Simple spectral subtraction method enhances speech intelligibility in noise for cochlear implant listeners

MATTHIAS LEIMEISTER[1,2], CSANÁD EGERVÁRI[1,3], FELIX KUHNKE[1,2], ANJA CHILIAN[1,4,5], CHARLOTT VOIGT[1], AND TAMAS HARCZOS[1,2,5,*]

[1] *Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany*

[2] *Institute for Media Technology, Faculty of Electrical Engineering and Information Technology, Ilmenau University of Technology, Ilmenau, Germany*

[3] *Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary*

[4] *Institute of Biomedical Engineering and Informatics, Faculty of Computer Science and Automation, Ilmenau University of Technology, Ilmenau, Germany*

[5] *Cochlear-Implant Rehabilitationszentrum Thüringen, Erfurt, Germany*

It has been demonstrated that while clean speech is well intelligible by most cochlear implant (CI) listeners, noise quickly degrades speech intelligibility. To remedy the situation, CI manufacturers integrate noise reduction (NR) algorithms (often using multiple microphones) in their CI processors, and they report that CI users benefit from this measure. We have implemented a single-microphone NR scheme based on spectral subtraction with minimum statistics to see if such a simple algorithm can also effectively increase speech intelligibility in noise. We measured speech reception thresholds using both speech-shaped and car noise in 5 CI users and 23 normal-hearing listeners. For the latter group, CI hearing was acoustically simulated. In case of the CI users, the performance of the proposed NR algorithm was also compared to that of the CI processor's built-in one. Our NR algorithm enhances intelligibility greatly in combination with the acoustic simulation regardless of the noise type; these effects are highly significant. For the CI users, trends agree with the above finding (for both the proposed and the built-in NR algorithms), however, due to low sample number, these differences did not reach statistical significance. We conclude that simple spectral subtraction can enhance speech intelligibility in noise for CI listeners and may even keep up with proprietary NR algorithms.

INTRODUCTION

Signal processing chains of modern cochlear implant (CI) processors (like the Nucleus® CP910 from Cochlear™ or the Naída CI *Q70* from Advanced Bionics) include noise reduction (NR) methods to enhance speech perception in noise. However, for studies involving novel speech processing strategies, the elements of

*Corresponding author: tamas.harczos@gmail.com

Proceedings of ISAAR 2015: Individual Hearing Loss – Characterization, Modelling, Compensation Strategies. 5th symposium on Auditory and Audiological Research. August 2015, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C. Dalsgaard, L. Tranebjærg, and T. Andersen. ISBN: 978-87-990013-5-4. The Danavox Jubilee Foundation, 2015.

the CI's built-in signal processing chain are typically not available. Since our research plans include the testing of novel strategies combined with noise reduction, we created our own plain NR implementation, which we will abbreviate as PNR all through this document. This paper describes the functional principle of PNR and elaborates on the study we did to test PNR with CI users and normal-hearing (NH) listeners.

METHODS

Noise reduction

PNR is based on a single-microphone spectral subtraction algorithm that was proposed by Martin (1994). The first algorithm of that kind was introduced by Boll (1979) and many variations are widely used in communications and audio processing. Given a speech signal that is corrupted by additive noise, spectral subtraction aims at estimating the magnitude power of the noise spectrum. By applying one of several subtraction rules on the frames of a short time Fourier transform (STFT), this noise estimate is subtracted from the mixture leading to an approximation of the clean sound. The resulting time domain signal is obtained by applying the overlap-add technique. The noise is commonly only estimated in the magnitude or power domain while the original phase values are not modified for the reconstruction. This is due to the observation that estimating the phase of the clean signal is not crucial for the intelligibility of the resulting output (Loizou, 2007).

Most variants of spectral subtraction use a speech activity detector in order to estimate the noise spectrum in speech pauses. However, this can be a source of error. If the detection does not work correctly, parts of the speech might contribute to the noise estimate and would be attenuated by the subtraction rule. The extension that is used in this work (Martin, 1994) circumvents this by estimating the noise spectrum at the minima of a smoothed power spectrum. Under the assumption that the noise can be observed in isolation within a certain search window, one arrives at a steadily updated noise floor.

