Generalized Multichannel Wiener Filter for Spatially Distributed Microphones

Toby Christian Lawin-Ore[1], *Sebastian Stenzel*[2], *Jürgen Freudenberger*[2], *Simon Doclo*[1]

[1]University of Oldenburg, Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, Germany
Email: {toby.chris.lawin.ore,simon.doclo}@uni-oldenburg.de
Web: www.sigproc.uni-oldenburg.de

[2]HTWG Konstanz, Institute for System Dynamics, Germany
Email: {sstenzel,jfreuden}@htwg-konstanz.de
Web: www.isd.htwg-konstanz.de

Abstract

Considerable noise reduction can be achieved with the multichannel Wiener filter (MWF), which aims to estimate an unknown desired signal. In practice, the standard MWF (S-MWF) estimates the speech component in one of the microphone signals, referred to as the reference microphone signal. Recently, a different formulation of the MWF, which uses the envelope of all acoustic transfer functions (ATFs) between the speech source and the microphones, has been presented. It has been shown that this formulation of the MWF leads to a higher broadband output SNR than the S-MWF, especially for spatially distributed microphones. In this paper, we show that all ATFs can be exploited in a more general way to derive a generalized formulation for the MWF (G-MWF). This generalized formulation can then be used to derive an alternative expression for the S-MWF, where experimental results show that this alternative formulation leads to similar results as the G-MWF using the envelope of all ATFs.

1 Introduction

In recent years, research on speech enhancement using so-called *acoustic sensor networks* consisting of spatially distributed microphones has gained significant interest [1–5]. Spatially distributed microphones are able to acquire more information about the sound field than a single microphone array at one position.

In speech enhancement applications, the multichannel Wiener filter (MWF) that produces a minimum-mean-square-error (MMSE) estimate of an unknown desired signal is widely used for noise reduction [6, 7]. The desired signal of the standard MWF (S-MWF) is usually the speech component in one of the microphone signals, referred to as the reference microphone signal. However, for spatially distributed microphones, the selection of the reference microphone may have a large influence on the performance of the MWF, moreover depending on the positions of the speech/noise sources and the microphones [5]. For example, a higher broadband output SNR is obtained when selecting a reference microphone close to the speech source than when selecting a reference microphone located at a large distance. This effect is due to estimation errors in the speech correlation matrix, which lead to different output SNRs for different reference microphones.

Estimating the speech component in a reference microphone signal is equivalent to defining the desired overall transfer function for the speech component to be equal to the ATF between the speech source and the selected reference microphone. Recently, it has been proposed to define the desired overall transfer function using the envelope of all ATFs between the speech source and the microphones [2]. This desired overall transfer function leads to a different formulation of the MWF, whose performance does not rely on an explicit selection of a reference microphone [2].

In this paper, we propose to exploit the diversity of all ATFs in a more general way for defining the desired overall transfer function. The resulting generalized formulation of the MWF (G-MWF) can then be used to derive an alternative formulation for the S-MWF estimating the speech component in a specific reference microphone. Experimental results show that this alternative formulation of the S-MWF yields similar results as the G-MWF using the envelope of all ATFs.

2 Signal model

We consider M spatially distributed microphones capturing a speech signal in some noise field. The received microphone signals can be described in the frequency-domain as

$$\mathbf{Y}(\omega) = \mathbf{A}(\omega)S(\omega) + \mathbf{V}(\omega) = \mathbf{X}(\omega) + \mathbf{V}(\omega), \quad (1)$$

where $\mathbf{Y}(\omega) = [Y_1(\omega) \cdots Y_M(\omega)]^T$ denotes the stacked vector of the microphone signals, $\mathbf{A}(\omega) = [A_1(\omega) \cdots A_M(\omega)]^T$ denotes the stacked vector of the ATFs between the speech source $S(\omega)$ and the microphone array, and $\mathbf{X}(\omega)$ and $\mathbf{V}(\omega)$ represent the speech and the noise component in the microphone signals. The output signal $Z(\omega)$ is obtained by filtering and summing the microphone signals, i.e.,

$$Z(\omega) = \mathbf{W}^H(\omega)\mathbf{X}(\omega) + \mathbf{W}^H(\omega)\mathbf{V}(\omega), \quad (2)$$

where $\mathbf{W}(\omega) = [W_1(\omega) \cdots W_M(\omega)]^T$ represents the stacked vector of the filter coefficients. For conciseness the frequency-domain variable ω will be omitted where possible in the remainder of this paper.

