INDIVIDUALIZING A MONAURAL BEAMFORMER FOR COCHLEAR IMPLANT USERS

Waldo Nogueira (1), Marta Lopez (1), Thilo Rode (1), Simon Doclo (2), Andreas Buechner (1)

(1) Dept. of Otolaryngology and Hearing4all, Medical University Hannover, Germany
(2) Dept. of Medical Physics and Acoustics and Hearing4all, University of Oldenburg, Germany

ABSTRACT

Speech intelligibility in noisy environments is still quite limited for cochlear implant (CI) users. Classical beamformers such as the Generalized Sidelobe Canceller (GSC) can provide large improvements in speech intelligibility for CI users. These algorithms have been adopted from hearing aids and multimedia applications into the CI field. However, their optimization taking into consideration the peculiarities of electrical hearing with a CI has not yet been completely investigated. This paper presents a novel method to optimize the performance of a GSC for each individual CI user. We show through a combination of objective and novel subjective measures, how much distortion can be tolerated by a CI user without decreasing speech intelligibility. Experimental results with 5 CI users show that a GSC delivering just noticeable distortion is the one maximizing speech intelligibility for CI users.

Index Terms— Beamformer, Cochlear Implant, Individualization, Speech Leakage, MUSHRA.

1. INTRODUCTION

A cochlear implant (CI) is a small electronic device that is surgically implanted into the inner ear and can restore the hearing of a profoundly deaf person. CI users need significantly higher signal-to-noise ratios (SNRs) to achieve the same speech intelligibility as normal-hearing listeners [1]. For this reason, speech enhancement techniques have emerged to improve the SNR in noisy acoustic conditions [2]. A well-known speech enhancement method is the use of adaptive directional microphones like Beamforming (BF). BF is a spatial filtering technique, which controls the directionality through the combination of multiple microphones. That way, the beampattern can be directed to the direction of the desired speech while signals coming from other directions are attenuated. The effectiveness of a CI highly varies for each patient and there is a wide variation of parameter calibration and sound perception for each individual CI user [3]. Although a large set of successful single- and multichannel noise reduction algorithms exist [4] [5] [2] [6] and almost all implants have some strategies implemented in their processors, noise reduction remains one of the big challenges of the acoustic processing in CIs. All algorithms and techniques have a good performance when the noise is coherent. However, their performance is reduced when the CI user is in a noisy environment with many incoherent noise sources, in reverberant rooms or in the presence of more interfering speech sources [7]. In this paper, we study a state-of-the-art monaural BF based on the Transfer Function Generalized Sidelobe Canceller TF-GSC [8] [9]. This algorithm has been shown to be more robust due to the use of relative transfer functions. However, a mismatch between the estimated transfer functions used by the BF and the real ones produce distortions in the desired target speech signal. We hypothesize that this distortion, which is higher in reverberant environments or in a multi-talker environment, decreases the speech intelligibility. The aim of this paper is to optimize the performance of the BF for each individual CI user. Section 2 presents the methodology, including the baseline BF and its individualization. Section 3 shows the experimental results in CI users and Section 4 presents the conclusions.

2. METHODS

2.1. Baseline Beamformer

Figure 1 shows the structure of a TF-GSC, which consists of 3 main parts: (1) A BF filter (W) including head related transfer functions (HRTFs) focusing the beam in the desired direction creating a speech reference; (2) A blocking matrix (B), which steers nulls in the direction of the speech source to create a noise reference; (3) Adaptive noise canceller (ANC), which reduces the noise components in the speech reference. In monaural CIs we are usually restricted to M=2

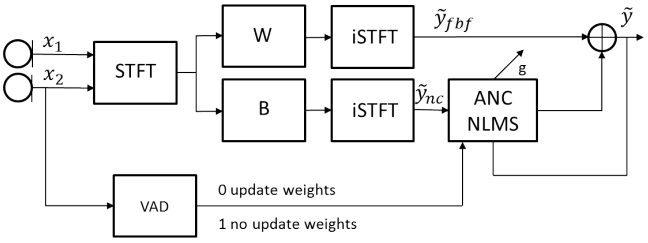

Fig. 1. Block diagram of a Generalized Sidelobe Canceller.

