Investigating the effects of noise-estimation errors in simulated cochlear implant speech intelligibility

ABIGAIL ANNE KRESSNER*, TOBIAS MAY, RASMUS MALIK THAARUP HØEGH, KRISTINE AAVILD JUHL, THOMAS BENTSEN, AND TORSTEN DAU

Hearing Systems, Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

A recent study suggested that the most important factor for obtaining high speech intelligibility in noise with cochlear implant recipients is to preserve the low-frequency amplitude modulations of speech across time and frequency by, for example, minimizing the amount of noise in speech gaps. In contrast, other studies have argued that the transients provide the most information. Thus, the present study investigates the relative impact of these two factors in the framework of noise reduction by systematically correcting noise-estimation errors within speech segments, speech gaps, and the transitions between them. Speech intelligibility in noise was measured using a cochlear implant simulation tested on normal-hearing listeners. The results suggest that minimizing noise in the speech gaps can substantially improve intelligibility, especially 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 correct identification of the boundaries between speech segments and speech gaps is the most important factor in maintaining high intelligibility in cochlear implants. Knowing the boundaries will make it possible for algorithms to both minimize the noise in the gaps and enhance the low-frequency amplitude modulations.

INTRODUCTION

Hochberg *et al.* (1992) reported that cochlear implant (CI) recipients typically had thresholds for speech reception in noise that were 10 to 25 dB poorer than normal-hearing listeners. Since then, there has been extensive research in the development of noise reduction algorithms and sound coding strategies in order to obtain an increased robustness to noise. Within this effort, speech intelligibility improvements have been demonstrated by applying both single-microphone noise reduction (e.g., Mauger *et al.*, 2012) and multi-microphone directional noise reduction (e.g., Hersbach *et al.*, 2013). In contrast, although many sound coding strategies have been proposed over the last few decades, none have been able to consistently produce a measured improvement in speech intelligibility in noisy environments over well-established strategies like continuous interleaved sampling (CIS) and the Advanced Combination

*Corresponding author: aakress@elektro.dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017.
 ISBN: 978-87-990013-6-1.

Encoder (ACE^{TM}, Cochlear Ltd., New South Wales, Australia). One potential reason for this lack of success is that relatively little is known about how different kinds of errors in CI stimulation specifically influence speech intelligibility outcomes.

In an effort to improve this understanding, Qazi *et al.* (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: (1) noise-related stimulation can fill the gaps between speech segments, (2) stimulation levels during speech segments can become distorted, and (3) channels which are dominated by noise can be selected for stimulation instead of channels which are dominated by speech. In order to measure the effect of each of these, 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 normal-hearing listeners with a vocoder, and measured speech intelligibility. 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 by, for example, minimizing 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. Accordingly, they demonstrated that enhancing speech onset cues alone improves speech intelligibility in CI recipients (Koning and 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, because they always identified these segments via ideal onset and offset detection. Therefore, 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 be able to perfectly identify the boundaries between speech segments and speech gaps.

The main motivation of the present study was to systematically quantify the relative impact of realistic noise-estimation errors occurring within speech segments, speech gaps, and speech transients. Specifically, this study investigated these distortions using a basic CI vocoder simulation tested with normal-hearing listeners, which provides insight into the impact of the spectro-temporal degradation in isolation from an impaired auditory system.

METHODS

A CI with an N-of-M strategy such as ACE 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, and then a Hann window and the short-time discrete Fourier Transform (STFT) was applied with $K = 128$ points to obtain the time-frequency representation of speech, $X(k,l)$. The STFT magnitudes were then combined into $M = 22$ channels using non-overlapping rectangular weights with spacing that matches Cochlear Ltd.'s (New South Wales, Australia) sound processor in order to obtain the time-frequency representation $X(m,l)$, where m represents the channel index and l represents the frame index. A new frame was calculated every 1 ms.

