Grimm *et al. EURASIP Journal on Advances in Signal Processing* (2016) 2016:78
DOI 10.1186/s13634-016-0375-6

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Phase reference for the generalized multichannel Wiener filter

Simon Grimm[1], Toby Christian Lawin-Ore[2], Simon Doclo[2] and Jürgen Freudenberger[1*]

Abstract

The multichannel Wiener filter (MWF) is a well-established noise reduction technique for speech processing. Most commonly, the speech component in a selected reference microphone is estimated. The choice of this reference microphone influences the broadband output signal-to-noise ratio (SNR) as well as the speech distortion. Recently, a generalized formulation for the MWF (G-MWF) was proposed that uses a weighted sum of the individual transfer functions from the speaker to the microphones to form a better speech reference resulting in an improved broadband output SNR. For the MWF, the influence of the phase reference is often neglected, because it has no impact on the narrow-band output SNR. The G-MWF allows an arbitrary choice of the phase reference especially in the context of spatially distributed microphones.

In this work, we demonstrate that the phase reference determines the overall transfer function and hence has an impact on both the speech distortion and the broadband output SNR. We propose two speech references that achieve a better signal-to-reverberation ratio (SRR) and an improvement in the broadband output SNR. Both proposed references are based on the phase of a delay-and-sum beamformer. Hence, the time-difference-of-arrival (TDOA) of the speech source is required to align the signals. The different techniques are compared in terms of SRR and SNR performance.

1 Introduction

Recently, research on speech enhancement using so-called acoustic sensor networks consisting of spatially distributed microphones has gained significant interest [1–12]. Compared with a microphone array at a single position, spatially distributed microphones are able to acquire more information about the sound field. The usage of spatially distributed microphones allows to employ beamforming techniques for speech quality improvement in reverberant and noisy conditions. Several methods were introduced that use a reference channel. These include the relative transfer function—generalized sidelobe canceler (RTF-GSC) [13], the minimum variance distortionless response (MVDR) beamformer [14], and the speech distortion weighted—multichannel Wiener filter (SDW-MWF) [15, 16].

The MWF is a well-established technique for speech enhancement. It produces a minimum-mean-squared error (MMSE) estimate of an unknown desired signal. 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. For spatially distributed microphones, the selection of the reference microphone may have a large influence on the performance of the MWF depending on the positions of the speech/noise sources and the microphones [5–7, 17].

With the S-MWF, the overall transfer function from the speakers to the output of the MWF equals the acoustic transfer function (ATF) from the speaker to the reference microphone. Hence, the reference microphone selection determines the amount of speech distortion. Moreover, the overall transfer function has an impact on the broadband output SNR of the MWF [17]. In [5], an MWF formulation with partial equalization (P-MWF) was presented, where the overall transfer function was chosen as the envelope of the individual ATFs with the phase of an arbitrary reference microphone. This results in a partial

*Correspondence: jfreuden@htwg-konstanz.de
This work was supported by the Research Unit FOR 1732 "Individualized Hearing Acoustics" and research grant FR 2673/2-3, both funded by the German Research Foundation (DFG).
[1] HTWG Konstanz, University of Applied Sciences Institute for System Dynamics—Signal Processing Group, 78462 Konstanz, Germany
Full list of author information is available at the end of the article

equalization of the acoustic system and an improved broadband output SNR. While this approach has advantages with respect to background noise reduction, the reverberation caused by the acoustic environment is not reduced.

Recently, the generalized MWF was proposed in order to improve the broadband output SNR [7] (see also [6]). With the G-MWF, the speech reference is a weighted sum of the speech components, such that the output signal has the same phase as the speech component in the reference microphone. The overall transfer function is the weighted sum of squared amplitudes of all ATFs.

In this work, we consider the phase of the speech reference. That is, we present a further generalization of the G-MWF approach in [7], which enables different phase references. We demonstrate that the phase of the speech reference shapes the overall transfer function and hence impacts the speech distortion. Moreover, the overall transfer function influences the broadband output SNR. We propose two speech references that achieve a better signal-to-reverberation ratio and an improvement in broadband output SNR. The proposed references are based on the phase of a delay-and-sum beamformer (DSB) [18].

