom, and the *p* values for the random effects were calculated based on likelihood ratio tests (Kuznetsova et al., 2015).

Post hoc analysis was performed through contrasts of estimated marginal means using the *emmeans* library (Lenth, 2018; Searle, Speed, & Milliken, 1980) and the *lme4* model object. The *p* values were calculated using the Kenward–Roger's degrees-of-freedom method, and a correction for the multiple comparisons was included using the Tukey method. Significant differences are reported using $\alpha = .05$.

Results

Figure 2 shows SRT scores for each individual listener, and Figure 3 shows the group distributions of SRT and SRT improvement (i.e., SRTs relative to the reference unprocessed condition). Group results were modeled using the aforementioned linear mixed-effects model. The model showed a significant main effect for the gains in the gap regions, $F(1,\ 196) = 757.54$, $p < .0001$; the gains in the transition regions, $F(1,\ 196) = 186.90$, $p < .0001$; the gains in the speech regions, $F(1,\ 196) = 392.25$, $p < .0001$; and the noise type, $F(1,\ 28) = 9.94$, $p < .01$. The interactions between the noise type and the gains in the gap regions and between the noise type and the gains in the speech regions were both significant, $F(1,\ 196) = 12.73$, $p < .001$; $F(1,\ 196) = 7.01$, $p < .01$, whereas the interaction between the noise type and the gains in the transition regions was not significant, $F(1,\ 196) = 1.06$, $p = .30$. Furthermore, the interactions

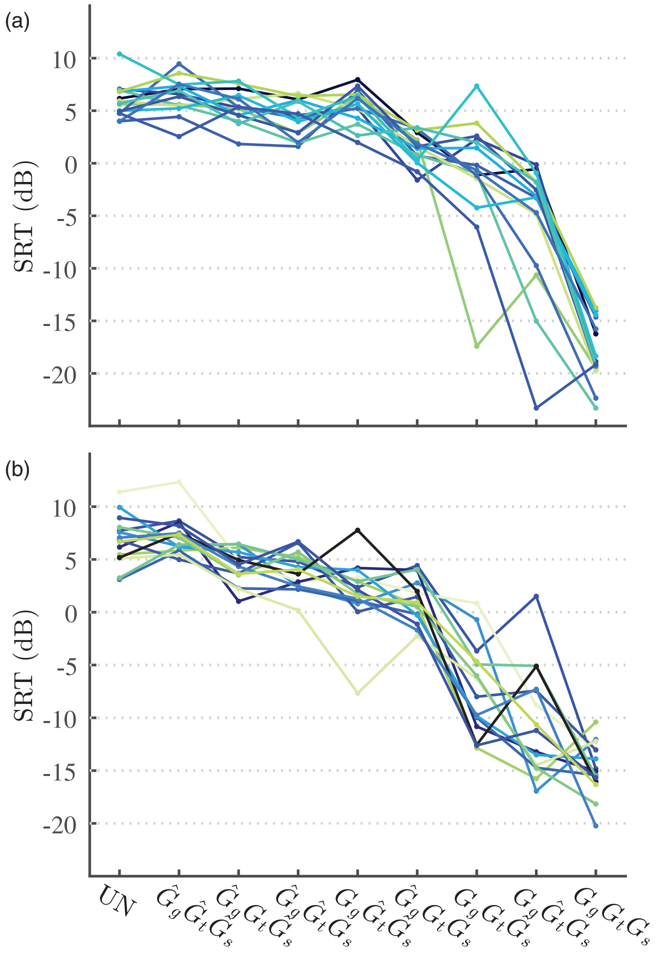

Figure 2. Individual SRTs for listeners who heard (a) stationary noise and (b) modulated noise. The condition labels along the abscissa are defined in the text, as well as in the caption of Figure 1. SRTs = speech reception thresholds; UN = unprocessed noisy speech.