The noisy speech correlation matrix $\mathbf{\Phi}_y$, the clean speech correlation matrix $\mathbf{\Phi}_x$ and the noise correlation matrix $\mathbf{\Phi}_v$ are defined as

$$\mathbf{\Phi}_y = \mathscr{E}\{\mathbf{Y}\mathbf{Y}^H\}, \mathbf{\Phi}_x = \mathscr{E}\{\mathbf{X}\mathbf{X}^H\}, \mathbf{\Phi}_v = \mathscr{E}\{\mathbf{V}\mathbf{V}^H\}, \quad (3)$$

where $\mathscr{E}\{\cdot\}$ denotes the expected value operator. Assuming that the speech and the noise components are uncorrelated, the correlation matrix $\mathbf{\Phi}_y$ can be expressed as $\mathbf{\Phi}_y = \mathbf{\Phi}_x + \mathbf{\Phi}_v$. For a single speech source, the speech correlation matrix $\mathbf{\Phi}_x$ is a rank-one matrix and is equal to

$$\mathbf{\Phi}_x = \phi_s \mathbf{A}\mathbf{A}^H, \quad (4)$$

with ϕ_s the power spectral density (PSD) of the source S,

[1]This work was partly supported by the Research Unit FOR 1732 "Individualized Hearing Acoustics" and the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG).

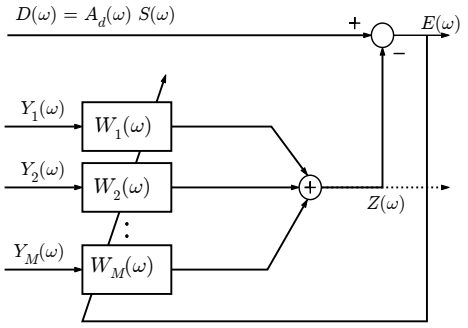


Figure 1: Multichannel Wiener filter.

i.e. $\phi_s = \mathscr{E}\{|S|^2\}$.

3 Multichannel Wiener filter

As illustrated in Figure 1, the multichannel Wiener filter (MWF) aims to estimate an unknown desired signal $D = A_d S$, where A_d is the desired overall transfer function of the speech component. To provide a trade-off between speech distortion and noise reduction, the speech-distortion-weighted multichannel Wiener filter has been proposed [6, 7], minimizing the weighted sum of the residual noise energy and the speech distortion energy, i.e.,

$$\xi(\mathbf{W}) = \mathscr{E}\{\left|A_d S - \mathbf{W}^H \mathbf{X}\right|^2\} + \mu \mathscr{E}\{|\mathbf{W}^H \mathbf{V}|^2\}, \quad (5)$$

where μ is a trade-off parameter between noise reduction and speech distortion. The filter minimizing (5) is given by

$$\mathbf{W} = (\mathbf{\Phi}_x + \mu \mathbf{\Phi}_v)^{-1} \phi_s \mathbf{A} A_d^*, \quad (6)$$

where we have used $\mathbf{X} = \mathbf{A}S$. Often, the MWF is used to estimate the speech component in one of the microphone signals, referred to as the reference microphone signal. This corresponds to defining

$$A_d = A_{m_0}, \quad (7)$$

i.e., the overall desired transfer function is equal to the ATF between the speech source and the selected reference microphone and $m_0 = 1 \dots M$. The resulting *standard* MWF (S-MWF) is then equal to

$$\boxed{\begin{aligned}\mathbf{W}_{\text{S-MWF}} &= (\mathbf{\Phi}_x + \mu \mathbf{\Phi}_v)^{-1} \phi_s \mathbf{A} A_{m_0}^* \\ &= (\mathbf{\Phi}_x + \mu \mathbf{\Phi}_v)^{-1} \mathbf{\Phi}_x \mathbf{e}_{m_0},\end{aligned}} \quad (8)$$

with $\mathbf{e}_{m_0}$ an M-dimensional vector with the m_0-th element equal to 1 and all other elements equal to 0, i.e., the vector selecting the column that corresponds to the reference microphone.