microphones placed in an ear piece positioned on the ear of a CI user. We assume that we have a single speech source $s(t)$ placed in a reverberant environment, the m^{th} microphone signal is given by:

$$x_m(t) = a_m(t) * s(t) + v_m(t) = z_m(t) + v_m(t), m = 1, 2. \quad (1)$$

where * is the convolution operator, a_m is the acoustic impulse response from the speech source signal to the m^{th} microphone, $z_m(t)$ and $v_m(t)$ are the clean speech and the noise components received at the m^{th} microphone. BFs are generally efficiently implemented in the frequency-domain using the Short Time Fourier Transform (STFT). The STFT is computed in frames of length N. Let $X_m(\omega)$, $A_m(\omega)$, $S(\omega)$, $Z_m(\omega)$ and $V_m(\omega)$ denote the STFTs of $x_m(t)$, $a_m(t)$, $s(t)$, $z_m(t)$ and $v_m(t)$, respectively. If N is sufficiently large, Eq. 1 can be approximated as:

$$\mathbf{X}(\omega) = \mathbf{A}(\omega)S(\omega) + \mathbf{V}(\omega) = \mathbf{Z}(\omega) + \mathbf{V}(\omega), \quad (2)$$

where, $\mathbf{X}(\omega) = [X_1(\omega), X_2(\omega)]^T$, and $\mathbf{A}(\omega)$, $\mathbf{Z}(\omega)$, and $\mathbf{V}(\omega)$ are defined similarly. It has been shown that a BF is equivalent

to a linear filter, i.e.: $\tilde{Y}(\omega) = \mathbf{H}^H(\omega)\mathbf{X}(\omega)$. The GSC [10] is an implementation of the minimum variance distortionless response (MVDR) beamformer [11] [12] and it can be viewed as a decomposition of the filter operation into two orthogonal subspaces, i.e.: $\mathbf{H}(\omega) = \mathbf{W}(\omega) - \mathbf{B}(\omega)\mathbf{G}(\omega)$, where $\mathbf{W}(\omega)$ is a fixed BF filter of size M, $\mathbf{B}(\omega)$ is a blocking matrix of size Mx$(M-1)$ that spans the null space of $\mathbf{A}(\omega)$ and $\mathbf{G}(\omega)$ represents the adaptive noise cancellation filter. The fixed BF can be designed such that the desired component in the output signal is equal to the speech source component m [13]: $Y_{FBF}(\omega) = \mathbf{W}(\omega)^H\mathbf{A}(\omega)S(\omega) \overset{constrain}{=} S(\omega)$, which is equivalent to $\mathbf{W}^H(\omega)\tilde{\mathbf{A}}(\omega) = 1$, where $\tilde{\mathbf{A}}(\omega)$ is defined as:

$$\tilde{\mathbf{A}}(\omega) = \Big[\frac{1}{A_1^*(\omega)} \quad \frac{1}{A_2^*(\omega)}\Big]^T. \tag{3}$$

Next the BF is normalized as:

$$\mathbf{W}_A(\omega) = \frac{\tilde{\mathbf{A}}(\omega)}{||\tilde{\mathbf{A}}(\omega)||^2}. \tag{4}$$

The blocking matrix is designed to block the target speech and create a noise reference. This is satisfied when $\mathbf{B}^H(\omega)\mathbf{A}(\omega) = 0$. For example, a blocking matrix can be created using estimated channel transfer-function ratios. For a two microphone BF, $\mathbf{B}(\omega)$ could be:

$$\mathbf{B}^H(\omega) = \Big[1 \quad -\frac{A_1^*(\omega)}{A_2^*(\omega)}\Big]^T. \tag{5}$$

The goal of the GSC is to compute an optimal adaptive noise cancellation (ANC) filter $G(\omega)$. This can be achieved by solving the following optimization problem:

$$\min_{g(\omega)} E\{|\mathbf{W}_A^{\mathbf{H}}(\omega)\mathbf{Y}(\omega) - G^*(\omega)\mathbf{B}^H(\omega)\mathbf{Y}(\omega)|^2\}. \tag{6}$$

The previous equation can be solved iteratively using the Normalized Least Mean Squares (NLMS) algorithm.