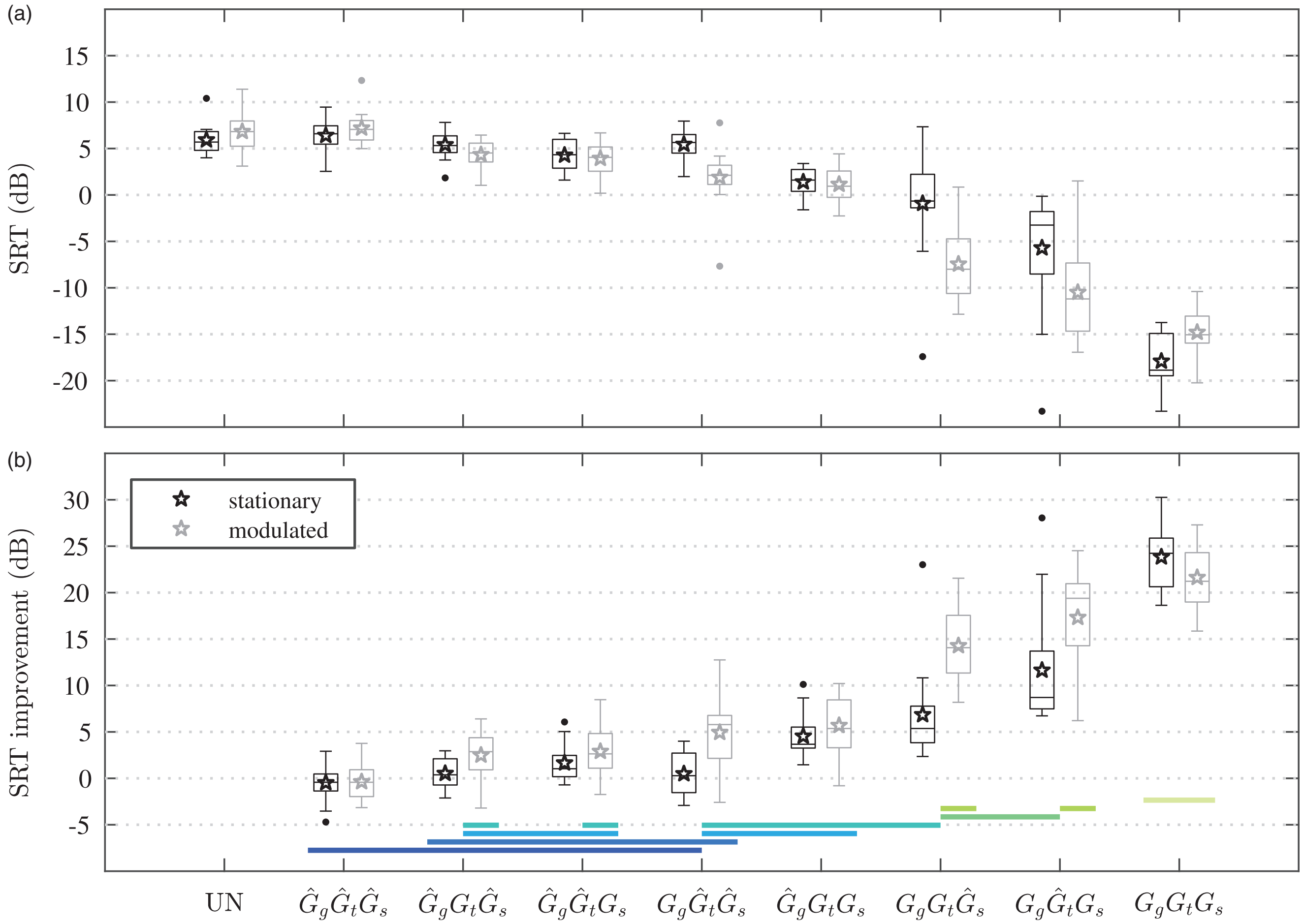

Figure 3. (a) SRT and (b) SRT improvements relative to the reference condition (UN). Means are marked with stars. Boxplots show the 25th, 50th, and 75th percentiles, together with whiskers that extend to cover all data points not considered outliers. Outliers are marked with circles. SRT improvements that were not significantly different from one another ($\alpha = .05$) are grouped via colored, horizontal lines at the bottom of the plot. The condition labels along the abscissa are defined in the text, as well as in the caption of Figure 1.
SRTs = speech reception thresholds; UN = unprocessed noisy speech.

between the gains in the gap regions and the gains in the transition regions and between the gains in the gap regions and the gains in the speech regions were both significant, $F(1, 196) = 55.09$, $p < .0001$; $F(1, 196) = 133.42$, $p < .0001$. On the other hand, the interaction between the gains in the transition regions and the gains in the speech regions was nonsignificant, $F(1, 196)$ 0.73, $p = .39$. All three-way interactions with the noise type were significant, $F(1, 196) = 4.80$, $p = .03$ with the gap and transition regions; $F(1, 196) = 8.19$, $p < .01$ with the gap and speech regions; and $F(1, 196) = 17.68$, $p < .0001$ with the transition and speech regions, but the three-way interaction between the gains in each of the three temporal regions was nonsignificant, $F(1, 196) = 0.11$, $p = .74$. Last, the four-way interaction was significant, $F(1, 196) = 8.55$, $p < .01$, meaning that none of the terms could be eliminated during model selection.