More detailed, the subband signal power P_x is computed from an STFT that is smoothed along the time axis by a first order low-pass. The estimated minimum power P_{min} is computed as the minimum within a given search window. By multiplying with a correction parameter *omin* that accounts for bias in the minimum estimate one arrives at the estimated noise power $P_n = omin \cdot P_{min}$ (for details see Martin, 1994). Given the STFT $X(t,k)$ of the noisy signal with time index t and subband index k, the output $Y(t,k)$ is computed as

$$|Y(t,k)| = \begin{cases} \sqrt{subf \cdot P_n(t,k)} & \text{if } |X(t,k)| \cdot Q(t,k) \leq \sqrt{subf \cdot P_n(t,k)} \\ |X(t,k)| \cdot Q(t,k) & \text{else} \end{cases} \quad \text{(Eq. 1)}$$

where the spectral weighting factor Q is given as

$$Q(t,k) = 1 - \sqrt{osub \cdot \frac{P_n(t,k)}{|X(t,k)|^2}} \qquad \text{(Eq. 2)}$$

To improve the quality of the reconstructed signal, some more parameters have been introduced. Because the noise cannot be estimated perfectly, $Y(t,k)$ contains spectral peaks that change rapidly between frames. In the reconstruction, this leads to audible tonal artefacts with fast changing frequencies that are known as musical noise. To reduce those peaks, the noise power is over-estimated by the factor *osub*. As this can lead to very small and even negative values, the reconstructed signal is bounded from below by a noise floor that can be adjusted by the factor *subf*. For our experiments, the following parameters are used: *subf*=0.001, *osub*=5.5, and *omin*=0.4. An overview of the proposed NR system is shown in Fig 1.

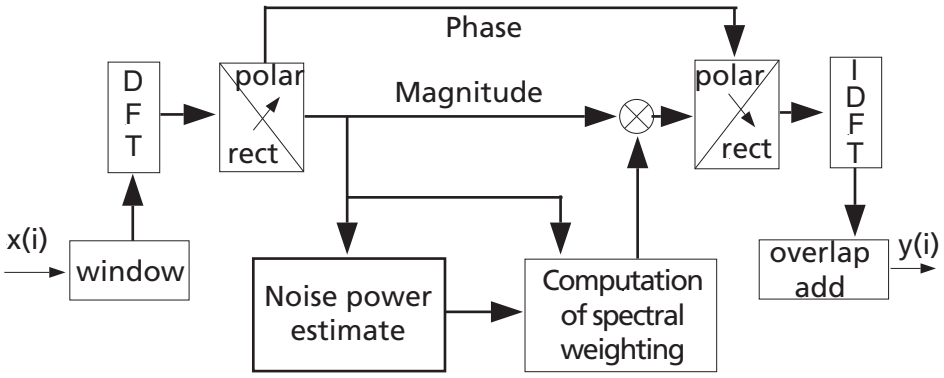

Fig 1: System overview of spectral subtraction algorithm (Martin, 1994).

Noise types

We used two noise types for our tests: a fluctuating speech-shaped noise (abbreviated *ols*) from the Oldenburg sentence test (OLSA, see Wagener *et al.*, 1999) and interior noise of a car driving steadily (abbreviated *car*). The two types of noise were normalized so that their A-weighted sound pressure level was the same (measured with a Phonic PAA3 handheld audio analyzer). The spectrograms are shown in Fig. 2.

Subjects

All subjects in this study were speaking German at the level of a native speaker. All CI users were fitted with a CP910 or CP920 processor using the ACE (Advanced Combination Encoder) strategy. Further details are listed in

Table **1**.

With the hearing subjects (age min=21, max=52.6, Md=27.9 years) we performed bilateral pure-tone audiometry at 500, 1000, and 2000 Hz, and calculated the pure-

tone average (PTA). Based on the results (PTA min=5, max=18.3, Md=8.3 dB HL), all subjects could be considered normal-hearing at the time of the study.

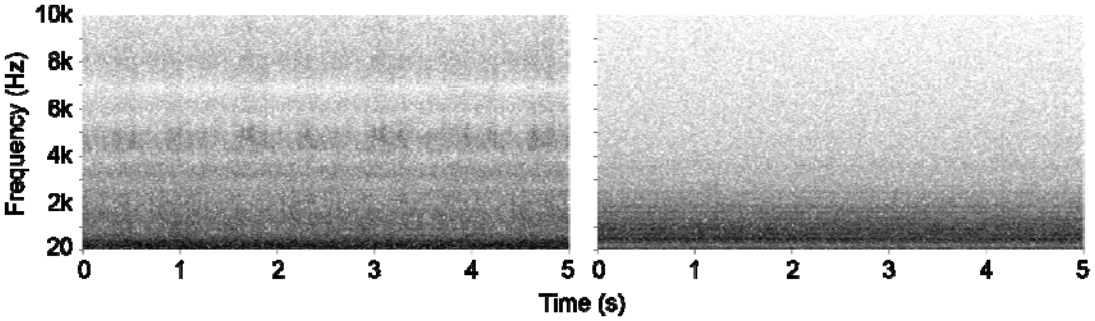

Fig. 2: Spectrograms of the two noise types: *ols* (left) and *car* (right).