The output of the GSC can be written as:

$$\tilde{Y}(\omega) = \tilde{Y}_{FBF}(\omega) - \tilde{Y}_{NC}(\omega), \tag{7}$$

with,

$$\tilde{Y}_{FBF}(\omega) = \mathbf{W}_A^H(\omega)\mathbf{X}(\omega) \tag{8}$$

$$\tilde{Y}_{NC}(\omega) = G^*(\omega)\mathbf{B}^H(\omega)\mathbf{X}(\omega) \tag{9}$$

It has been shown [3] that the GSC output signal can be decomposed as:

$$\tilde{Y}(\omega) = \tilde{S}(\omega) - \tilde{S}_N(\omega) + \tilde{V}(\omega) - \tilde{V}_N(\omega), \tag{10}$$

where

$$\tilde{S}(\omega) = \mathbf{W}_A^H(\omega)\mathbf{A}(\omega)S(\omega),$$
$$\tilde{V}(\omega) = \mathbf{W}_A^H(\omega)\mathbf{V}(\omega),$$
$$\tilde{S}_N(\omega) = G^*(\omega)\mathbf{B}^H(\omega)\mathbf{A}(\omega)S(\omega),$$
$$\tilde{V}_N(\omega) = G^*(\omega)\mathbf{B}^H(\omega)\mathbf{V}(\omega),$$

are the BF speech component, BF noise component, speech leakage and residual noise component respectively. If $\mathbf{W}(\omega) \neq \mathbf{W}_A(\omega)$, the speech component is not perfectly dereverberated and therefore $\tilde{S}(\omega) \neq S(\omega)$. If $\mathbf{B}^H(\omega)\mathbf{A}(\omega) \neq 0$, the speech signal leaks into the noise reference and causes that $\tilde{S}_N(\omega) \neq 0$, which usually results in distortion of the target speech component. In this paper we want to characterize the perception of speech leakage and distortion tolerated by a CI user. Additionally we will investigate the optimal individual trade-off between speech distortion and noise cancellation.

2.1.1. VOICE ACTIVITY DETECTOR (VAD)

Several techniques have been proposed to limit the speech distortion. Some aim at reducing the speech leakage in the noise references e.g. by constructing a more robust BM [14] or using adaptive filters [13]. Another possibility is to limit the distorting effect of the remaining speech leakage components by using a voice activity detector (VAD) to update the ANC only during noise periods and fix the weights when there is speech [15]. We included a VAD to the TF-GSC to reduce the remaining speech distortion at the output. The performance of the BF is strongly influenced by the accuracy of the speech/non-speech classification of the VAD. In this paper we used a VAD implementation as proposed by [16]. We have chosen this algorithm because it is robust and allows us to control its accuracy by changing a single parameter, namely the speech probability threshold (SPT). The SPT is used to decide whether a given frame contains speech or not. Small SPT values make the VAD more conservative often labelling the signal as speech. We expect that when the speech is wrongly detected as noise, the ANC will adapt the desired signal producing more distortion at the BF output. In contrast, if the noise is misclassified as speech, the ANC will stop the adaptation, and noise that changes quickly in time will not be attenuated. Therefore, the SPT enables us to trade-off speech leakage and noise reduction.

2.2. Beamformer Individualization

In order to optimize the SPT parameter of the VAD such that it produces maximum noise reduction but with non-perceivable speech leakage distortion we designed the procedure presented in Figure 2.

Fig. 2. Process to individualize the SPT value for a BF.

Speech Database: The set of selected speech signals is the German HSM Sentence Test [16], which consists of 30 lists of 20 everyday sentences.

2.2.1. Environment Simulation

An office room was simulated generating three sources at -90^0, 0^0 and 90^0 at 1m distance from the center of the head of an artificial listener. All three babble sources had the same level. We used the HRTF dataset from [17] because it takes into consideration the effect of the head as in a conventional hearing aid. The three sources were presenting a 4-talker babble noise, uncorrelated between each source. The frontal speaker was also used to present the target speech signal. The input SNR at the microphones was adjusted at 0dB. Although the HRTFs provided three omnidirectional microphones (front, middle, back) signals, only the front and back microphones of the one ear, which are separated by 14.9 mm, were used.

2.2.2. Beamformer Implementation

The BF was based on the TF-GSC BF. It was implemented using 50% overlapping sine windows prior to applying STFT. The frame length was set to N=1024 and the sampling frequency f_s was 16 kHz. We averaged the anechoic HRTFs [17]. These filters were used in the frequency domain and applied to match the fixed BF and the

Table 1. Relation between speech distortion (SD) and noise reduction (NR) using babble noise in office room.