4 Generalized MWF formulation

In this section, we show that by defining a generalized overall desired transfer function A_d, a generalized formulation of the MWF can be derived, which incorporates different special cases.

To overcome the drawback of having to select a specific reference microphone, a multichannel Wiener filter whose performance does not rely on an explicit selection of a reference microphone has been introduced in [2]. It has been proposed to define the desired overall transfer function using the envelope of the ATFs as

$$A_d = ||\mathbf{A}||e^{j\psi_{m_0}} = \sqrt{\sum_{m=1}^{M} |A_m|^2} e^{j\psi_{m_0}}, \quad (9)$$

where $\psi_{m_0} = \arg(A_{m_0})$ is the phase of the ATF of an arbitrarily selected reference microphone. However, it should be noted that the phase of the desired transfer function has no impact on the narrowband nor on the broadband output SNR [2]. It has been experimentally shown in [2] that compared to the S-MWF filter, the MWF using the envelope of the ATFs leads to an improved broadband output SNR.

As a generalization of (9), the desired overall transfer function using a weighted combination of the amplitudes $|A_m|$ of the ATFs can be defined, i.e.,

$$A_d = \sqrt{\sum_{m=1}^{M} \alpha_m |A_m|^2}\ e^{j\psi_{m_0}}, \quad (10)$$

with $0 \leq \alpha_m \leq 1$. The parameters $\alpha_m, m = 1 \dots M$ introduce an additional degree of freedom in defining the desired overall transfer function.

By plugging (10) into (6), the resulting MWF referred to as the *generalized* MWF (G-MWF) can be expressed as

$$\boxed{\mathbf{W}_{\text{G-MWF}} = (\mathbf{\Phi}_x + \mu \mathbf{\Phi}_v)^{-1} \mathbf{\Phi}_x \mathbf{g}} \quad (11)$$

with

$$\mathbf{g} = \frac{\mathbf{A}\sqrt{\sum \alpha_m |A_m|^2}}{\mathbf{A}^H \mathbf{A}} e^{-j\psi_{m_0}}, \quad (12)$$

an M-dimensional vector where the m-th element is equal to

$$g_m = \frac{|A_m|\sqrt{\sum \alpha_m |A_m|^2}}{\mathbf{A}^H \mathbf{A}} e^{j(\psi_m - \psi_{m_0})}. \quad (13)$$

Since a rank-one speech correlation matrix $\mathbf{\Phi}_x$ is assumed (cf. 4), the phase difference $\psi_m - \psi_{m_0}$ can be computed using $\mathbf{\Phi}_x$ as

$$e^{j(\psi_m - \psi_{m_0})} = \frac{\mathbf{\Phi}_x(m, m_0)}{|\mathbf{\Phi}_x(m, m_0)|}. \quad (14)$$

Furthermore, using $\mathbf{\Phi}_x(m,m) = \phi_s |A_m|^2$ and $\text{tr}(\mathbf{\Phi}_x) = \phi_s \mathbf{A}^H \mathbf{A}$, g_m in (13) can be computed as

$$g_m = \frac{\sqrt{\mathbf{\Phi}_x(m,m)}\sqrt{\sum \alpha_m \mathbf{\Phi}_x(m,m)}}{\text{tr}(\mathbf{\Phi}_x)} \frac{\mathbf{\Phi}_x(m, m_0)}{|\mathbf{\Phi}_x(m, m_0)|}. \quad (15)$$

such that, the generalized MWF in (11) and (15) can be completely computed using the speech and noise correlation matrices $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_v$. It should be noted that although a rank-one speech correlation matrix has been assumed in the derivations, the MWF formulation can also be used for cases where $\mathbf{\Phi}_x$ is not a rank-one matrix.

