A MULTICHANNEL WIENER FILTER WITH PARTIAL EQUALIZATION FOR DISTRIBUTED MICROPHONES

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

[1] HTWG Konstanz, Germany, Institute for System Dynamics, {sstenzel, jfreuden}@htwg-konstanz.de
[2] University of Oldenburg, Germany, Department of Medical Physics and Acoustics - Signal Processing Group, {toby.chris.lawin.ore, simon.doclo}@uni-oldenburg.de

ABSTRACT

In speech enhancement applications, the multichannel Wiener filter (MWF) is widely used to reduce noise and thus improve signal quality. The MWF performs noise reduction by estimating the desired signal component in one of the microphones, referred to as the reference microphone. However, for distributed microphones, the selection of the reference microphone has a significant impact on the broadband output SNR of the MWF, largely depending on the acoustical transfer function (ATF) between the desired source and the reference microphone. In this paper, a multichannel Wiener filtering approach using a soft combined reference is presented. Simulation results show that the proposed scheme leads to a higher broadband output SNR compared to an arbitrarily selected reference microphone, moreover achieving a partial equalization of the overall acoustic system.

Index Terms— Multichannel Wiener filter (MWF), spatially distributed microphones, equalization

1. INTRODUCTION

In many hands-free speech communication applications the desired speech signal is linearly corrupted by the acoustics of the environment. Furthermore there are often additive distortions due to background noise. Multi-microphone systems can benefit from the directivity which is achieved by a proper combining of the spatially sampled wave field. The performance of most speech enhancement algorithms depends on the spatial correlation properties of the noise field, but also on the environment's acoustics and therefore on the acoustic transfer functions (ATF) which corrupt the speech signal.

For a proper design, the room acoustics should be taken into account to guarantee an optimal signal combining in the sense of a maximal narrowband output SNR. Thus the system has to compensate the differences in the speech signals at the sensors, which implies knowledge of the relative transfer functions (RTFs). With the transfer function - generalize sidelobe canceler (TF-GSC) a beamformer design that relies on the relative transfer functions between the sensors was proposed [1]. Recently in [2] a minimum variance distortion-less response (MVDR) beamformer was presented, where a reference channel is used to derive the beamformer coefficients. In [3] the speech distortion weighted - multichannel Wiener filter (SDW-MWF) was proposed, where signal distortions are taken into account in the optimization. In this approach the distortion is measured as the distance between the speech component of the output signal and the speech component of a reference input channel.

*Research for this article was supported by DFG (FR 2673/2-1).

These approaches have some aspects in common, they all maximize the narrowband output SNR. Furthermore, all these approaches rely on the relative transfer functions between the sensors and an arbitrarily selected reference channel. Therefore the ATF of the reference channel remains as the overall transfer function (TF) (the TF between the speaker and the output of the combining system). In contrast to the narrowband output SNR, the overall transfer function may have an impact on the broadband output SNR, as shown in [4]. Especially in setups with widely distributed microphones the signal conditions at the individual microphones may highly vary. Thus for practical implementations the choice of the reference channel impacts the broadband output SNR.

In this contribution, we present a multichannel Wiener filter that does not rely on an explicit reference channel selection. In the proposed approach the overall transfer function is chosen as the envelope of the individual transfer functions. Hence, a soft weighted reference of the different channels is obtained. For diverse transfer functions the approach achieves a partial equalization of the acoustic system. This results in an improvement of the broadband output SNR in comparison to the MWF with a fixed reference microphone. The improvement is especially evident if we compare source positions nearby non-reference channels of the MWF. Considering different speech source positions, we can show that the spatially averaged broadband output SNR (averaged over all possible source positions) is improved.

2. SIGNAL MODEL

In this section, we briefly introduce the notation. In general, we consider M microphones and assume that the acoustic system is linear and time-invariant. Hence, the microphone signals $y_i(k)$ can be modeled by the convolution of the speech signal $x(k)$ with the impulse response $h_i(k)$ of the acoustic system plus an additive noise term $n_i(k)$. The M microphone signals $y_i(k)$ can be expressed in the short-time frequency domain as

$$Y_i(\kappa,\nu) = H_i(\nu)X(\kappa,\nu) + N_i(\kappa,\nu)\,, \tag{1}$$

where $Y_i(\kappa,\nu)$, $X(\kappa,\nu)$ and $N_i(\kappa,\nu)$ denote the corresponding short-time spectra and $H_i(\nu)$ the acoustic transfer functions. $S_i(\kappa,\nu) = H_i(\nu)X(\kappa,\nu)$ is the speech component of the i^{th} microphone signal. The subsampled time index and the frequency bin index are denoted by κ and ν, respectively. In the remainder of this paper the dependencies on κ and ν are omitted. We define the M-dimensional vectors $\mathbf{S}$, $\mathbf{N}$ and $\mathbf{Y}$, in which the signals are stacked