Pairwise comparisons were subsequently conducted between each of the conditions. Groups of conditions that did not have significantly different means from one another are indicated via the colored, horizontal lines in the bottom of Figure 3(b). Neither $\hat{G}_gG_t\hat{G}_s$ nor $\hat{G}_g\hat{G}_tG_s$ yielded scores with means that were significantly different from the scores with the baseline $\hat{G}_g\hat{G}_t\hat{G}_s$ gains. However, the mean SRT improvement after applying the $G_g\hat{G}_t\hat{G}_s$ gain matrix was significantly different from that of $\hat{G}_g\hat{G}_t\hat{G}_s$, but only in the case of the modulated noise type. Furthermore, each of the gain matrices with two of the three regions corrected yielded improvements with means that were significantly different from both the baseline $\hat{G}_g\hat{G}_t\hat{G}_s$ gain matrix and the ideal $G_gG_tG_s$ gain matrix. Thus, the introduction of any errors, even in temporal regions with a relatively short duration such as the transition regions, significantly influences the intelligibility of the processed speech.

Because normal-hearing listeners generally do not benefit from single-microphone noise reduction algorithms (Hu & Loizou, 2007), it is not surprising that the $\hat{G}_g\hat{G}_t\hat{G}_s$ gain matrix did not provide an SRT improvement for either noise type, despite that the gain matrix was estimated with a CI-optimized noise reduction algorithm that has been shown to improve speech intelligibility for CI recipients (Mauger et al., 2012). Similarly, it is not surprising that the SRTs improved by as much as 30 dB when a priori information about the local SNRs was used, as this was the maximum possible benefit given the starting SNR and adaptation rules of the SRT measurement method (Nielsen & Dau, 2011).

Gap Regions

The impact of reducing errors in the gap regions can be evaluated in two ways: (a) by comparing the SRT improvements with the $\hat{G}_gG_tG_s$ gain matrix to those with the $G_gG_tG_s$ gain matrix and (b) by comparing the SRT improvements with the $G_g\hat{G}_t\hat{G}_s$ gain matrix to those with the $\hat{G}_g\hat{G}_t\hat{G}_s$ gain matrix. In the former case where the only errors that were present were those in the gaps, SRT improvements were smaller than for the gain matrices which contained errors either only in the speech regions or only in the transition regions. These results suggest that errors in the gap regions are more detrimental than errors in the other regions.

In the latter case where only errors in the gaps were removed from the estimated gain matrix, the mean change in SRT in the presence of stationary noise was not significantly different from the mean change in SRT with the baseline gain matrix. In the presence of the modulated noise, however, the mean SRT improvement was significantly different—though the magnitude of the change in SRT varied widely across participants. These results suggest that, when the detection of the transitions between the gaps and speech segments is imprecise, minimizing noise reduction errors in the gaps may only be beneficial in nonstationary noise.

Transition Regions

The impact of removing the gain errors in the transition regions (i.e., $\hat{G}_gG_t\hat{G}_s$ compared with $\hat{G}_g\hat{G}_t\hat{G}_s$) was nonsignificant with both noise types. Furthermore, the highest mean SRT improvements obtained, except those obtained with the ideal gain matrices, were obtained with the gain matrices that contained errors only in the transition regions (i.e., $G_g\hat{G}_tG_s$). These results point toward the conclusion that errors in the transition regions have limited impact compared with those in the gap or speech regions. However, the outcomes are likely confounded by the fact that the duration of the transition regions is relatively short compared with the other two temporal regions.

Despite their relatively limited duration, the interaction of the gains in the transition regions and the gains in the gap regions was highly significant. This interaction is further highlighted by the comparison between the $G_g\hat{G}_t\hat{G}_s$ and $G_gG_t\hat{G}_s$ gain matrices, where the paired comparisons within each noise type revealed significant differences. These results imply that the potential benefit of minimizing the stimulation in the gap regions largely depends on how accurately the boundaries between the gaps and segments of speech are encoded.