The desired overall transfer function A_d in (10) incorporates different special cases. By setting $\alpha_m = 1$, $\forall m$, the overall transfer function is equal to $A_d = ||\mathbf{A}||e^{j\psi_{m_0}}$, cf. (9), corresponding to the desired overall transfer function proposed in [2], and the vector $\mathbf{g}_1$ is then given by

$$\boxed{g_{1,m} = \frac{\sqrt{\mathbf{\Phi}_x(m,m)}\sqrt{\sum \mathbf{\Phi}_x(m,m)}}{\text{tr}(\mathbf{\Phi}_x)} \frac{\mathbf{\Phi}_x(m, m_0)}{|\mathbf{\Phi}_x(m, m_0)|}} \quad (16)$$

On the other hand, by setting $\alpha_{m_0} = 1$ and $\alpha_m = 0$, $\forall m \neq m_0$, it can be easily seen that the desired overall transfer function corresponds to $A_d = A_{m_0}$ and the vector $\mathbf{g}_2$ can be computed as

$$\boxed{g_{2,m} = \frac{\sqrt{\mathbf{\Phi}_x(m,m)}\sqrt{\mathbf{\Phi}_x(m_0,m_0)}}{\text{tr}(\mathbf{\Phi}_x)} \frac{\mathbf{\Phi}_x(m,m_0)}{|\mathbf{\Phi}_x(m,m_0)|}} \tag{17}$$

Interestingly, by setting the desired overall transfer function equal to the ATF of a reference microphone, the MWF can now be computed in two different ways, i.e., either using the standard formulation of the MWF (S-MWF) in (8) or the generalized MWF (G-MWF) in (11) and (17). For a rank-one speech correlation matrix, i.e., $\mathbf{\Phi}_x = \phi_s \mathbf{A}\mathbf{A}^H$, it can be easily shown that

$$\mathbf{\Phi}_x \mathbf{e}_{m_0} = \phi_s \mathbf{A} A^*_{m_0} \frac{\mathbf{A}^H\mathbf{A}}{\mathbf{A}^H\mathbf{A}} = \phi_s \mathbf{A}\mathbf{A}^H \frac{\mathbf{A} A^*_{m_0}}{\mathbf{A}^H\mathbf{A}} = \mathbf{\Phi}_x \mathbf{g}_2. \tag{18}$$

such that both MWF formulations lead to the same filter coefficients. However, if the speech correlation matrix is not a rank-one matrix, e.g., due to estimation errors, different filter coefficients can be obtained, resulting in a different performance for both MWF formulations.

5 Experimental results

In this section, we will experimentally compare the performance of the S-MWF using $A_d = A_{m_0}$ and the G-MWF using $A_d = A_{m_0}$ and $A_d = ||\mathbf{A}||e^{j\psi_{m_0}}$.

5.1 Setup and performance measures

In a room with dimensions 4.8m×4.8m×3m and $T_{60} =$ 400ms, we consider the acoustic scenario depicted in Figure 2 with $M = 6$ spatially distributed microphones. The circles represent the microphone positions and the cross markers various possible positions of the desired source. The inter-microphone distance of the first (microphones # $1 \ldots 3$) and second (microphones # $4 \ldots 6$) microphone array are set to 10cm and 20cm, respectively. We consider a scenario with a single speech source, and diffuse noise generated using the method presented in [8]. The desired signal has been generated by convolving a clean speech signal from the HINT-database [9] with impulse responses simulated using the image model [10, 11]. The sampling frequency is $f_s = 16$kHz. For each position of the desired source, the a-priori input SNR is set to 5 dB.

In our STFT-based implementation, we have used the overlap/add method with a Hann analysis and synthesis window, and 50% overlap. The used FFT length is $N_{\text{FFT}} =$ 1024. For estimating the speech and noise correlation matrices, we have used a perfect voice activity detector to classify signal frames as speech dominant frames or noise dominant frames. The correlation matrices $\mathbf{\Phi}_y(\omega)$ and $\mathbf{\Phi}_v(\omega)$ are estimated in batch mode by using all speech + noise frames and all noise-only frames respectively, i.e.,

$$\hat{\mathbf{\Phi}}_y(\omega) = \frac{1}{F_x} \sum_{F_x} \mathbf{Y}(\omega)\mathbf{Y}^H(\omega), \tag{19}$$

$$\hat{\mathbf{\Phi}}_v(\omega) = \frac{1}{F_v} \sum_{F_v} \mathbf{V}(\omega)\mathbf{V}^H(\omega), \tag{20}$$

where F_x and F_v are the number of frames during speech + noise and noise-only periods. The speech correlation matrix $\hat{\mathbf{\Phi}}_x(\omega)$ is estimated as $\hat{\mathbf{\Phi}}_x(\omega) = \hat{\mathbf{\Phi}}_y(\omega) - \hat{\mathbf{\Phi}}_v(\omega)$. The trade-off parameter is set to $\mu = 1$. To evaluate the performance of the different MWF formulations, we have considered the broadband output SNR defined as