978-1-4799-0972-8/13/$31.00 ©2013IEEE

as follows:

$$\mathbf{S} = [S_1\ S_2\ \cdots\ S_M]^T \quad (2)$$
$$\mathbf{N} = [N_1\ N_2\ \cdots\ N_M]^T \quad (3)$$
$$\mathbf{Y} = \mathbf{S} + \mathbf{N}, \quad (4)$$

Note that T denotes the transpose of a vector or matrix, whereas the conjugate transpose is denoted by $\dagger$ and conjugation by $*$, respectively. $\mathbf{H}$ denotes the vector of channel coefficients

$$\mathbf{H} = [H_1\ H_2\ \cdots\ H_M]^T. \quad (5)$$

We assume that the noise signals are zero-mean random processes with the variances $\sigma^2_{N_1}, \ldots, \sigma^2_{N_M}$. Furthermore we assume that the single speaker speech signal is a zero-mean random process with PSD σ^2_X and a time-invariant acoustic system $\mathbf{H}$. The correlation matrix of the speech signal can be written as

$$\mathbf{R}_S = \mathbb{E}\left\{\mathbf{S}\mathbf{S}^\dagger\right\} = \sigma^2_X \mathbf{H}\mathbf{H}^\dagger. \quad (6)$$

3. MULTICHANNEL MMSE CRITERION

In the following section the multichannel Wiener filter is derived. Similar to [2], we constrain the resulting transfer function of the overall system to be equal to $\tilde{H}$. With this parameter we introduce a degree of freedom in our filter design. Thus we are able to explicitly design the overall transfer function of our system. Note that the overall transfer function $\tilde{H}$ does not affect the narrowband output SNR but it has an influence on the broadband output SNR, as we will show later on.

To calculate the minimum mean squared error (MMSE) estimate of the target speech signal $\tilde{H}X$, one has to minimize the following cost function

$$\mathbf{G}^{\mathrm{MWF}} = \arg\min_{\mathbf{G}} \mathbb{E}\left\{|\mathbf{G}^\dagger\mathbf{Y} - \tilde{H}X|^2\right\}. \quad (7)$$

For this minimization we can rewrite the error signal ε as follows

$$\varepsilon = \mathbf{G}^\dagger\mathbf{Y} - \tilde{H}X = \underbrace{(\mathbf{G}^\dagger\mathbf{S} - \tilde{H}X)}_{\varepsilon_x} + \underbrace{\mathbf{G}^\dagger\mathbf{N}}_{\varepsilon_n}. \quad (8)$$

Using the two MSE cost functions

$$J_n(\mathbf{G}) = \mathbb{E}\left\{|\varepsilon_n|^2\right\}, \quad (9)$$
$$J_x(\mathbf{G}) = \mathbb{E}\left\{|\varepsilon_x|^2\right\} \quad (10)$$

the unconstrained minimization criterion for the parametric MWF is defined by

$$\mathbf{G}^{\mathrm{MWF}} = \arg\min_{\mathbf{G}} J_n(\mathbf{G}) + \frac{1}{\mu_{\mathrm{W}}} J_x(\mathbf{G}), \quad (11)$$

where $\frac{1}{\mu_{\mathrm{W}}}$ is a Lagrange multiplier. This results in the solution

$$\mathbf{G}^{\mathrm{MWF}} = (\mathbf{R}_S + \mu_{\mathrm{W}}\mathbf{R}_N)^{-1}\sigma^2_X\mathbf{H}\tilde{H}^*, \quad (12)$$

where $\mathbf{R}_S$ and $\mathbf{R}_N$ are the speech and noise correlation matrix, respectively. The parameter μ_{W} allows a trade-off between noise reduction and speech distortion with respect to our target signal $\tilde{H}X$ (for details cf. [3]).