As shown in [19], the temporal smearing and therefore the reverberation relies on the all-pass component of the overall transfer function. This suggests that a suitable phase reference can improve the output SRR of the system. As a consequence, the phase term of a delay-and-sum beamformer is applied as a phase reference of the G-MWF. Similar concepts were proposed in [20–22]. The DSB needs an estimate of the TDOA to align the signals properly. In the literature, several methods for TDOA estimation were proposed [23–30]. Many of these techniques are summarized in [29].

The work is a sequel to [21]. In addition to the concept proposed in [21], we present a new approach that combines the delay-and-sum beamformer and the P-MWF. Both approaches for the G-MWF can improve the SRR and SNR compared with the S-MWF and P-MWF. Furthermore, we present a theoretical analysis of the broadband output SNR of the G-MWF.

The paper is organized as follows: in Section 2, we introduce the signal model and notation. The G-MWF formulation and the analysis of the output SNR are presented in Sections 3 and 4, respectively. The design of the overall transfer function is explained in Section 5. The block diagram structure of the system is presented in Section 6, together with the necessary TDOA estimation and the challenge of acquiring these estimates in noisy and reverberated environments. In Section 7, the simulation results in terms of SNR and SRR improvement are given, followed by a conclusion in Section 8.

2 Signal model and notation

We consider a linear and time-invariant acoustic system. The beamformer array consists of M microphones. The ith microphone signal $y_i(k)$ can be expressed as the convolution of the speech signal $s(k)$ with the acoustic impulse response $h_i(k)$ from the speech source to the ith microphone plus an additive noise term $n_i(k)$. In the short time frequency domain, the resulting microphone signals can be written as follows

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

$Y_i(\kappa,\nu)$, $S(\kappa,\nu)$, and $N_i(\kappa,\nu)$ correspond to the short time spectra of the time domain signals. $H_i(\nu)$ represents the ATF corresponding to the the acoustic impulse response and $X_i(\kappa,\nu) = H_i(\nu)S(\kappa,\nu)$ is the speech component at the ith microphone. κ and ν denote the subsampled time index and the frequency bin index, respectively. In the following, these indices are often omitted when possible. The short time spectra and the ATF can be written as M-dimensional vectors:

$$\boldsymbol{X} = [X_1, X_2, \ldots, X_M]^T \tag{2}$$
$$\boldsymbol{N} = [N_1, N_2, \ldots, N_M]^T \tag{3}$$
$$\boldsymbol{H} = [H_1, H_2, \ldots, H_M]^T \tag{4}$$
$$\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_M]^T \tag{5}$$
$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N} \tag{6}$$

^{T} denotes the transpose of a vector, * the complex conjugate, and † denotes the conjugate transpose. Vectors and matrices are written in bold and scalars are normal letters.

We assume that the speech and noise signals are zero-mean random processes with the power spectral densities (PSDs) ${\Phi_{Ni}}^2$ and ${\Phi_S}^2$. Assuming a single speech source, the speech correlation matrix $\boldsymbol{R}_S$ has rank one and therefore can be expressed as

$$\boldsymbol{R}_S = \mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\dagger\right\} = P_S\boldsymbol{H}\boldsymbol{H}^\dagger, \tag{7}$$

where $\mathbb{E}$ denotes the mathematical expectation. Similarly, $\boldsymbol{R}_N = \mathbb{E}\{\boldsymbol{N}\boldsymbol{N}^\dagger\}$ denotes the noise correlation matrix. It is assumed, that the speech and noise terms are uncorrelated.

The output signal Z of the beamformer with filter coefficients $\mathbf{G} = [G_1, G_2, \ldots, G_M]^T$ is obtained by filtering and summing the microphone signals, i.e.,

$$\begin{aligned} Z &= \mathbf{G}^\dagger\mathbf{Y} = \mathbf{G}^\dagger\mathbf{X} + \mathbf{G}^\dagger\mathbf{N} \\ &= Z_S + Z_N \end{aligned} \tag{8}$$

where Z_S and Z_N denote the speech and the noise components at the beamformer output.