In the Qazi *et al.* (2013) study, sentences were divided temporally into speech segments and speech gaps. Artificial sequences were then synthesized by copying segments from the clean speech sequence and noisy speech sequence. In the present study, sentences were instead divided into three temporal regions (i.e., speech segments, speech gaps, and speech transitions). This protocol allows for the separation of the reduction of noise in the speech gaps from the encoding of the transitions. In order to do this segmentation, broadband channel activity, $A(l)$ was defined for each frame as the number of channels containing speech above a threshold:

$$A(l) = \sum_{m=1}^{M} T_{\lambda}\left(X(m,l)\right), \qquad \text{(Eq. 1)}$$

where the function $T_{\lambda}(\cdot)$ performs element-wise thresholding and returns a value of one for elements that are above 25 dB sound pressure level (i.e., the default threshold level in ACE). As in the Qazi *et al.* (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 close in time to another speech segment were merged together. The merging prevented rapid switches between speech and non-speech 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 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 a clean sentence is shown in Fig. 1(a), with the temporal regions indicated by the underlying shading. The 20-ms duration for the transition region was heuristically chosen in order to ensure the transition regions were long enough to be perceptible, but short enough to maintain a segmentation that was still comparable to the segmentation in Qazi *et al.* (2013).

Whereas Qazi *et al.* (2013) primarily manipulated channel selection and current levels within each temporal region in order to investigate the impact of noise-induced errors in sound coding strategies, the present study manipulated the gains that are applied in a preceding noise reduction stage in order to investigate the impact of noise-induced errors in noise reduction algorithms. Therefore, instead of synthesizing stimulation patterns from the clean and noisy speech, artificial gain matrices were synthesized

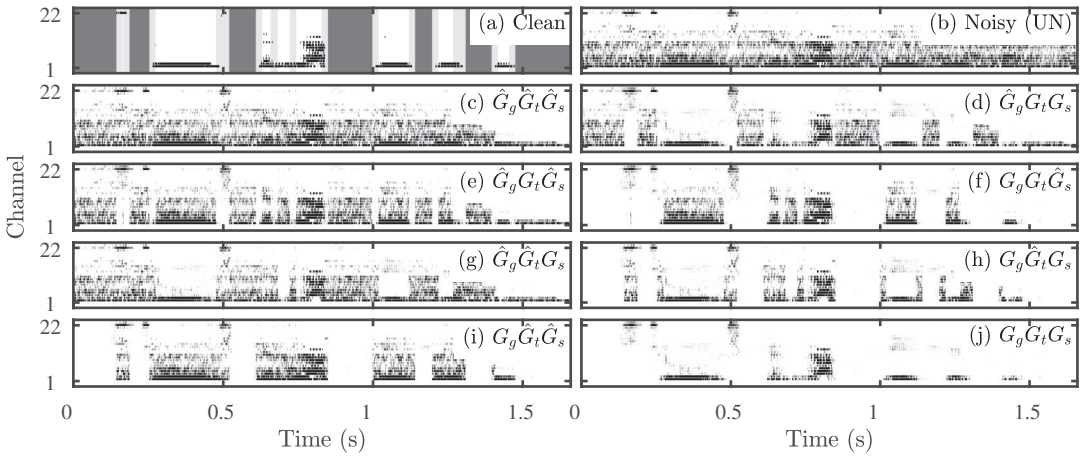

Fig. 1: Electrodograms showing (a) stimulation levels above threshold for the Danish sentence, *Stuen skal nok blive hyggelig* and (b-j) unthresholded levels for the same sentence mixed with speech-weighted noise and then de-noised using the indicated gain matrix. Speech segments, transitions, and gaps are identified in (a) by the white, light gray, and dark gray shading, respectively.

from either the *a priori* local signal-to-noise ratios (SNRs) or from estimated SNRs using a CI-optimized noise reduction algorithm (Mauger *et al.*, 2012). An underlying assumption in this study then is that a maxima selection strategy, such as ACE, will stimulate the correct set of channels if it chooses channels from a representation that has been sufficiently de-noised.

The following general signal model was thereby considered: $Y(k,l) = X(k,l) + D(k,l)$, with $X(k,l)$ 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 minimum controlled recursive algorithm (Cohen, 2003). $\hat{D}(m,l)$ was then computed 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). From $\hat{\xi}(m,l)$, gains $\hat{G}(m,l)$ were obtained using the CI-optimized gain function (Mauger *et al.*, 2012),

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

Additionally, the ideal gains $G(m,l)$ were computed using the *a priori* instantaneous signal-to-noise ratio $\xi(m,l)$.