Using the matrix inversion lemma the MWF in eq. (12) can be rewritten as [5]

$$\mathbf{G}^{\mathrm{MWF}} = \frac{\sigma^2_X}{\sigma^2_X + \mu_{\mathrm{W}}(\mathbf{H}^\dagger\mathbf{R}_N^{-1}\mathbf{H})^{-1}} \frac{\mathbf{R}_N^{-1}\mathbf{H}}{\mathbf{H}^\dagger\mathbf{R}_N^{-1}\mathbf{H}} \tilde{H}^*$$
$$\mathbf{G}^{\mathrm{MWF}} = G^{\mathrm{WF}}\mathbf{G}^{\mathrm{MVDR}}\tilde{H}^*. \quad (13)$$

As one can see, the parametric MWF can be decomposed as an MVDR beamformer $\mathbf{G}^{\mathrm{MVDR}}$, a filter that is equal to the overall transfer function $\tilde{H}$, and a single-channel Wiener filter

$$G^{\mathrm{WF}} = \frac{\sigma^2_X}{\sigma^2_X + \mu_{\mathrm{W}}\sigma^2{}_{N_{\mathrm{MVDR}}}}, \quad (14)$$

where μ_{W} can be interpreted as noise overestimation factor and $\sigma^2{}_{N_{\mathrm{MVDR}}}$ is the noise variance at the output of $\mathbf{G}^{\mathrm{MVDR}}$

$$\sigma^2{}_{N_{\mathrm{MVDR}}} = \mathbf{G}^{\mathrm{MVDR}\dagger}\mathbf{R}_N\mathbf{G}^{\mathrm{MVDR}} = (\mathbf{H}^\dagger\mathbf{R}_N^{-1}\mathbf{H})^{-1}. \quad (15)$$

The broadband output SNR is defined as

$$\gamma_{\mathrm{out}} = \frac{\sum_\nu \mathbf{G}(\nu)^\dagger\mathbf{R}_S(\nu)\mathbf{G}(\nu)}{\sum_\nu \mathbf{G}(\nu)^\dagger\mathbf{R}_N(\nu)\mathbf{G}(\nu)} = \frac{\sum_\nu \sigma^2_X(\nu)|G^{\mathrm{WF}}(\nu)|^2|\tilde{H}(\nu)|^2}{\sum_\nu (\mathbf{H}(\nu)^\dagger\mathbf{R}_N^{-1}(\nu)\mathbf{H}(\nu))^{-1}|G^{\mathrm{WF}}(\nu)|^2|\tilde{H}(\nu)|^2} = \frac{\sum_\nu \sigma^2_X(\nu)|G^{\mathrm{WF}}(\nu)|^2|\tilde{H}(\nu)|^2}{\sum_\nu \sigma^2{}_{N_{\mathrm{MVDR}}}(\nu)|G^{\mathrm{WF}}(\nu)|^2|\tilde{H}(\nu)|^2}. \quad (16)$$

From this equation we can observe that the overall transfer function influences the weighting between the SNR values of the individual frequency-bands and thus impacts the broadband output SNR [2].

4. REFERENCE SELECTION FOR THE MULTICHANNEL WIENER FILTER

In this section different choices for the overall transfer function $\tilde{H}$ are discussed. Setting $\tilde{H}$ to one, the MWF achieves a complete dereverberation, i.e the clean speech signal x is estimated. However the filter coefficients depend on the unavailable cross-correlation vector of the clean speech and the microphone input signal.

Choosing $\tilde{H} = H_{\mathrm{ref}}$ the filter $\mathbf{G}^{\mathrm{MWF}}$ can be calculated by knowledge of the second-order statistics, i.e. the noise and the speech correlation matrices. But due to the reference channel selection, the ATF of the reference channel remains as the overall TF. As shown in the section before, the overall TF influences the broadband output SNR. Especially for distributed microphones the individual TFs differ and thus for some frequency-bands the arbitrarily selected reference may not be the channel with the best energy among all channels. This can degrade the broadband output SNR. Therefore we propose to choose $\tilde{H}$ as the envelope of the individual transfer functions, which improves the broadband output SNR. Furthermore a partial equalization of the acoustic system is achieved.