3 Generalized MWF

The MWF aims to estimate an unknown signal $\tilde{H}_d S$, where $\tilde{H}_d$ denotes the overall transfer function of the speech component [15, 16, 31]. The parametric MWF minimizes the weighted sum of the residual noise energy and the speech distortion energy, i.e., the cost function

$$\xi(\mathbf{G}) = \mathbb{E}\left\{\left|\tilde{H}_d S - \mathbf{G}^\dagger \mathbf{X}\right|^2\right\} + \mu \mathbb{E}\left\{|\mathbf{G}^\dagger \mathbf{N}|^2\right\}, \tag{9}$$

where μ is a trade-off parameter between noise reduction and speech distortion. The filter minimizing (9) is given by

$$\mathbf{G} = (\boldsymbol{R}_S + \mu \boldsymbol{R}_N)^{-1} P_S \boldsymbol{H} \tilde{H}_d^*. \tag{10}$$

Commonly, the MWF is implemented as

$$\mathbf{G} = (\boldsymbol{R}_S + \mu \boldsymbol{R}_N)^{-1} \boldsymbol{R}_S \mathbf{u}, \tag{11}$$

where $\mathbf{u}$ is a vector that selects the reference microphone, i.e., the vector $\mathbf{u}$ contains a single one and all other elements are zero. Therefore, the overall transfer function is equal to the ATF of a reference microphone, i.e. $H_d = H_{\text{ref}}$.

Since, $\boldsymbol{R}_S$ is a rank one matrix, it should be noted that any non-zero vector $\mathbf{u}$ achieves the same (optimal) narrow-band output SNR. In [7], the generalized MWF was presented, where the elements u_i of the vector $\mathbf{u}$ define a speech reference for the MWF which is a weighted sum of the speech components in the different microphones with the phase of the speech component in the reference microphone signal. The vector $\mathbf{u}$ can be used to define the desired complex-valued response as

$$\tilde{H}_d = \mathbf{u}^\dagger \boldsymbol{H} = \sum_i u_i^* \cdot H_i \;\text{ for }\; u_i \in \mathbb{C}. \tag{12}$$

In [7], the magnitude of the response $\tilde{H}_d$ was designed to improve the broadband output SNR, whereas the phase term of $\tilde{H}_d$ was set equal to the phase of the ATF in the reference microphone. In contrast to the approach in [7], we consider a complex-valued selection vector $\mathbf{u}$ which enables different phase references. In the following, we demonstrate that $\tilde{H}_d$ can be considered as the overall transfer function.

3.1 MWF overall transfer function

According to [5] and many others, the MWF in (10) can be decomposed using the matrix inversion lemma as

$$\mathbf{G} = \frac{P_S}{P_S + \mu(\boldsymbol{H}^\dagger \boldsymbol{R}_N^{-1} \boldsymbol{H})^{-1}} \frac{\boldsymbol{R}_N^{-1}\boldsymbol{H}}{\boldsymbol{H}^\dagger \boldsymbol{R}_N^{-1}\boldsymbol{H}} \tilde{H}_d^* \tag{13}$$

$$= G_{WF}\, \mathbf{G}_{MVDR}\, \tilde{H}_d^*, \tag{14}$$

i.e., a MVDR beamformer

$$\mathbf{G}_{MVDR} = \frac{\boldsymbol{R}_N^{-1}\boldsymbol{H}}{\boldsymbol{H}^\dagger \boldsymbol{R}_N^{-1}\boldsymbol{H}}, \tag{15}$$

a filter $\tilde{H}_d$, and a single-channel Wiener post filter

$$G_{WF} = \frac{P_S}{P_S + \mu(\boldsymbol{H}^\dagger \boldsymbol{R}_N^{-1} \boldsymbol{H})^{-1}}. \tag{16}$$

Without noise reduction, i.e., for $\mu = 0$, the overall transfer function equals $\tilde{H}_d$, because $\mathbf{G}^{\text{MVDR}}$ has a unity gain transfer function. The output signal can be written as

$$Z_S = \tilde{H}_d \cdot S. \tag{17}$$

In the following, we consider some special cases of the G-MWF. Note that the different formulations of the G-MWF differ only with respect to the vector $\mathbf{u}$ and the corresponding transfer function $\tilde{H}_d$.

3.2 MVDR beamformer

The MVDR beamformer obtains perfect equalization of the acoustic system, where the overall transfer function is chosen to be $\tilde{H}_d = 1$. Hence, the elements of the vector $\mathbf{u}$ are

$$u_i = \frac{H_i}{\boldsymbol{H}^\dagger \boldsymbol{H}}. \tag{18}$$

However, the resulting G-MWF requires perfect knowledge about the ATF from the speaker to the microphones. The corresponding issue of blind channel estimation is a challenging task in noisy environments and so far an unsolved problem. A further issue is the inversion of the squared norm of the ATFs, since they may contain zeros in their magnitude response.