Babble SNR = 0 dB	**TF-GSC**	
SPT	**SD [dB]**	**NR [dB]**
0	0.21	-5.29
0.3	0.85	0.95
0.5	0.96	1.18
0.75	1.02	1.34
1	1.4	1.39

blocking matrix. Once the signal was filtered in the FFT domain they were transformed back into the time domain using overlap and add. Next ANC based on Normalized Least Mean Squares (NLMS) adaptive filter was applied to adaptively solve Eq. 6 in the time domain. The NLMS steps size (μ=0.01) was chosen such that it produced maximum SNR for an $SPT_{VAD} = 1$ in the simulated environment. It has to be noted that the NLMS could have been implemented in the frequency domain saving an FFT operation. The output of the BF was then fed directly to the input of a Nucleus CI which applied the ACE sound coding strategy [18].

2.2.3. Objective Measures

Speech Distortion (SD): Our aim is to quantify the SD at the BF output. The speech distortion is defined as:

$$SD = \sum_{\forall\omega} 10log_{10}\Big(\frac{P_S(\omega)}{P_{\tilde{S}}(\omega)}\Big)U(\omega), \tag{11}$$

where $P_{\tilde{S}}(\omega)$ and $P_S(\omega)$ are the PSD of $\tilde{S}(\omega)$ and $S(\omega)$ respectively. We weight the distortion measure according to the frequencies that are most important for speech intelligibility $U(\omega)$ [19].

Noise Reduction (NR):NR is evaluated as:

$$NR = \sum_{\forall\omega} 10log_{10}\Big(\frac{P_V(\omega)}{P_{\tilde{V}}(\omega)}\Big)U(\omega), \tag{12}$$

where $P_{\tilde{V}}(\omega)$ and $P_V(\omega)$ are the PSD of $\tilde{V}(\omega)$ and $V(\omega)$ respectively. Table 1 presents the SD and NR obtained for different SPT_{VAD} values and averaged for the whole HSM test.

2.3. Subjective Experiments (*jnd*SPT)

The goal is to subjectively characterize for each individual the trade-off between SD and NR for different SPT_{VAD} values. For this we measured the just noticeable (jnd) distortion depending on the SPT_{VAD}. We used a 3-Alternative Forced Choice (3-AFC) procedure with adaptive control of the SPT_{VAD} parameter. Three alternative choices were presented randomly, where two of them are the clean reference speech signal while the other is one of the test signals processed with a given SPT. The subject selects which of the sounds is different. The initial value was set to a SPT_{VAD} of 0.85. We used a 1up- 2down- method to obtain unbiased results. The subject had to answer two times correctly to reproduce a less distorted signal. If the subject answers wrong the algorithm goes 1 step behind, where the distortion is more perceptible. A maximum of 12 reversals was established.

3. RESULTS

A subjective test was performed using 5 post-locutive adult CI users. The mean age was 55 years. All of them were excellent performers (i.e. having more than 50% speech intelligibility at 10dB SNR). If the CI users had a bilateral implant we only tested the best ear. Signals were pre-processed in a PC and delivered to the CI speech processor through a direct-in audio cable.

3.1. Experiment 1 *jnd*SPT

Figure 3 presents the results for the *jnd*SPT experiment. The results show that 3 out of 5 subjects could perceive the distortion caused by the BF ($SPT_{VAD} < 1$). We also performed some tests with normal hearing listeners resulting in *jnd*SPT of around 0.5.

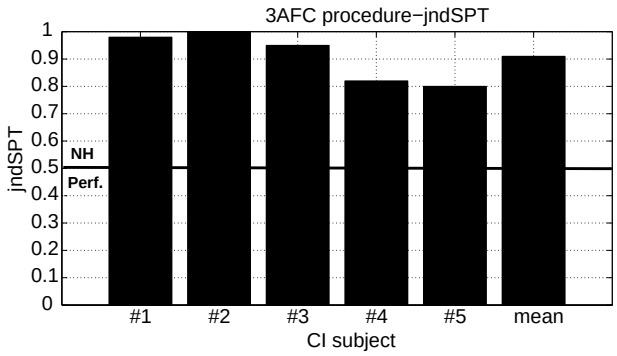

Fig. 3. Results of the jndSPT subjective experiment.

3.2. Experiment 2 Sound Quality

The quality of the sound delivered by the BF was assessed using two MUlti Stimulus with Hidden Reference and Anchor test (MUSHRA) [20] experiments. The first one, termed $MUSHRA_{SD}$, assessed the quality of the speech distorted signal $\tilde{s}(t)$. The second one, termed $MUSHRA_{SD-NR}$ assessed the quality of the signal $\tilde{y}(t)$.