4.1. Using an Explicit Reference Channel

By setting $\tilde{H} = H_{\mathrm{ref}}$, the MWF minimizes the mean squared error with respect to the speech signal of a reference microphone signal

S_{ref}. This results in the following solution

$$\begin{aligned}\mathbf{G}^{\mathrm{MWF}} &= (\mathbf{R}_S + \mu_{\mathrm{W}}\mathbf{R}_N)^{-1}\sigma_X^2\mathbf{H}\tilde{H}^* \\ &= (\mathbf{R}_S + \mu_{\mathrm{W}}\mathbf{R}_N)^{-1}\mathbf{R}_S\mathbf{u}\,, \end{aligned} \quad (17)$$

where $\mathbf{u}$ is an M-dimensional vector which selects the reference channel, i.e. the corresponding entry is set to one and the others are set to zero.

Even if this selection of the overall transfer function provides a maximal narrowband output SNR, the speech signal at the output is as reverberant as the speech signal of the reference microphone. The weighting of the frequency-bands with the resulting transfer function can degrade the broadband output SNR (see eq. (16)). E.g. a frequency-band having a higher SNR in a non-reference channel as in the reference channel, will have also a high SNR at the output, but due to the explicit reference selection of this method it may be attenuated by the channel coefficient H_{ref}.

4.2. Multichannel Wiener Filtering with Partial Equalization

In this subsection we show that using the second-order statistic of the input signals, an improved overall transfer function can be defined. The overall transfer function is chosen as the envelop of ATFs to obtain a partial equalization of the acoustic system.

For a coherent combining of the speech signals we have to compensate the phase difference between the speech signals at each microphone. As with the MWF in eq. (17), it is sufficient to estimate the phase differences to a reference microphone. Let $\phi_i(\nu)$ be the phase of the complex channel coefficient $H_i(\nu)$, i.e. $H_i(\nu) = |H_i(\nu)|e^{j\phi_i(\nu)}$. Then the phase differences to a reference microphone are given by $\Delta_i(\nu) = \phi_{\mathrm{ref}}(\nu) - \phi_i(\nu)$.

We define the overall transfer function as

$$\tilde{H} = \sqrt{\mathbf{H}^\dagger\mathbf{H}}\, e^{j\phi_{\mathrm{ref}}}\,, \quad (18)$$

where the phase ϕ_{ref} of the reference microphone is selected as the phase of the output signal [6]. Note, estimating the phase according to a reference channel has no influence on the broadband output SNR (see eq. (16)). But choosing the magnitude of the overall transfer function as $\sqrt{\mathbf{H}^\dagger\mathbf{H}}$ ensures that the speech energy of a non-reference channel is not attenuated by the TF of the reference channel. This is shown in Figure 1, the proposed overall transfer function corresponds to the envelope of the individual transfer functions. Here we have plotted three transfer functions (dotted and dashed) measured with an artificial head in a conference room of a size of 4.5 m×4.5 m×3 m. The distance between the first and the second microphone was set to 1.2 m and 1 m for the other microphone pairs. The solid line is the magnitude of the overall transfer function $\tilde{H}$. As we can see, the deep dips in the individual ATFs are equalized by the contribution of the other channels (e.g. around 1.2kHz for the channel H_2).

Using the overall transfer function defined in (18), eq. (12) can be rewritten as

$$\begin{aligned}\mathbf{G}^{\mathrm{MWF-P}} &= (\mathbf{R}_S + \mu_{\mathrm{W}}\mathbf{R}_N)^{-1}\sigma_X^2\mathbf{H}\sqrt{\mathbf{H}^\dagger\mathbf{H}}\, e^{-j\phi_{\mathrm{ref}}} \\ &= (\mathbf{R}_S + \mu_{\mathrm{W}}\mathbf{R}_N)^{-1}\mathbf{R}_S\frac{\mathbf{H}}{\sqrt{\mathbf{H}^\dagger\mathbf{H}}}\, e^{-j\phi_{\mathrm{ref}}} \\ &= (\mathbf{R}_S + \mu_{\mathrm{W}}\mathbf{R}_N)^{-1}\mathbf{R}_S\mathbf{u}^{\mathrm{p}}\,. \end{aligned} \quad (19)$$