3.3 Selection of a reference channel

In the S-MWF, the overall transfer function $\tilde{H}_d$ is equal to the ATF from the speaker to one of the microphones, i.e., $\tilde{H}_d = H_{\text{ref}}$ where ref denotes the index of the reference microphone. In this case, the numerator of the S-MWF can be written as

$$\boldsymbol{R}_S \boldsymbol{u} = P_S \boldsymbol{H} \boldsymbol{H}^\dagger \boldsymbol{u} = P_S \boldsymbol{H} \tilde{H}_{\text{ref}}^*, \tag{19}$$

where $\mathbf{u}$ is a column vector of length M that selects the reference microphone, i.e., the corresponding entry is equal to one, while all other entries are equal to zero. As a result, the corresponding ATF remains as the overall transfer function.

Compared to the MVDR beamformer in Section 3.2, the advantage of the S-MWF is that it only depends on estimates of the signal statistics, i.e., $\boldsymbol{R}_S$ and $\boldsymbol{R}_N$ and no explicit knowledge of the ATFs is required. However, it should be noted that the output signal is as reverberant as the input signal.

3.4 Partial equalization approach

In [5], the P-MWF has been presented, where the amplitude of the overall transfer function is defined as the

envelope of the individual ATFs, and the phase is chosen as the phase ϕ_{ref} of an arbitrary (reference) ATF, i.e.,

$$\tilde{H}_d = \sqrt{\boldsymbol{H}^\dagger \boldsymbol{H}}\, e^{j\phi_{\text{ref}}}. \tag{20}$$

This formulation results in a partial equalization of the acoustic system, since the dips in the magnitude response of the individual ATFs can be avoided. The elements of the vector $\mathbf{u}$ can be computed as

$$u_i = \sqrt{\frac{r_{S_{i,i}}}{\text{tr}(\boldsymbol{R}_S)}}\frac{r_{S_{i,\text{ref}}}}{|r_{S_{i,\text{ref}}}|} = \frac{H_i}{\sqrt{\boldsymbol{H}^\dagger \boldsymbol{H}}}\, e^{-j\phi_{\text{ref}}}, \tag{21}$$

where $\text{tr}(\cdot)$ denotes the trace of the matrix and $r_{S_{i,j}}$ denotes the element of $\mathbf{R}_S$ in the ith row and jth column. Hence, for the P-MWF, we have

$$\boldsymbol{R}_S \boldsymbol{u} = \boldsymbol{R}_S \frac{\boldsymbol{H}}{\sqrt{\boldsymbol{H}^\dagger \boldsymbol{H}}}\, e^{-j\phi_{\text{ref}}} = P_S \boldsymbol{H} \sqrt{\boldsymbol{H}^\dagger \boldsymbol{H}}\, e^{-j\phi_{\text{ref}}}. \tag{22}$$

Similar to the S-MWF, the P-MWF only depends on the signal statistics and therefore no explicit knowledge of the ATFs is required. It should be noted that the phase of the output speech component is equal to the phase of the reverberant speech component in the reference microphone signal. As a result, the P-MWF approach equalizes the amplitude of the desired overall transfer function, but the output signal is as reverberant as the selected microphone signal.

4 Output SNR

In this section, we investigate the narrow-band and broadband output SNR of the different MWF formulations. Firstly, we consider the narrow-band output SNR

$$\gamma(\nu) = \frac{\mathbb{E}\left\{|Z_S(\nu)|^2\right\}}{\mathbb{E}\left\{|Z_N(\nu)|^2\right\}} = \frac{\boldsymbol{G}^\dagger \boldsymbol{R}_S \boldsymbol{G}}{\boldsymbol{G}^\dagger \boldsymbol{R}_N \boldsymbol{G}}. \tag{23}$$