Subject	Age (years)	CI (months)	Nucleus implant type	Lateralization	Note
S1	69	11	CI422	Bimodal	Ménière's disease
S2	42	86	CI24RE (CA)	Bilateral	Congenital
S3	13	87	CI24RE (CA)	Bilateral	Congenital
S4	63	14	CI422	Bimodal	Cause unknown
S5	22	15	CI24RE	Bilateral	Meningitis

Table 1: Demographics of cochlear-implanted subjects of the study.

Acoustic simulation of cochlear implant hearing

For normal-hearing listeners, we simulated cochlear implant hearing using the ACE strategy (channel stimulation rate of 900 pps with N=8 selected channels) as described in Chilian *et al.* (2011). Chilian *et al.* extended the signal synthesis of the general vocoder approach by combining two different carrier signals. As a result, both place and rate pitch mechanisms could be simulated. The algorithm also includes models of the electrode-tissue-interface and loudness perception.

In this study, we used the following parameters for the acoustic simulation: λ=8 mm (range of current spread), s=0.25 (synchronisation factor), PLL=300 Hz (phase-locking limit), α_p=0.75 mm, and α_s=4.5 mm (pass-band and stop-band filter bandwidths, respectively, as measured along the cochlea).

Test environment

Listening tests were performed in a soundproof booth (in accordance with the guidelines of ITU-R BS.1116) using a pair of Tapco S5 studio monitors (frequency response flatness: ±3 dB for 64 to 20000 Hz, according to the specifications) with an approximate loudspeakers-to-ears distance of 1 meter, driven by a Creative Sound

Blaster Live! 24-bit external (USB) sound card having excellent frequency response (within ±0.2 dB for 20 to 20000 Hz, measured with RightMark Audio Analyzer 5.5 using an external loopback, at 48-kHz sampling rate and 24-bit resolution).

Test procedure

During the listening tests, we captured the speech reception threshold (SRT) using the Oldenburg sentence test. However, we applied some modifications to the original test procedure, as follows. First, we embedded the sentences in either the original noise (*ols*) or the *car* noise. Second, speech and noise were not spatially separated, but mixed and played back from both loudspeakers. Third, the volume of each processed sentence was set so that the sound pressure level at the ears reached but did not exceed 70 dB SPL(A) during the playback (based on 100-ms measurement windows). Processing steps are shown in Fig. 3.

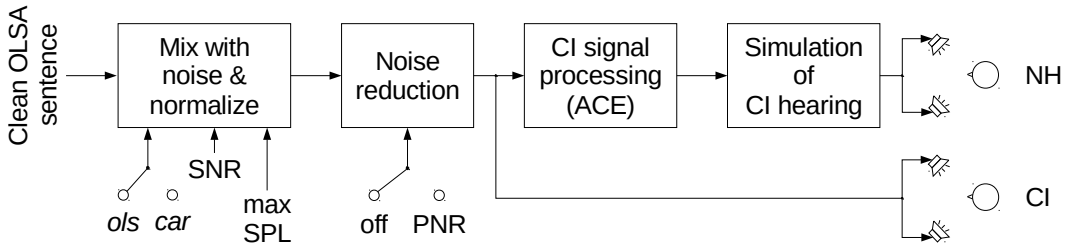

Fig. 3: System overview. NH and CI abbreviates normal-hearing and cochlear-implanted listener, respectively.

Before the listening tests, all subjects were made comfortable with OLSA: After the explanation of the test procedure, subjects examined a table showing all possible words of the OLSA sentences for 3 minutes, which was then followed by a warm-up list of 30 sentences (with feedback). The results of the warm-up list were excluded from any evaluation. During the subsequent actual tests, no feedback was provided. Between lists of 30 sentences, subjects could choose to have a short break for refreshments.

For each CI user, two variants of their everyday CI setting (map) were created: one with built-in NR disabled and one (otherwise identical copy) with built-in NR enabled. The CI processor's program was then switched between the OLSA sentence lists according to the desired test condition.