Speech Regions

In the presence of both the stationary and modulated noise, the impact of removing the gain errors in the speech regions (i.e., $\hat{G}_g\hat{G}_tG_s$ compared with $\hat{G}_g\hat{G}_t\hat{G}_s$) was nonsignificant. On the other hand, when the only errors that were present were those that were in the speech regions (i.e., $G_gG_t\hat{G}_s$ compared with $G_gG_tG_s$), SRT improvements were greater than when the gain matrices contained only errors in the gap regions but smaller than when the gain matrices contained only errors in the transition regions.

Interestingly, the interaction between the gains in the transition regions and the gains in the speech regions was nonsignificant, implying that—unlike with the gap regions—the magnitude of the benefit from correcting gain errors in the speech regions is not dependent on how accurately the boundaries between the speech and gap regions are identified. This is made especially clear by the nonsignificance of the difference between the $\hat{G}_g\hat{G}_tG_s$ outcomes and the $\hat{G}_gG_tG_s$ outcomes.

Discussion

The primary objective for this investigation was to determine which noise reduction gain errors are most responsible for limiting the benefit CI recipients receive from noise reduction algorithms, especially in modulated noise where a clinically significant benefit has not yet been shown. In modulated noise, errors in the gap regions had the most impact because correcting these errors led to a significant improvement. However, in stationary noise, these differences were nonsignificant. Thus, it seems that the region with the most detrimental effect depends on the temporal characteristics of the interfering noise. Despite this inconsistency, removing errors in both the transitions and the gaps simultaneously had a large impact in both noise types. Therefore, correctly encoding these two regions together seems to contribute substantially to understanding speech in noise. Overall though, the largest mean SRT improvements were obtained when both the speech and gap regions were restored. However, this phenomenon may, at least in part, be explained by the fact that the remaining distortions were restricted in time due to the relatively short duration of the transition regions compared with the other two regions.

Noise Reduction Errors Versus Stimulation Errors

In this study, artificial noise reduction gain matrices were created to systematically investigate the effects of noise reduction errors on speech intelligibility in noise. Despite the fact that Qazi et al. (2013) focused instead on errors in the stimulation pattern itself rather than in the noise reduction gain matrix, many comparisons can be made between the results in this study with those in Qazi et al. (2013).

For example, Qazi et al. (2013) measured the impact of ideal Wiener filtering and obtained a mean SRT of −16.0 dB for their normal-hearing listeners tested with a vocoder simulation. This ideal Wiener filtering condition matches closely to the $G_gG_tG_s$ condition in this study, with small differences existing only in the gain function and recursive smoothing that was applied. Listeners in this study obtained a mean SRT of −18.0 dB in stationary noise, which aligns relatively well.

Since the gains in the speech and transition regions in the $\hat{G}_gG_tG_s$ condition in this study were determined based on a priori SNRs, one can assume that the envelopes in the speech segments were completely restored by the noise reduction. The gap regions, on the other hand, included noise reduction that was based on estimated SNRs and, thereby, likely contained noise-dominated stimulation. Therefore, this condition corresponds to that of the "noise in the gaps" condition in Qazi et al. (2013), where the stimulation pattern from the clean sentence was presented during speech segments, while the stimulation pattern from the noise was presented in the gaps. However, Qazi et al. (2013) reported SRT improvements of about 11.5 dB for the normal-hearing listeners when the "noise in the gaps" stimuli were presented in stationary noise. In comparison, the normal-hearing listeners in this study obtained a mean SRT improvement of only 4.5 dB when presented with the stationary noise. It is not entirely clear from where this large difference arises, but one contributing factor could be differences between the speech and how their respective interfering noises are shaped. Specifically, the Danish HINT material was recorded with a male talker whose speech is dominated by low-frequency information that falls primarily within only the first channel. In comparison, results collected using a corpus with a female talker as in Qazi et al. (2013) would likely show a much stronger effect, particularly when using a vocoder, as the fundamental frequency and formant frequencies of a female talker are likely to be represented by multiple CI channels. A second contributing factor could be that the estimated gains that were applied within the gap regions may actually have introduced distortions that were more detrimental than presenting the unaltered noise-dominated envelopes.