Using Eq. (14), we have

$$\begin{aligned}\gamma(\nu) &= \frac{\left(G_{\text{WF}}\, \mathbf{G}_{\text{MVDR}}\, \tilde{H}_d^*\right)^\dagger \boldsymbol{R}_S \left(G_{\text{WF}}\, \mathbf{G}_{\text{MVDR}}\, \tilde{H}_d^*\right)}{\left(G_{\text{WF}}\, \mathbf{G}_{\text{MVDR}}\, \tilde{H}_d^*\right)^\dagger \boldsymbol{R}_N \left(G_{\text{WF}}\, \mathbf{G}_{\text{MVDR}}\, \tilde{H}_d^*\right)} \\ &= \frac{|G_{\text{WF}}|^2 |\tilde{H}_d|^2 \mathbf{G}_{\text{MVDR}}^\dagger \boldsymbol{R}_S \mathbf{G}_{\text{MVDR}}}{|G_{\text{WF}}|^2 |\tilde{H}_d|^2 \mathbf{G}_{\text{MVDR}}^\dagger \boldsymbol{R}_N \mathbf{G}_{\text{MVDR}}} \\ &= \frac{\mathbf{G}_{\text{MVDR}}^\dagger \boldsymbol{R}_S \mathbf{G}_{\text{MVDR}}}{\mathbf{G}_{\text{MVDR}}^\dagger \boldsymbol{R}_N \mathbf{G}_{\text{MVDR}}}.\end{aligned} \tag{24}$$

Consequently, the narrow-band output SNR is independent of the particular choice of $\tilde{H}_d$. Nevertheless, $\tilde{H}_d$ impacts the broadband output SNR, which is defined as

$$\begin{aligned}\gamma_{\text{out}} &= \frac{\sum_\nu \mathbb{E}\left\{|Z_S(\nu)|^2\right\}}{\sum_\nu \mathbb{E}\left\{|Z_N(\nu)|^2\right\}} \\ &= \frac{\sum_\nu \boldsymbol{G}(\nu)^\dagger \boldsymbol{R}_S(\nu) \boldsymbol{G}(\nu)}{\sum_\nu \boldsymbol{G}(\nu)^\dagger \boldsymbol{R}_N(\nu) \boldsymbol{G}(\nu)}.\end{aligned} \tag{25}$$

Note that the PSD of the speech component at the output of the MVDR beamformer is P_S. Hence, the PSD of the speech component Z_S at the output of the G-MWF is $\mathbb{E}\left\{|Z_S(\nu)|^2\right\} = |G_{WF}|^2 |\tilde{H}_d|^2 P_S$. Similarly, the PSD of the noise component at the output of the MVDR beamformer is $P_{N,\text{MVDR}} = \mathbf{G}_{\text{MVDR}}^\dagger \boldsymbol{R}_N \mathbf{G}_{\text{MVDR}}$, such that the PSD of the noise component at the output of the G-MWF is $\mathbb{E}\left\{|Z_N(\nu)|^2\right\} = |G_{WF}|^2 |\tilde{H}_d|^2 P_{N,\text{MVDR}}$ and

$$\gamma_{\text{out}} = \frac{\sum_\nu P_S(\nu) |G_{\text{WF}}(\nu)|^2 |\tilde{H}_d(\nu)|^2}{\sum_\nu P_{N,\text{MVDR}}(\nu) |G_{\text{WF}}(\nu)|^2 |\tilde{H}_d(\nu)|^2}. \tag{26}$$

From this equation, it can be seen that the overall transfer function as well as the single-channel Wiener post filter impact the broadband output SNR.

Next, we consider the response $\tilde{H}_d$ that maximizes the broadband output SNR. Equation (26) can be written as

$$\gamma_{\text{out}} = \frac{\sum_\nu \alpha_\nu |\tilde{H}_d(\nu)|^2}{\sum_\nu \beta_\nu |\tilde{H}_d(\nu)|^2} = \frac{\tilde{\mathbf{H}}^\dagger \mathbf{A} \tilde{\mathbf{H}}}{\tilde{\mathbf{H}}^\dagger \mathbf{B} \tilde{\mathbf{H}}}. \tag{27}$$

with

$$\begin{aligned}\alpha_\nu &= P_S(\nu) |G_{\text{WF}}(\nu)|^2 \\ \beta_\nu &= P_{N,\text{MVDR}}(\nu) |G_{\text{WF}}(\nu)|^2 \\ \tilde{\mathbf{H}} &= [\tilde{H}_d(0), \ldots, \tilde{H}_d(F-1)]^T \\ A &= \begin{pmatrix} \alpha_0 & 0 & \ldots & 0 \\ 0 & \alpha_1 & \ldots & 0 \\ & & & \\ 0 & \ldots & \ddots & 0 \\ 0 & \ldots & 0 & \alpha_{F-1} \end{pmatrix} \\ B &= \begin{pmatrix} \beta_0 & 0 & \ldots & 0 \\ 0 & \beta_1 & \ldots & 0 \\ & & & \\ 0 & \ldots & \ddots & 0 \\ 0 & \ldots & 0 & \beta_{F-1} \end{pmatrix}.\end{aligned}$$