RESULTS

The results of the listening tests are displayed in Fig. 4. For the NH listeners, the evaluation of the speech reception threshold, which was measured using acoustic simulation of CI hearing with ACE, shows statistically significant benefit with PNR over non-processed noise corrupted speech (for both noise types; statistical test used: paired-sample Wilcoxon signed rank test). The improvements in SRT (median

differences) with PNR were 2.45 dB for *car* noise and 1.025 dB for *ols* noise. Speech intelligibility in the presence of *car* noise was better for both unprocessed and processed signals, which we will further elaborate on in the discussion section.

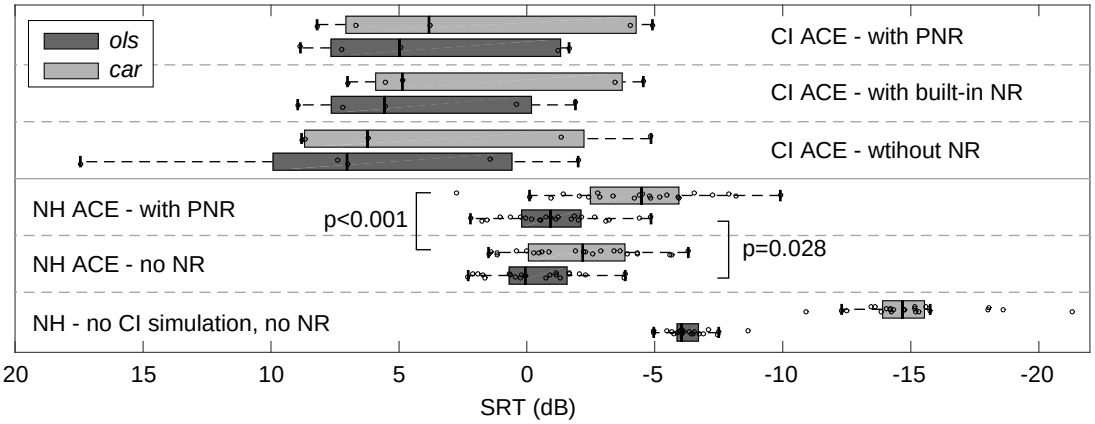

Fig. 4: Overview of the main results.

For CI users, a similar trend can be observed when comparing SRTs for speech corrupted with noise (CI ACE – without NR) and that for PNR applied to the signal before playback (CI ACE – with PNR). However, no statistical significance could be shown, which was likely due to the small sample number. The measured median improvements with PNR in SRT were 2.6 dB for *car* noise and 2.4 dB for *ols* noise, respectively. *Car* noise always allowed for better intelligibility than *ols* noise.

Finally, when comparing the CI's built-in noise reduction stage with PNR, the evaluation showed that both approaches result in improved SRT and that PNR seems on par with the built-in algorithm. The built-in method achieved median SRT improvements of 1.7 dB for *car* noise and 1.3 dB for *ols* noise. Again, to establish statistically significant results, a higher number of test subjects would be desirable.

DISCUSSION

Comparison of OLSA and car noise

Both for CI users and NH subjects, better intelligibility could be observed in the case of *car* noise. Further analysis of the disturbed signals and intermediate stages of the PNR algorithm showed that the spectral shape of the used *car* noise is less destructive to the clean speech signal than that of the speech-shaped noise. The energy of the *car* noise is more stable over time than in the case of *ols* noise so that spectral subtraction can distinguish better between the noise floor and the signal of interest. Furthermore, it is concentrated mostly outside of the individual bands that are important for speech intelligibility, whereas the *ols* noise is concentrated in exactly this region of the spectrum. Fig. 5 illustrates the effect of the two noise types.

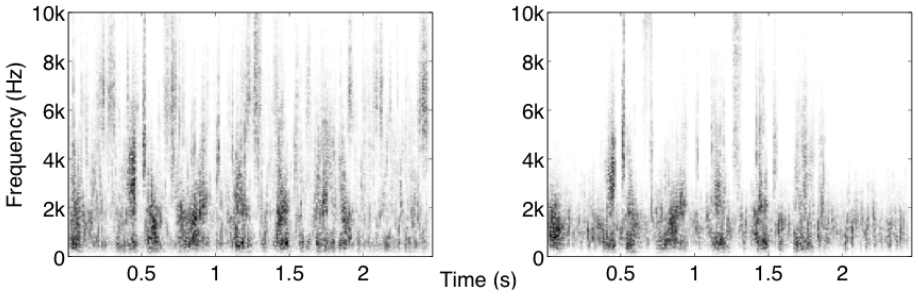

Fig. 5: Spectrograms of noisy speech after noise reduction and CI processing for *ols* noise (left) and *car* noise (right).