An additional comparison can be made between the $G_g\hat{G}_t\hat{G}_s$ and $G_gG_t\hat{G}_s$ conditions in this study with the "ideal voice activity detector" condition in Qazi et al. (2013). In the "ideal voice activity detector" condition, stimulation patterns were synthesized by combining the channel selection pattern as specified by the clean sentence with the current levels as specified by the noisy mixture. Therefore, these stimulation patterns contained zero stimulation in the gap regions and ideal transition encoding, as well as ideal channel selection in the speech regions together with distorted current levels. The primary differences between these stimulation patterns and the sentences processed with $G_g\hat{G}_t\hat{G}_s$ gain matrix in the current study then are whether the transitions were encoded accurately, as well as whether the channels were correctly selected within the speech regions. On the other hand, the sentences processed with the $G_gG_t\hat{G}_s$ gain matrix contained accurate transition encoding and, therefore, only differ in whether the channels were correctly selected within the speech regions. In Qazi et al. (2013), the normal-hearing listeners obtained a mean SRT improvement of 19.0 dB with these stimulation patterns. In comparison, SRT improvements in stationary noise in the current study were on average 0.5 dB with the $G_g\hat{G}_t\hat{G}_s$ gain matrix and 7.0 dB with the $G_gG_t\hat{G}_s$ gain matrix. Therefore, even when noise-dominated stimulation in the gap regions was minimized and the transitions were ideally encoded, the listeners in the current study obtained a much smaller benefit than the normal-hearing listeners listening to the stimulation patterns with ideal channel selection. This large difference suggests that there remained enough distortions in the denoised envelopes within the speech regions to lead to adversely inaccurate channel selection. Furthermore, when the transition encoding is imprecise, the remaining benefit from minimizing noise-dominated stimulation in the gap regions was close to negligible.

Transient- and Onset-Enhancing Stimulation

Comparisons can also be made between this study and some of the previous studies that argued for the importance of the transition region. Vandali (2001) proposed a speech coding strategy called the transient emphasis spectral maxima (TESM) strategy, which was developed specifically to emphasize short-duration onset cues in speech. This strategy applied additional gain to a channel whenever there was a rapid rise in the channel's envelope. Furthermore, higher gain was applied when there was a rapid rise followed by a rapid fall (e.g., as might occur for a consonant burst) when compared with a rapid rise followed by a steady envelope level (e.g., as might occur at the onset of a vowel). In comparison with the recipients' everyday strategy, the TESM strategy provided significant improvements in the perception of nasal, stop, and fricative consonants. When full sentences were presented in multitalker noise at either 5 or 10 dB SNR—where the SNR presented depended on whether the recipient was a "good" performer—there was a statistically significant mean increase in word recognition of 5.7%. A similar trend was reported in Bhattacharya, Vandali, and Zeng (2011), where recipients received a benefit of approximately 8.5% with sentences mixed with stationary, speech-shaped noise at 10 dB SNR, 1.5% with sentences mixed at 5 dB SNR, and 0% with sentences mixed at 0 dB SNR. In contrast, however, Holden, Vandali, Skinner, Fourakis, and Holden (2005) found no significant difference in speech intelligibility with sentences presented in noise when this strategy was compared with ACE.

Another stimulation strategy called the envelope enhancement (EE) strategy focuses on onset enhancement (Geurts & Wouters, 1999; Koning & Wouters, 2012, 2016). Similar to the TESM strategy that enhances rapid increases in a channel's envelope, the EE strategy uses peak detection to enhance rapid increases in the envelopes. CI recipients in Koning and Wouters (2016) received a mean improvement of 25.6% in keyword understanding for sentences mixed with stationary, speech-shaped noise at −2 dB SNR, a 17.7% mean improvement at 2 dB SNR, and a 11.7% mean improvement at 6 dB SNR. For speech presented with an interfering talker, there was a significant improvement of 1 dB with this strategy.

To summarize the results from both the TESM and EE strategies, there seems to be a small, but significant benefit with the enhancement of onset information. In the present study, correcting gain errors in just the transition regions (i.e., $\hat{G}_g G_t \hat{G}_s$) with a priori information yielded, on average, a 0.5 dB and 2.5 dB SRT benefit for stationary and modulated noise, respectively. Therefore, the results in this study support those of the previous studies. One important difference between the $\hat{G}_g G_t \hat{G}_s$ condition in this study and the onset-enhancement-based stimulation strategies though is that manipulations have also occurred at the offsets as opposed to only at the onsets. In addition, nonideal noise reduction has been applied within the gap and speech regions, which likely introduces detrimental distortions that would otherwise not be present.