where F denotes the total number of frequency bins. Maximizing γ_{out} is equivalent to solving the generalized eigenvalue problem $\mathbf{A}\tilde{\mathbf{H}} = \lambda \mathbf{B}\tilde{\mathbf{H}}$ or $\mathbf{B}^{-1}\mathbf{A}\tilde{\mathbf{H}} = \lambda \tilde{\mathbf{H}}$. The solution to the eigenvalue problem is the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}$. Since $\mathbf{B}^{-1}\mathbf{A}$ is a diagonal matrix, the largest eigenvalue is

$$\lambda_{\max} = \max_\nu \frac{\alpha_\nu}{\beta_\nu} = \max_\nu \frac{P_S(\nu)}{P_{N,\text{MVDR}}(\nu)}. \tag{28}$$

Comparing Eqs. (28) with (26), we obtain the corresponding eigenvector $\tilde{\mathbf{H}} = [0, \ldots, 1, \ldots, 0]^T$, with a one in the frequency bin corresponding to the largest eigenvalue and zero elsewhere. Although this overall transfer function maximizes the broadband output SNR, the corresponding speech distortion will not be acceptable, because only one frequency bin will pass the beamformer.

Hence, we conclude that the design of the desired response $\tilde{H}_d$ requires additional constraints on the speech distortion. The optimal solution with respect to speech distortion is the MVDR beamformer which is, however, hardly attainable in practice.

5 MWF reference selection

It was shown in [19] that the temporal smearing and therefore the reverberation relies on the all-pass component of the overall ATF. This suggests that a suitable phase reference can improve the output SRR. In this section, we present two formulations of the G-MWF that improve the SRR and the broadband output SNR compared with the S-MWF or the P-MWF. Both formulations use a phase reference from a DSB, which delays the microphone signals to compensate for the different times of arrival. Hence, the DSB enhances the direct path component and, as we will see in Section 7, improves the SRR.

5.1 Delay-and-sum beamformer

In the first approach, we propose to simply use the output of a delay-and-sum beamformer as the speech reference. The corresponding elements of the vector $\mathbf{u}$ can be described as

$$u_i = \frac{1}{M} \cdot e^{j2\pi \frac{\nu}{F} \tau_i} \text{ for } \nu \in 0, \ldots, F-1, \tag{29}$$

where τ_i is a delay (in samples), which compensates the TDOA of the direct path speech components at the microphones. The speech components are typically aligned to the microphone with the latest arrival time to obtain a causal DSB. Using (12) we obtain the overall transfer function

$$\tilde{H}_d = \frac{1}{M} \sum_i H_i e^{-j2\pi \frac{\nu}{F} \tau_i}. \tag{30}$$

5.2 Partial equalization with DSB phase reference

The second approach is a combination of the P-MWF with the DSB as the phase reference. As already described in Section 3.4, the phase reference of the P-MWF is the phase of an arbitrary ATF. In order to improve the SRR, we can use the DSB as the phase reference. The resulting vector $\mathbf{u}$ can be described as

$$u_i = \sqrt{\frac{r_{S_{i,i}}}{\mathrm{tr}(\boldsymbol{R}_S)}} \cdot e^{j2\pi \frac{\nu}{F} \tau_i} \text{ for } \nu \in 0, \ldots, F-1. \tag{31}$$

Note that the phase term impacts the magnitude of the overall transfer function $\tilde{H}_d$, cf. (12). Comparing (21) and (31), we have $u_i = \frac{|H_i|}{\sqrt{\boldsymbol{H}^\dagger \boldsymbol{H}}} e^{j2\pi \frac{\nu}{F} \tau_i}$ and

$$\tilde{H}_d = \frac{1}{\sqrt{\boldsymbol{H}^\dagger \boldsymbol{H}}} \sum_i |H_i| H_i e^{-j2\pi \frac{\nu}{F} \tau_i}. \tag{32}$$

Hence, the direct path speech components in the microphones are aligned, but additionally the microphone signals are weighted with the magnitude of the ATFs similar to the P-MWF approach.