3.2.1. $MUSHRA_{SD}$ description

The MUSHRA experiment was designed to evaluate the signal $\tilde{s}(t)$ processed with values SPT_{VAD} 0.4, 0.55, *jnd*SPT and 1. An anchor signal and the original clean and unprocessed signal were added to the experiment. The anchor signal is a low pass filtered signal of the most distorted signal (SPT_{VAD} =1). The experiment consists of rating the quality perception between 0 bad and 100 excellent for all presented signals. Figure 4 presents the results. The original clean speech signal obtained the highest scores. As expected, the signal processed with SPT_{VAD} =0.4 obtained the same ratings as the original signal because this VAD configuration does not produce noticeable speech leakage distortion. Unexpectedly, the *jnd*SPT was rated lower than the original speech signal. We think that the flexibility of the MUSHRA test, where the signals can be reproduced several times, makes this test more sensitive to perceive the differences in perceptual distortion than the *jnd*SPT. As expected, the anchor signal obtained the lowest ratings and the signals processed with a SPT_{VAD} =1.0 obtained lower ratings than the signals processed with lower SPT_{VAD} values.

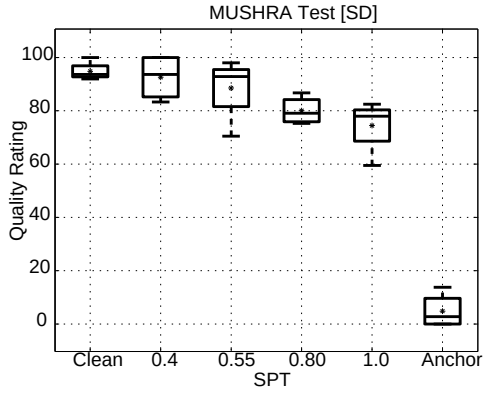

Fig. 4. Results of the $MUSHRA_{SD}$ test averaged for all CI users.

3.2.2. $MUSHRA_{SD-NR}$ description

The $\tilde{y}(t)$ signal was processed using the same SPT_{VAD} as in previous experiment. The original clean speech signal, the original noisy signal and the anchor signal were also added to the experiment. Here the subjects had to take into account both, the amount of noise reduction provided by the BF and the distortion produced by speech leakage. The results are presented in Figure 5. It can be observed that the differences between different SPT conditions are small. In some cases the distortions (produced by leakage) might be masked by the noise.

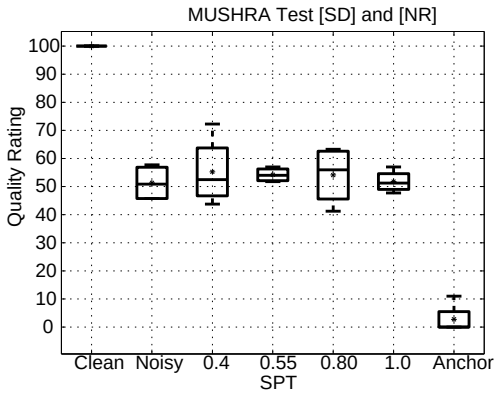

Fig. 5. Results of the $MUSHRA_{SD-NR}$ test averaged for all CI users.

3.3. Experiment 3 Speech Intelligibility

Speech intelligibility was measured by means of the HSM sentence test [21]. Two lists for each SPT_{VAD} condition were presented. The test was performed using the same simulated environment as in the previous experiments. Figure 6 presents the speech intelligibility scores. The results show that subjects obtained an improvement of 50% in speech intelligibility using a BF with high noise reduction (SPT_{VAD} = 1) with respect to not using a BF or using a BF with almost no noticeable speech leakage distortion (SPT_{VAD} = 0.4). The *jnd*SPT BF obtained the best performance, obtaining a good compromise between distortion and noise reduction. The *jnd*SPT was individually adjusted to each CI user using the results presented in Figure 3. In terms of speech intelligibility it hence seems more important for CI users to reduce the background noise than to limit the distortion produced by the BF. However, subjects who were more sensitive to perceive the speech distortion (lower *jnd*SPT) decreased their speech intelligibility performance when using SPT_{VAD} larger than their *jnd*SPT.