Similarly to the methods proposed in [4], the vector $\mathbf{u}^{\mathrm{p}}$ selects the reference channel, but now using a soft weighted selection with respect to the magnitudes of the input channels. The elements of the vector $\mathbf{u}^{\mathrm{p}}$ can be rewritten and computed as

$$\begin{aligned}u_i^{\mathrm{p}} &= \frac{H_i}{\sqrt{\mathbf{H}^\dagger\mathbf{H}}}\, e^{-j\phi_{\mathrm{ref}}} \\ &= \sqrt{\frac{\sigma_X^2|H_i|^2}{\sigma_X^2\mathbf{H}^\dagger\mathbf{H}}}\, e^{-j\Delta_i} \\ &= \sqrt{\frac{\mathrm{r}_{s_{ii}}}{\mathrm{tr}(\mathbf{R}_{\mathrm{S}})}}\,\frac{\sigma_X^2 H_i H_{\mathrm{ref}}^*}{|\sigma_X^2 H_i H_{\mathrm{ref}}^*|} \\ &= \sqrt{\frac{\mathrm{r}_{s_{ii}}}{\mathrm{tr}(\mathbf{R}_{\mathrm{S}})}}\,\frac{\mathrm{r}_{s_{i\,\mathrm{ref}}}}{|\mathrm{r}_{s_{i\,\mathrm{ref}}}|}\,, \end{aligned} \quad (20)$$

where $\mathrm{r}_{s_{ij}}$ denotes the $(i,j)^{\mathrm{th}}$ element of the speech correlation matrix $\mathbf{R}_S$ and tr is the trace operator. Thus the phase differences are taken from the speech correlation matrix. As it can be seen from eq. (19) and eq. (20) we get rid of all direct dependencies on the acoustic transfer functions, i.e. knowledge of the second-order statistics is sufficient for the filter computation.

Considering the decomposition of the parametric MWF eq. (13), we notice that this MWF partially equalizes the acoustic system. The transfer function $\tilde{H} = \sqrt{\mathbf{H}^\dagger\mathbf{H}}\, e^{j\phi_{\mathrm{ref}}}$ remains as overall transfer function.

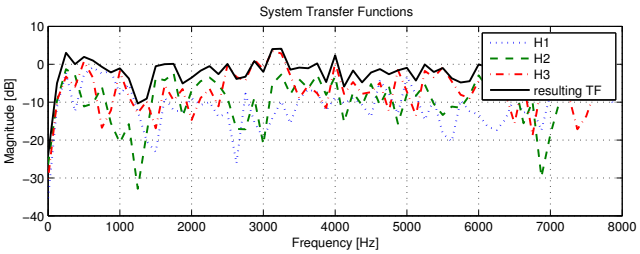


Figure 1: The individual transfer functions and the resulting one.

5. SIMULATION RESULTS

In this section we show simulation results for the new derived multichannel Wiener filter (eq. (19)). We consider a three-microphone setup in a room with the same geometry as for the ATFs plotted in Figure 1. Furthermore a six-microphone setup is simulated. The broadband output SNR of the MWF has been evaluated for various positions of the desired source. For each position of the desired source (every 0.21 m), impulse responses have been generated using the image method. The reverberation time was chosen to $T_{60} = 400$ ms and the SNR at the speech source was set to 20 dB for each possible speaker position. As noise source we considered a diffuse noise field.

The algorithms were simulated in a batch-mode. The correlation matrices are estimated in advance with full access to the input data, but based on the microphone signals $\mathbf{Y}$ and an ideal voice activity detection. The speech correlation matrix is estimated as $\mathbf{R}_S = \mathbf{R}_Y - \mathbf{R}_N$. For practical implementations, negative values on the main diagonal of $\mathbf{R}_S$ (due to the estimation of the correlation matrix $\mathbf{R}_S$) are floored to a small positive constant. For all simulations we set the parameter $\mu_{\mathrm{W}} = 1$.

The MWF using an explicit reference channel was simulated M times, thus each channel was selected as a reference input once. The left plot in Figure 2(a) shows the position dependent broadband output SNR for the MWF with channel 1 as a fixed reference. As one can see, the SNR is very high for source positions near the