6 System structure of the G-MWF

Figure 1 depicts the block diagram of the G-MWF for an array with two microphones. Since the filtering is performed in the frequency domain, the microphone signals are first windowed and then transformed using the fast Fourier transform (FFT).

A frequency-dependent voice activity detector (VAD) as proposed in [32] is used to estimate the required correlation matrices for the G-MWF. During speech pauses and in frequency bins where no speech activity is detected, the estimate of the noise correlation matrix $\boldsymbol{R}_N$ is updated. The estimate of the speech correlation matrix $\mathbf{R}_S$ is obtained from the input correlation matrix $\mathbf{R}_Y$ as

$$\mathbf{R}_S = \mathbf{R}_Y - \mathbf{R}_N. \tag{33}$$

Furthermore, for the phase reference proposed in Section 5, the TDOA from the speaker to the microphones is required, to achieve a coherent summation of the microphone signals. Depending on the TDOA, a suitable vector $\mathbf{u}$ is derived to compensate the phase differences of the microphone signals, as calculated in Eq. (29). A very popular TDOA estimation approach

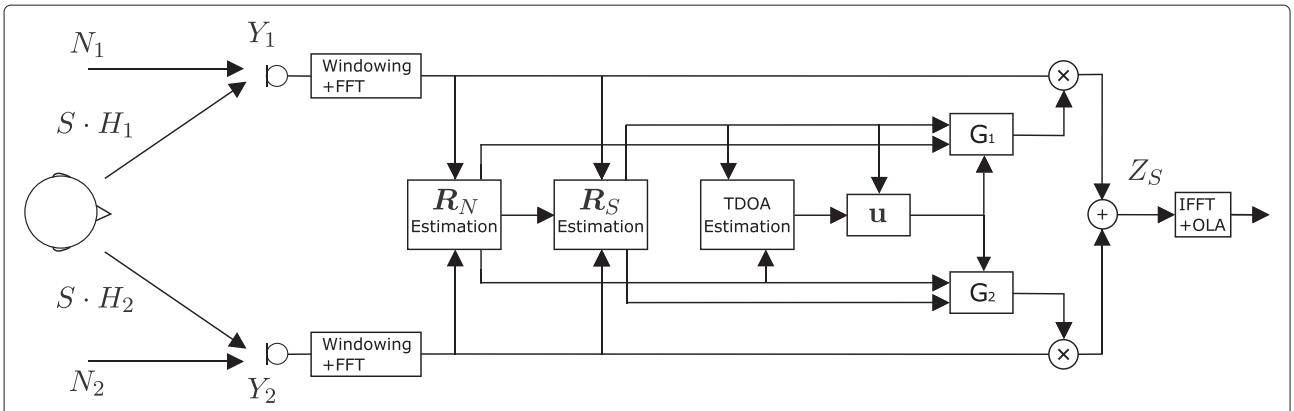

Fig. 1 System structure for multichannel Wiener filtering with vector $\mathbf{u}$

is the generalized cross correlation (GCC) method [23, 28, 29], where the cross-correlation between the microphone signals is calculated in the frequency domain as the cross power spectral density (CPSD). Depending on the application and the environmental conditions, the CPSD is typically weighted with a coherence or noise-based weighting using the magnitude spectrum of the CPSD. The weighted CPSD is transformed to the time domain using the inverse Fourier transform, resulting in the cross correlation vector. The main peak in the cross correlation vector indicates the time delay. It should be noted that the TDOA estimate is only valid in signal blocks where the speaker is active, which can be determined based on a VAD.

It should be noted that the phase of the CPSD is equal to the phase of the relative transfer function (RTF) between the microphones, since both only differ from a different magnitude response. Since in general the microphone signals contain correlated noise components, estimating the RTFs directly from the noisy microphone signals leads to biased RTF estimates. Several methods for unbiased RTF estimation have been proposed, e.g., by exploiting the non-stationarity of speech signals [13, 33] or based on the generalized eigenvalue decomposition of $\boldsymbol{R}_Y$ and $\boldsymbol{R}_N$ [34, 35]. In [36], an approach for unbiased RTF estimation was proposed, requiring estimates of the PSDs and CPSDs of the speech and noise components, which can be obtained from the estimated speech and noise correlation matrices $\boldsymbol{R}_S$ and $\boldsymbol{R}_N$. The RTF estimate between microphones i and j is computed as a combination of two weighted coefficients