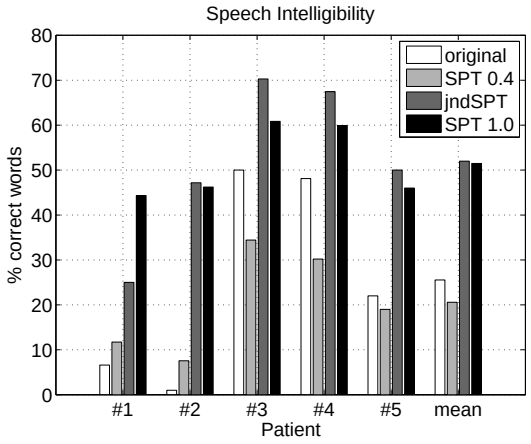

Fig. 6. HSM Speech Intelligibility scores.

4. CONCLUSIONS

The main goal of this paper was to investigate if a state-of-the-art beamformer (TF-GSC) produces perceived distortion in CI users and to find out if this distortion affects sound quality and speech intelligibility. With the 3AFC procedure we observed that some CI users perceive the distortion with different sensitivity than others. Furthermore, some subjects do not perceive the distortion produced by the BF. Second, we used a MUSHRA test to assess the quality of different BFs. We observed that MUSHRA is more sensitive to assess the distortion than 3AFC. However, the MUSHRA experiment presenting both the target signal and the background noise signal does not show different user preferences of BF configuration in terms of quality. Third, a speech intelligibility tests shows that speech leakage reduces speech intelligibility. For 4 out of 5 CI subjects the BF that produces just noticeable distortion (*jnd*SPT) is the one producing best speech intelligibility scores.

Acknowledgments

The authors would like to thank the subjects who have participated in the experiments. This work was supported by the DFG Cluster of Excellence EXC 1077/1 Hearing4all.

5. REFERENCES

[1] I. Hochberg, A. Boothroyd, M. Weiss, and S. Hellman, "Effects of noise and noise suppression on speech perception by cochlear implant users.," *Ear and hearing*, vol. 13, pp. 263–271, 1992.

[2] A. Buechner, K-H. Dyballa, P. Hehrmann, S. Fredelake, and Th. Lenarz, "Advanced beamformers for cochlear implant users: Acute measurement of speech perception in challenging listening conditions," *PLoS ONE*, vol. 9, 2014.

[3] A. Borowicz and A. Petrovsky, "Incorporating auditory properties into generalised sidelobe canceller," in *Signal Processing Conference (EUSIPCO), Proceedings of the 20th European*, Aug 2012, pp. 589–593.

[4] S. J. Mauger, P. W. Dawson, and A. Hersbach, "Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction," *The Journal of the Acoustical Society of America*, , no. November 2011, pp. 327.

[5] P. W. Dawson, S. J. Mauger, and A. Hersbach, "Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus cochlear implant recipients.," *Ear and hearing*, vol. 32, pp. 382–390, 2011.

[6] A. Spriet, L. Van Deun, K. Eftaxiadis, J. Laneau, M. Moonen, B. van Dijk, A. van Wieringen, and J. Wouters, "Speech understanding in background noise with the two-microphone adaptive beamformer BEAM in the Nucleus Freedom Cochlear Implant System.," *Ear and hearing*, vol. 28, pp. 62–72, 2007.

[7] B. L. Fetterman and E. H. Domico, "Speech recognition in background noise of cochlear implant patients," *Otolaryngol Head Neck Surg*, vol. 126, no. 3, pp. 257–263, Mar 2002.

[8] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, 2004.

[9] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, pp. 269–302. John Wiley and Sons, Inc., 2010.

[10] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, pp. 27–34, 1982.

[11] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," in *Proceedings of the IEEE*, vol. 60, pp. 926–935. 1972.

[12] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, 1969.

[13] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, pp. 1614–1626, 2001.

[14] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2677–2684, 1999.

[15] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, pp. 2230–2244, 2002.

[16] J. Sohn, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[17] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *Eurasip Journal on Advances in Signal Processing*, vol. 2009, 2009.

[18] W. Nogueira, A. Büchner, Th. Lenarz, and B. Edler, "A psychoacoustic NofM-type speech coding strategy for cochlear implants," *Eurasip Journal on Applied Signal Processing*, vol. 2005, pp. 3044–3059, 2005.

[19] P. C. Loizou and J. Ma, "Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms.," *The Journal of the Acoustical Society of America*, vol. 130, pp. 986, 2011.

[20] Recommendation-BS ITU-R, "1534-1: Method for the subjective assessment of intermediate quality levels of coding systems," *International Telecommunication Union, Geneva*, 2003.

[21] I. Hochmair-Desoyer, E. Schulz, L. Moser, and M. Schmidt, "The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users.," *The American Journal of Otology*, vol. 18, pp. S83, 1997.