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 understand

the impact of errors specifically in the speech gaps, gains from $G(m,l)$ were applied to the noisy signal $Y(m,l)$ in all of the speech gaps, and 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 signify that the estimated gains were corrected in the speech gaps, but not in the transitions and the speech segments. Accordingly, the condition $\hat{G}_gG_t\hat{G}_s$ signifies 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. An acoustic signal was then constructed from the stimulation sequence using a 22-channel noise vocoder. Figure 1(b) shows the sequences for a noisy version of the sentence in Fig. 1(a), and Figs. 1(b-j) show the sequences after de-noising with each type of gain matrix. A visual comparison between Figs. 1(c) and 1(j) highlights the extent of the estimation errors in $\hat{G}_g\hat{G}_t\hat{G}_s$. Subsequently, the remaining figures in the left column contain the stimulation patterns for the conditions where just one of the temporal regions of the gain matrix have been corrected. Lastly, the remaining plots in the right column each show the stimulation patterns for the conditions where two of the temporal regions have been corrected.

Speech intelligibility was evaluated in six participants by obtaining speech reception thresholds (SRTs) of sentences in noise via the Danish hearing in noise test (HINT) (Nielsen and Dau, 2011). Through an adaptive procedure, HINT determines the SNR at which the participants were able to understand 50% of the sentence material. Testing was carried out in a double-walled booth, using equalized Sennheiser HD-650 circumaural headphones. Participants were at least 18 years of age, had audiometric thresholds of less than or equal to 20 dB HL 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.

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 stationary speech-weighted noise (Nielsen and Dau, 2011) or the International Speech Test Signal (Holube *et al.*, 2010) (i.e., a modulated noise that is speech-like but unintelligible), and then testing proceeded with the other. The presentation order of the noise types was counterbalanced across participants. There were eight noise reduction conditions, and together with the reference, noisy condition (i.e., unity gains), there were nine test conditions for each noise type. Two SRTs were collected per condition, and the mean of the two was used for analysis. For two of the participants, only one SRT was collected for a small subset of the test conditions, and therefore, these three data points did not include test-retest averaging. None of these points were outliers. Since the Danish HINT contains only ten lists, participants heard the first nine lists multiple times, in a random order each time.

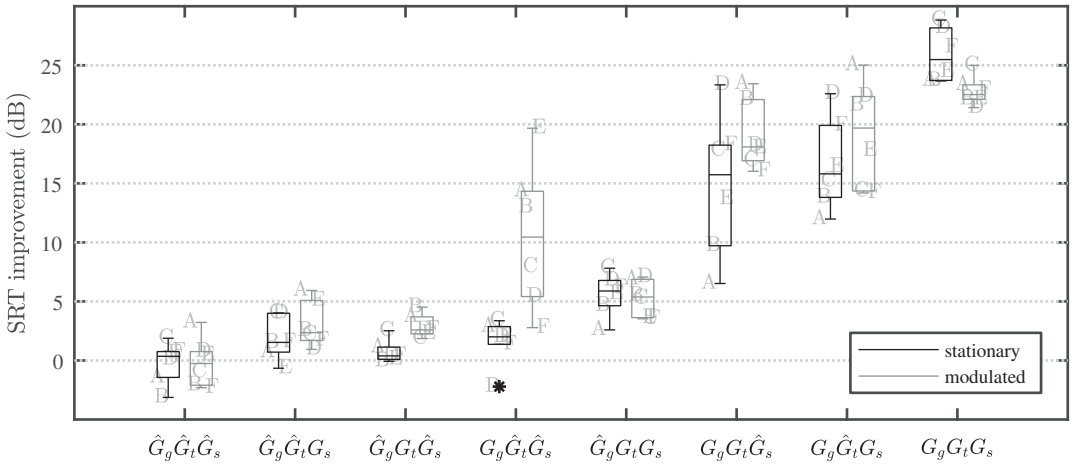

Fig. 2: Speech reception threshold (SRT) improvements relative to the reference noisy condition. Box plots show the 25^{th}, 50^{th}, and 75^{th} percentiles, together with whiskers that extend to extreme data points not considered outliers. Outliers are marked with an asterisk. Letters correspond to individual participants.

RESULTS

Figure 2 shows the improvement in SRT for each individual relative to their average SRT in the reference noisy condition. Because normal-hearing listeners generally do not benefit from single-microphone noise reduction algorithms (Hu and Loizou, 2007), it is not surprising that the CI-optimized noise reduction algorithm (i.e., $\hat{G}_g\hat{G}_t\hat{G}_s$) did not provide an SRT improvement, on average. Similarly, it is not surprising that the average SRT improvement was around 25 dB when *a priori* information about the local SNRs was used (i.e., $G_gG_tG_s$), as this was the maximum possible benefit given the constraints of the testing software.