Influence of processing order

Commercial CI processors typically implement a pre-emphasis filter, which behaves like a high-pass filter on the audio input of the CI, just before further processing and filtering steps. One issue that needs further analysis is the order of pre-emphasis filtering and noise reduction. Because, for the sake of this study, it was not feasible to implement the PNR algorithm within the processing chain of commercial CI processors, the algorithm was applied to the audio signals before playing them back to the CI users. The built-in pre-emphasis filter of the CI was therefore applied after PNR processing. In a real-world scenario the noise reduction would run within the CI processing after the pre-emphasis.

Because PNR involves taking the maximum between the noise-subtracted spectrum and zero, the operation is non-linear and cannot be exchanged with the pre-emphasis filter, as in the case of linear time-invariant filters. However, in an informal analysis, the signals showed only minor differences when the two processing steps were swapped. Visual inspection of the resulting spectrograms suggests even a slightly improved noise reduction for the order of pre-emphasis followed by noise reduction followed by the CI processing, as can be seen in Fig. 6. Given this, it would be interesting to implement the proposed noise reduction stage within the CI hardware for further analysis.

Future directions

There are several possible extensions to the basic PNR algorithm that could improve its performance, such as multi-band processing and psycho-acoustically motivated spectral subtraction techniques (Zoghlami *et al.*, 2010). Among new approaches in the field of speech enhancement, deep neural networks become more and more popular. In recent studies they showed superior performance to classic methods as well as matrix factorization approaches (e.g., Liu *et al.*, 2014). In addition to using them as a pre-processing stage, such machine learning methods might provide the possibility to work well directly in the coded domain of the CI, which can be an interesting topic for future research.

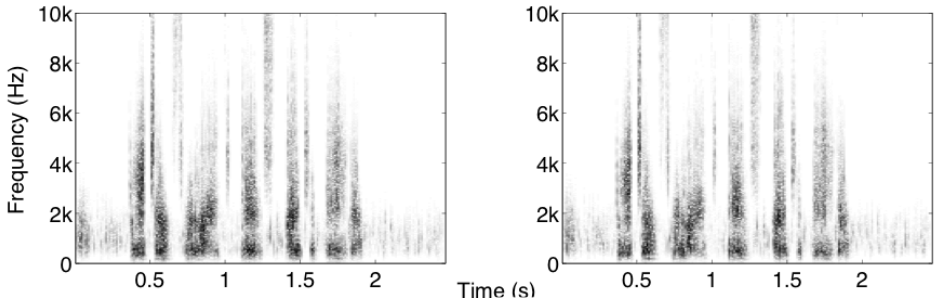

Fig. 6: Comparison of processing orders. Left: noise reduction – pre-emphasis – CI processing. Right: pre-emphasis – noise reduction – CI processing.

REFERENCES

Boll, S. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE T. Acoust. Speech, **27**, 113-120.

Chilian A., Braun E., and Harczos T. (**2011**). "Acoustic simulation of cochlear implant hearing," Proc. of 3rd International Symposium on Auditory and Audiological Research (ISAAR 2011) – *Speech Perception and Auditory Disorders*, Nyborg, Denmark, pp. 425-432.

Liu, D., Smaragdis P., and Kim M. (**2014**). "Experiments on deep learning for speech denoising," Proc Annual Conference of the International Speech Communication Association (INTERSPEECH), Singapore.

Loizou, P.C. (**2007**). *Speech Enhancement: Theory and Practice.* CRC Press: Boca Raton, FL, USA.

Martin, R. (**1994**). "Spectral subtraction based on minimum statistics," Proc. European Signal Processing Conference (EUSIPCO) '94, Edinburgh, Scotland, nr. 1, pp. 1182-1185.

Wagener, K., Kühnel, V., and Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests (Development and evaluation of a sentence test in German language I: Design of the Oldenburg sentence test)," Z. Audiol, **38**, 4-15.

Zoghlami, N., Lachiri Z., and Ellouze N. (**2010**). "Perceptually motivated generalized spectral subtraction for speech enhancement," Advances in Nonlinear Speech Processing, Lecture Notes Artif. Int., **5933**, 136-143.