CI Simulation

The individual SRTs measured for unprocessed speech in the current study ranged between + 3 dB and + 11 dB. For comparison, SRTs for CI recipients often range anywhere between −5 dB and + 10 dB (see, e.g., Mauger, Warren, Knight, Goorevich, & Nel, 2014). However, the mean SRT for the Danish HINT corpus with the speech-shaped stationary noise is reported to be −2.52 dB for normal-hearing listeners (Nielsen & Dau, 2011). Thus, the CI simulation (i.e., vocoder processing together with N-of-M channel selection) increased the mean SRT by roughly 8.4 dB (i.e., from −2.5 dB to + 5.9 dB, albeit with different listeners). Although this elevation in SRT due to the CI simulation is higher than could be expected based on vocoder studies in the literature that do not include a channel selection stage, Qazi et al. (2013) have shown that the combination of vocoder processing with N-of-M processing elevated SRTs in their study with the Flemish/Dutch LIST by 6.3 dB (from −8.5 dB in the unprocessed condition to −2.2 dB in the simulated condition). Therefore, a change in SRT on the order of 8.4 dB is not unprecedented.

Dip Listening

Typically, normal-hearing listeners are able to extract information related to the target speech during temporal dips in the interfering noise (Duquesnoy, 1983; Festen & Plomp, 1990). This process is sometimes called *listening in the dips*. However, the normal-hearing listeners in the current study did not exhibit dip listening, as evidenced by the fact that SRTs were worse in the presence of modulated noise than in stationary noise in the unprocessed condition.

Bernstein and Grant (2009), among others, have demonstrated that hearing-impaired listeners exhibit difficulties with dip listening, and they have furthermore suggested that these difficulties may be attributed to the reduced fluctuating-masker benefit that is associated with the higher SNRs they require to obtain 50% speech recognition. Given that the normal-hearing listeners in the current study had elevated SRTs due to the CI simulation, one may have expected to observe *reduced* dip listening rather than a lack thereof. Based on this line of thought, inducing lower SRTs by, for example, using a simpler speech corpus or a different CI simulation technique, would lead to dip listening. However, Qin and Oxenham (2003) investigated dip listening for a range of vocoders, and by increasing the number of channels in their vocoder processing, they effectively lowered the range of SRTs observed among their listeners, and despite this lowering of the operating range of SNRs, single-talker interference was still more detrimental to speech recognition than the steady-state noise.

Fu and Nogaki (2005) further investigated whether the lack of dip listening due to vocoder processing is a result of the reduced number of spectral channels or the channel interactions. They found that, as long as the spectral channels in their vocoder did not overlap, the normal-hearing listeners were able to obtain a significant masking release; however, whenever crossover between the carrier bands was introduced, masking release was absent. Therefore, the lack of dip listening in the current study can likely be attributed specifically to the presence of crossover between carrier bands in the vocoder.

Visual Cues

Bernstein and Grant (2009) showed that both normal-hearing and hearing-impaired listeners obtain a significant improvement in their ability to listen in the dips of fluctuating maskers when they are presented with both audio and visual cues when compared with audio cues alone. Their study, among others, highlights the importance of visual cues, as well as the interaction between audio and visual cues, in the perception of speech in noise in more realistic environments. Relatively little is known, however, about the influence of visual cues on speech perception specifically in CI recipients. Depending on factors such as whether a recipient was pre- or postlingually deafened, how early a recipient was implanted postdeafening, and whether a recipient can lip-read, the integration of and reliance on visual cues can vary drastically among CI recipients (see, e.g., Champoux, Lepore, Gagné, & Théoret, 2009; Schorr, Fox, van Wassenhove, & Knudsen, 2005). It is clear nonetheless that a deeper understanding of the interaction between audio and visual cues in general will become increasingly more relevant as focus turns to the investigation of more realistic listening scenarios.

An interesting extension of the current investigation would be to identify whether visual cues influence the contribution of each of the different temporal regions of speech to intelligibility. Such an investigation would help to identify the relative contribution of each of these audio cues in more realistic listening scenarios. For example, it could be expected that an enhancement of the temporal cues which aid in the segmentation of words would provide a smaller benefit to CI recipients when the audio cues are presented in combination with visual cues. This hypothesis is supported by the fact that Dorman et al. (2016) have shown that CI recipients obtain improved lexical segmentation when they are provided with visual information alongside the acoustic information.