Focusing first on the impact of errors in the speech gaps (i.e., $G_g\hat{G}_t\hat{G}_s$ versus $\hat{G}_g\hat{G}_t\hat{G}_s$), SRTs tended to improve in the stationary noise, and substantially improved in the modulated noise —though to varying degrees across participants— when the errors in the gaps were removed. This result suggests that minimizing noise-dominated stimulation in the speech gaps is an important factor for improving intelligibility, which is in line with the conclusions in Qazi *et al.* (2013).

However, in comparison to correcting the errors in the speech gaps, correcting errors in the speech segments (i.e., $\hat{G}_g\hat{G}_tG_s$) yielded, on average, a smaller SRT benefit, especially with regard to the modulated noise type. In a similar manner, correcting gain errors in the transition regions (i.e., $\hat{G}_gG_t\hat{G}_s$) yielded a relatively small SRT benefit, particularly in the stationary noise. This result was unexpected, however, given that the previous body of literature suggests that increased gain in transition regions (e.g., Vandali, 2001), or specifically at the onsets (e.g., Koning and Wouters,

2016), significantly improves speech intelligibility for CI recipients in both stationary noise and in the presence of a competing talker. Thus, it is possible that CI listeners rely more on these cues than normal-hearing listeners with a vocoder simulation. Alternatively, it may be that the detrimental effect of the sudden changes in gains in these stimuli were larger than the benefit of encoding the transitions correctly.

Despite the relatively small impact when only correcting gain errors in the transitions alone, the combination of correcting errors in the transitions and correcting errors in the gaps resulted in substantial improvements in SRTs. Furthermore, the benefit from correcting gain errors in both of these regions is much larger than the sum of the benefit from each in isolation. This result suggests that there is a strong interaction between gain errors in speech gaps and gain errors in the transitions, which implies that the potential benefit of minimizing stimulation from noise-dominated channels in speech gaps largely depends on how accurately the boundaries between the gaps and segments of speech are encoded.

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 CI recipients and normal-hearing listeners tested with a vocoder simulation 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 noise-dominated stimulation presented in 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 very large benefit in modulated noise with the minimization of gain errors in the gaps, even when errors in the transitions remained present, their intelligibility improvement is likely attributed to the fact that they could listen in the dips for salient onset cues. Since CI recipients are typically less able to listen in the dips (Nelson *et al.*, 2003), this benefit is likely to be less pronounced in CI listeners. Therefore, removing stimulation in the speech gaps may not itself be such a key component to improving speech intelligibility in noise in CI listeners. 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 the most to understanding speech in noise, at least with normal-hearing listeners tested with speech degraded by a vocoder simulation.

ACKNOWLEDGMENT

This work was supported by the Danish Council for Independent Research (DFF) with grant number DFF-5054-00072.

REFERENCES

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

Hersbach, A.A., Grayden, D.B., Fallon, J.B., and McDermott, H.J. (**2013**). "A beamformer post-filter for cochlear implant noise reduction," J. Acoust. Soc. Am., **133**, 2412-2420.

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

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). "Development and analysis of an international speech test signal (ISTS)," Int. J. Audiol., **49**, 891-903.

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

Koning, R., and Wouters, J. (**2012**). "The potential of onset enhancement for increased speech intelligibility in auditory prostheses," J. Acoust. Soc. Am., **132**, 2569-2581.

Koning, R., and Wouters, J. (**2016**). "Speech onset enhancement improves intelligibility in adverse listening conditions for cochlear implant users," Hear. Res., **342**, 13-22.

Mauger, S.J., Arora, K., and Dawson, P.W. (**2012**). "Cochlear implant optimized noise reduction," J. Neural Eng., **9**, 1-9.

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

Nielsen, J.B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol., **50**, 202-208.

Qazi, O. ur R., van Dijk, B., Moonen, M., and Wouters, J. (**2013**). "Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility," Hear. Res., **299**, 79-87.

Vandali, A.E. (**2001**). "Emphasis of short-duration acoustic speech cues for cochlear implant users," J. Acoust. Soc. Am., **109**, 2049-2061.