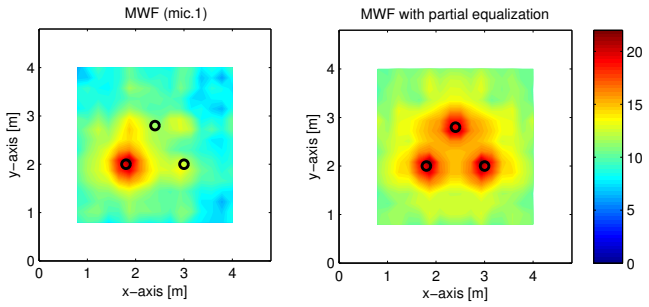
Setup	SNR [dB]									
	ch.1	ch.2	ch.3	ch.4	ch.5	ch.6	$\mathrm{SNR}_{\mathrm{in}}^{\mathrm{max}}$	Energy	$\mathrm{SNR}_{\mathrm{out}}^{\mathrm{max}}$	p. eq.
3 mics	10.7	10.9	10.8	n/a	n/a	n/a	12.7	12.7	13.1	13.3
6 mics	9.7	10.2	9.7	9.6	9.7	10.2	12.7	12.8	12.8	15.0

Table 1: Spatially averaged output SNR for the two simulated scenarios. The first six columns correspond to the MWF using a fixed reference channel, followed by results of the approaches using a frequency-selective reference. The last column presents the results for the approach with partial equalization.

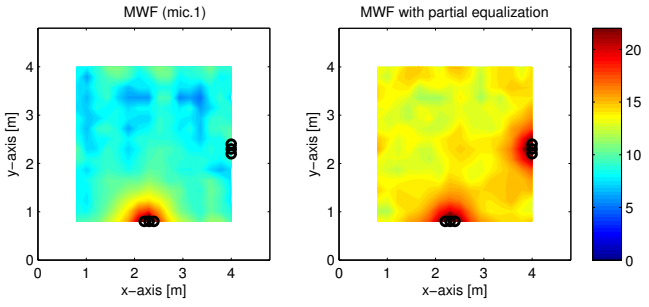
reference microphones, but with an increasing distance the SNR of the processed signal decreases rapidly. The broadband output SNR is also low for source positions close to non-reference microphone positions. This effect can be explain as follows: a frequency-band having a higher SNR in non-reference channels than in the reference channel will have also a high SNR at the output, but due to the explicit reference selection of the MWF it is attenuated by the channel coefficient H_{ref} (see eq. (16)).

In the right plot of Figure 2(a), the position dependent output SNR of the MWF using the proposed partial equalization approach is depicted. Compared to the case when one of the microphones is selected as reference (in this case the first microphone), a higher or equal broadband output SNR is obtained at all positions. Especially, a higher broadband output SNR is always obtained for positions of the desired source close to non-reference microphones.

In Figure 2(b) the results for the microphone setup using six microphones are shown. Here, two microphone arrays have been considered. Each array consists of three microphones with an inter-microphone distance set to 0.1 m. In this setup the effect of the soft combined reference is even more obvious.



(a) Three microphone setup.



(b) Six microphone setup.

Figure 2: Position dependent broadband output SNR of the different MWF. The black circles represents the microphone positions.

Table 1 shows the spatially averaged output SNR (averaged over all possible source positions). Here the presented approach is also compared to other frequency-dependent schemes, which are based on the highest narrowband input SNR, the highest narrowband output SNR and the highest narrowband energy [4]. As expected, an arbitrary selected reference microphone yields poor performance compared to the other reference selection procedures. Furthermore, we notice that the proposed approach leads to the highest spatially averaged broadband output SNR.

6. CONCLUSIONS

In this paper, a soft weighted reference selection procedure for spatially distributed microphones has been presented. Simulation results have shown that compared to an arbitrarily selected reference microphone, the novel frequency-dependent method leads to an improved broadband output SNR, moreover providing a partial equalization of the acoustic system. The improvement is especially evident if we compare source positions close to non-reference channels of the MWF.

7. REFERENCES

[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[2] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New Insights Into the MVDR Beamformer in Room Acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2010.

[3] 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*, vol. 49, no. 7-8, pp. 636–656, July 2007.

[4] T. C. Lawin-Ore and S. Doclo, "Reference Microphone Selection for MWF-based Noise Reduction Using Distributed Microphone Arrays," in *Proceedings of 10. ITG Symposium on Speech Communication*, Sept. 2012, pp. 31–34.

[5] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer Berlin Heidelberg, 2008, pp. 945–978.

[6] S. Stenzel and J. Freudenberger, "Blind Matched Filtering for Speech Enhancement with Distributed Microphones," *Journal of Electrical and Computer Engineering*, 2012, Article ID 169853, 15 pages.