Implications and Limitations

The results in the current study provide a framework for hypothesizing how CI recipients would be affected by noise reduction errors in the speech, gap, and transition regions. One of the primary conclusions by Qazi et al. (2013) was that CI recipients can tolerate significantly less noise in the gap regions when compared with their normal-hearing counterparts. Therefore, a logical hypothesis would be that CI recipients would actually benefit more from minimizing noise reduction errors in the gaps between speech segments than the normal-hearing listeners in this study did. On the other hand, because CI recipients rely so heavily on the low-frequency amplitude modulations of speech, presumably more so than normal-hearing listeners, it is likely that the magnitude of the benefit from the suppression of noise in the gaps will be substantially smaller than suggested by Qazi et al. (2013) in realistic algorithms, given that realistic algorithms will be unable to detect onsets and offsets precisely. However, it is important to test with CI recipients rather than with normal-hearing listeners and a vocoder simulation, as results obtained with CI simulations can be misleading.

An additional, yet potentially important, limitation of the experimental design in this study is that the sentences were segmented into speech, gap, and transitions regions using a heuristically designed method. The transition regions were fixed to be 20 ms in duration, but even short-duration speech signals range in duration from just 5 ms to as long as 50 ms (Vandali, 2001). Therefore, the current segmentation method may have led to the labeling of transition regions that did not accurately reflect the location of the true transition regions and, thereby, may have led to an over- or underestimation of the impact of errors in these regions. On the other hand, segmenting sentences in this way facilitated comparisons with the segmentation of sentences in Qazi et al. (2013), which was largely advantageous.

Conclusion

Qazi et al. (2013) suggested that the most important factor for attaining high speech intelligibility in noise with CI listeners is to preserve the low-frequency amplitude modulations of speech across time and frequency in the stimulation patterns. In their study, both normal-hearing listeners tested with a vocoder simulation and CI recipients achieved the largest improvement in intelligibility when there was no stimulation in the gaps between speech segments. In a realistic algorithm, however, the identification of these regions will be imperfect, and the results from the current study suggest that the benefit of attenuating stimulation during speech gaps is largely diminished when the transitions between the speech and speech gaps are distorted.

Although some listeners in the current study obtained a large benefit in modulated noise with the minimization of gain errors in the gaps while errors in the transitions remained present, their intelligibility improvement can likely be attributed to the fact that they could listen in the dips for salient onset cues. Because CI recipients are typically less able to listen in the dips (Nelson, Jin, Carney, & Nelson, 2003), this benefit is likely to be less pronounced in CI listeners. Therefore, removing stimulation in the speech gaps may itself not be such a key component to improving speech intelligibility in noise with CI recipients. Instead, a more effective goal may be to identify the boundaries between the speech and gaps so that, while minimizing the stimulation of noise-dominated channels in the gaps, it will also be possible to deliver salient cues related to the transients. These two components together seem to contribute substantially to understanding speech in noise, at least with the normal-hearing listeners tested in the current study using speech degraded by a vocoder simulation.

Acknowledgments

The authors would like to thank all of the subjects who participated in the experiment, as well as Kristine Aavild Juhl and Rasmus Malik Thaarup Høegh, for helping to conduct the testing.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Danish Council for Independent Research with grant number DFF-5054-00072.

ORCID iD

Abigail A. Kressner http://orcid.org/0000-0003-4274-3948
Tobias May http://orcid.org/0000-0002-5463-5509

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bernstein, J. G., & Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *125*(5), 3358–3372.

Bhattacharya, A., Vandali, A., & Zeng, F.-G. (2011). Combined spectral and temporal enhancement to improve cochlear-implant speech perception. *The Journal of the Acoustical Society of America*, *130*(5), 2951–2960.

Champoux, F., Lepore, F., Gagné, J.-P., & Théoret, H. (2009). Visual stimuli can impair auditory processing in cochlear implant users. *Neuropsychologia*, *47*(1), 17–22.

Cohen, I. (2003). Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Speech Audio Process*, *11*(5), 466–475.

Dorman, M. F., Liss, J., Wang, S., Berisha, V., Ludwig, C., & Natae, S. C. (2016). Experiments on auditory-visual perception of sentences by users of unilateral, bimodal, and bilateral cochlear implants. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1505–1519.

Duquesnoy, A. (1983). Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons. *The Journal of the Acoustical Society of America*, *74*(3), 739–743.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*(4), 1725–1736.