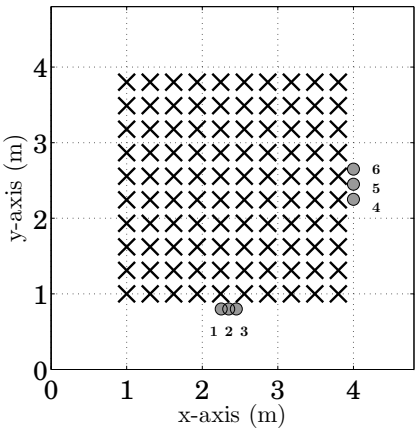


Figure 2: The scenario of an acoustic sensor network with $M = 6$ microphones.

$$\text{oSNR}^B = \frac{\sum_\omega \hat{\mathbf{W}}^H(\omega)\mathbf{\Phi}_x(\omega)\hat{\mathbf{W}}(\omega)}{\sum_\omega \hat{\mathbf{W}}^H(\omega)\mathbf{\Phi}_v(\omega)\hat{\mathbf{W}}(\omega)}, \tag{21}$$

which make use of the theoretical rank-one speech correlation matrix in (4). The ATFs $\mathbf{A}$ is computed using the simulated room impulse responses, and ϕ_s is estimated by calculating the PSD of the clean speech signal. The noise correlation matrix is calculated using the complete noise signal. In order to describe the performance of all considered MWF formulations for all positions using a single number, we define the spatially averaged broadband output SNR as

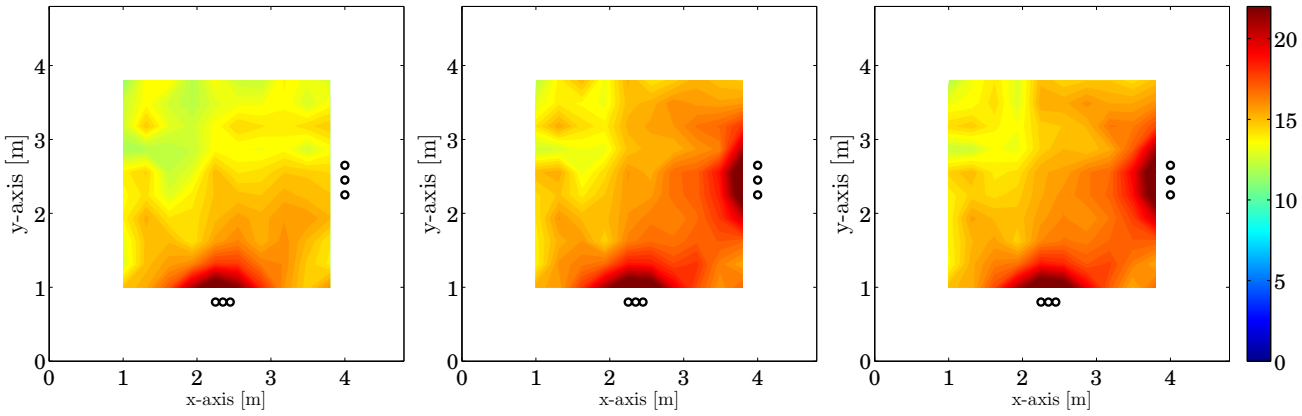
$$\text{oSNR}_{\text{avg}} = \frac{1}{N_s} \sum_{N_s} \text{oSNR}^B,$$

which averages the broadband output SNR over the considered N_s source positions.