Fu, Q.-J., & Nogaki, G. (2005). Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing. *Journal of the Association for Research in Otolaryngology*, *6*(1), 19–27.

Geurts, L., & Wouters, J. (1999). Enhancing the speech envelope of continuous interleaved sampling processors for cochlear implants. *The Journal of the Acoustical Society of America*, *105*(4), 2476–2484.

Goehring, T., Bolner, F., Monaghan, J. J., van Dijk, B., Zarowski, A., & Beeck, S. (2017). Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing Research*, *344*, 183–194.

Hersbach, A. A., Grayden, D. B., Fallon, J. B., & McDermott, H. J. (2013). A beamformer post-filter for cochlear implant noise reduction. *The Journal of the Acoustical Society of America*, *133*(4), 2412–2420.

Hochberg, I., Boothroyd, A., Weiss, M., & Hellman, S. (1992). Effects of noise and noise suppression on speech perception by cochlear implant users. *Ear and Hearing*, *13*(4), 263–271.

Holden, L. K., Vandali, A. E., Skinner, M. W., Fourakis, M. S., & Holden, T. A. (2005). Speech recognition with the advanced combination encoder and transient emphasis spectral maxima strategies in nucleus 24 recipients. *Journal of Speech, Language, and Hearing Research*, *48*(3), 681–701.

Holube, I., Fredelake, S., Vlaming, M., & Kollmeier, B. (2010). Development and analysis of an international speech test signal (ISTS). *International Journal of Audiology*, *49*(12), 891–903.

Hu, Y., & Loizou, P. C. (2007). A comparative intelligibility study of single-microphone noise reduction algorithms. *The Journal of the Acoustical Society of America*, *122*(3), 1777–1786.

Koning, R., & Wouters, J. (2012). The potential of onset enhancement for increased speech intelligibility in auditory prostheses. *The Journal of the Acoustical Society of America*, *132*(4), 2569–2581.

Koning, R., & Wouters, J. (2016). Speech onset enhancement improves intelligibility in adverse listening conditions for cochlear implant users. *Hearing Research*, *342*, 13–22.

Kressner, A. A., May, T., Høegh, R. M. T., Juhl, K. A., Bentsen, T., & Dau, T. (2017). Investigating the effects of noise-estimation errors in simuated cochlear implant speech intelligibility. *Proceedings of the International Symposium on Auditory and Audiological Research*, *6*, 295–302.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Kuznetsova, A., Christensen, R. H., Bavay, C., & Brockhoff, P. B. (2015). Automated mixed ANOVA modeling of sensory and consumer data. *Food Quality and Preference*, *40*, 31–38.

Lenth, R. (2018). *Emmeans: Estimated marginal means, aka least-squares means*. Retrieved from https://cran.r-project.org/web/packages/emmeans.

Mauger, S. J., Arora, K., & Dawson, P. W. (2012). Cochlear implant optimized noise reduction. *Journal of Neural Engineering*, *9*(6), 1–9.

Mauger, S. J., Warren, C. D., Knight, M. R., Goorevich, M., & Nel, E. (2014). Clinical evaluation of the Nucleus 6 cochlear implant system: Performance improvements with SmartSound iQ. *International Journal of Audiology*, *53*(8), 564–576.

Nelson, P. B., Jin, S.-H., Carney, A. E., & Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *113*(2), 961–968.

Nielsen, J. B., & Dau, T. (2011). The Danish hearing in noise test. *International Journal of Audiology*, *50*(3), 202–208.

Qazi, O. U., van Dijk, B., Moonen, M., & Wouters, J. (2013). Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility. *Hearing Research*, *299*, 79–87.

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *The Journal of the Acoustical Society of America*, *114*(1), 446–454.

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(51), 18748–18750.

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, *34*(4), 216–221.

Spriet, A., Van Deun, L., Eftaxiadis, K., Laneau, J., Moonen, M., Van Dijk, B., . . . Wouters, J. (2007). Speech understanding in background noise with the two-microphone adaptive beamformer beam^{TM} in the nucleus freedom^{TM} cochlear implant system. *Ear and Hearing*, *28*(1), 62–72.

Vandali, A. E. (2001). Emphasis of short-duration acoustic speech cues for cochlear implant users. *The Journal of the Acoustical Society of America*, *109*(5), 2049–2061.

Yund, E. W., & Woods, D. L. (2010). Content and procedural learning in repeated sentence tests of speech perception. *Ear and Hearing*, *31*(6), 769–778.