5.2 Results

For $A_d = A_{m_0}$, the S-MWF and the G-MWF have been simulated for each possible choice of the reference microphone m_0, i.e., the ATF of each microphone has been selected as the desired overall transfer function. Figure 3a shows the broadband output SNR of the S-MWF for different positions of the desired speech source with the first microphone selected as the reference microphone, i.e., $m_0 = 1$. As expected, the S-MWF leads to good results at some positions of the desired source but to poor results at other positions. For example, a relatively small output SNR is achieved when the speech source is located in the area close to the microphones 4 to 6. Similar results are obtained for other reference microphones, e.g., by selecting one of the microphones of the second array (microphones # $4 \ldots 6$) as the reference microphone, a smaller output SNR is obtained when the speech source is located in the area close to the microphones 1 to 3.

Figure 3b shows the position-dependent broadband output SNR of the G-MWF with the first microphone selected as the reference microphone. In contrast to the S-MWF, the G-MWF leads to a higher broadband output SNR, es-



(a) (b) (c)

Figure 3: Position-dependent broadband output SNR of the different MWF filters: (a) S-MWF with $A_d = A_1$, (b) G-MWF with $A_d = A_1$, (c) G-MWF with $A_d = ||\mathbf{A}||e^{j\psi_1}$.

S-MWF						G-MWF						
A_d=A_{m_0}						A_d=A_{m_0}						A_d=$\|\|\mathbf{A}\|\|e^{j\psi_{m_0}}$
m_0=1	m_0=2	m_0=3	m_0=4	m_0=5	m_0=6	m_0=1	m_0=2	m_0=3	m_0=4	m_0=5	m_0=6	m_0=1…6
14.18	13.73	13.42	13.75	14.14	13.55	**16.08**	15.68	15.62	15.70	15.87	15.71	**15.90**

Table 1: Output SNR (oSNR_{avg} [dB]) of the S-MWF and G-MWF filters, averaged over all considered source positions.

pecially at positions of the speech source located close to non-reference microphones (microphones # 4…6).

Figure 3c depicts the position-dependent broadband output SNR of the G-MWF filter with the desired overall transfer function $A_d = ||\mathbf{A}||e^{j\psi_1}$. Interestingly, by comparing Figure 3b with Figure 3c, it can be seen that the obtained results using both formulations are very similar.

Furthermore, as can be observed in Table 1, the G-MWF computed using specific reference microphones yields a higher spatially averaged broadband output SNR than the S-MWF. Moreover, the G-MWF using specific reference microphones yields similar results as the G-MWF using $A_d = ||\mathbf{A}||e^{j\psi_1}$, whose performance is independent of the reference microphone.

6 Conclusion

In this paper, a generalized desired overall transfer function has been used to derive a generalized formulation of the MWF (G-MWF). It has been shown that the G-MWF can be used to derive an alternative formulation for the standard MWF (S-MWF), which estimates the speech component in a specific reference microphone. Simulation results have shown that the G-MWF filter leads to a higher broadband output SNR than the S-MWF. Furthermore, it has been shown that the G-MWF using a specific reference microphone yields similar results as the G-MWF using the envelope of all ATFs.

References

[1] S. Doclo, T. van den Bogaert, J. Wouters, and M. Moonen, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 38–51, Jan. 2009.

[2] S. Stenzel, T. C. Lawin-Ore, J. Freudenberger, and S. Doclo, "A Multichannel Wiener Filter with Partial Equalization for Distributed Microphones," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz NY, USA), Oct. 2013.

[3] A. Bertrand and M. Moonen, "Distributed LCMV beamforming in a wireless sensor network with single-channel per-node signal transmission," *IEEE Transactions on Signal Processing*, vol. 61, pp. 3447–3459, July 2013.

[4] S. M. Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 343–356, Feb. 2013.

[5] T. C. Lawin-Ore and S. Doclo, "Reference microphone selection for MWF-based noise reduction using distributed microphone arrays," in *Proc. ITG Conference Speech Communication*, (Braunschweig, Germany), pp. 31–34, Sept. 2012.

[6] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication, special issue on Speech Enhancement*, vol. 49, pp. 636–656, July 2007.

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

[8] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *Journal of the Acoustical Society of America*, vol. 122, pp. 3464–3470, Dec. 2007.

[9] M. Nilsson, S. D. Soli, and A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *Journal of the Acoustical Society of America*, vol. 95, pp. 1085–1099, Feb. 1994.

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, June 1979.

[11] E. A. P. Habets, "Room impulse response (RIR) generator," *available: http://home.tiscali.nl/ehabets/rir generator.html*, 2008.