$$\hat{W}_{\text{unbiased}} = f_i \frac{r_{S_{i,j}}}{r_{S_{i,i}}} + f_j \frac{r_{S_{j,j}}}{r_{S_{j,i}}}, \tag{34}$$

where the terms f_i and f_j are SNR-based weighting coefficients which are defined as

$$f_i = \frac{\frac{r_{S_{i,i}}}{r_{N_{i,i}}}}{\frac{r_{S_{i,i}}}{r_{N_{i,i}}} + \frac{r_{S_{j,j}}}{r_{N_{j,j}}}} \tag{35}$$

$$f_j = \frac{\frac{r_{S_{j,j}}}{r_{N_{j,j}}}}{\frac{r_{S_{i,i}}}{r_{N_{i,i}}} + \frac{r_{S_{j,j}}}{r_{N_{j,j}}}}. \tag{36}$$

We propose a slightly modified approach based on frequency-dependent VAD [32], where the RTF estimate is updated only in frequency bins where speech activity is detected. Furthermore, a smoothing parameter to average the RTF estimate is used, which is the rate of all frequency bins where speech activity is detected. By applying the inverse Fourier transform, $\hat{W}_{\text{unbiased}}$ can be transformed back into the time domain, which results in the vector $\hat{w}_{\text{unbiased}}$. The location of the peak value that indicates the delay to the microphone j can be calculated as

$$\tau_i = \arg \max_{n=0,\ldots,F-1} \hat{w}_{\text{unbiased}}(n), \tag{37}$$

where $\hat{w}_{\text{unbiased}}(n)$ is the nth element of the vector $\hat{w}_{\text{unbiased}}$.

7 Simulation results

To verify the SRR and SNR improvements provided by the proposed approaches, different simulations were carried out. In the following, G-MWF-1 denotes the G-MWF that uses the DSB as the speech reference, i.e., (29), whereas G-MWF-2 denotes the partial equalization approach, using the DSB only as a phase reference, i.e., (31). For the S-MWF and the P-MWF, the first microphone was used as the reference. All simulations were performed with a sampling rate of 16 kHz and an FFT length $F = 512$. We consider a noisy car environment as well as a reverberant classroom. The signals for testing the algorithms are ITU speech signals convolved with measured impulse responses. For the car scenario, this was done with an artificial head and two cardioid microphones that were mounted close to the rear-view mirror. For the classroom scenario [37] impulse responses were recorded with a loudspeaker and omnidirectional microphones at two different spatial locations with a microphone distance of 0.5 m. The reverberation time RT_{60} of the classroom has a value between 1.5 and 1.8 s over all frequencies. To evaluate the dereverberation capabilities of the algorithms, the energy decay curves (EDCs) [38] of the resulting overall transfer functions $\tilde{H}_d$ using the measured impulse responses were calculated (for $\mu = 0$). For the car environment, the resulting EDCs are shown in Fig. 2.

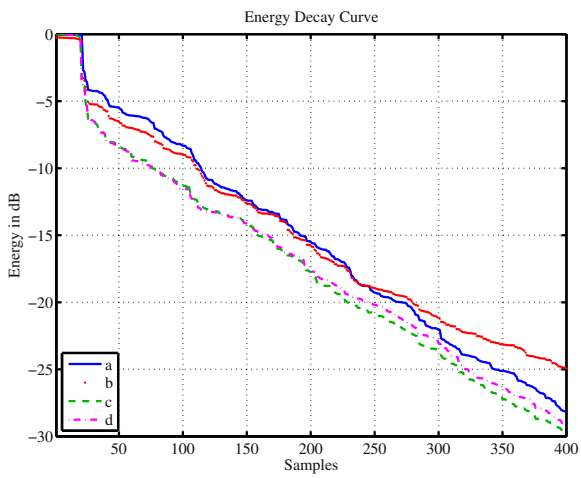

Fig. 2 EDC of the resulting acoustic transfer functions of the car environment: (*a*) ATF from the speech signal source to microphone 1 (S-MWF), (*b*) overall transfer function of P-MWF with phase reference of microphone 1, (*c*) overall transfer function of G-MWF-1, (*d*) overall transfer function of G-MWF-2

Curve (a) depicts the EDC of the overall transfer function for the S-MWF. Curve (b) depicts the resulting EDC of the overall transfer function of the P-MWF. Compared with (a), it can be observed that the decay time is increased, but the energy of the first reflections is reduced due to the partial equalization as can be seen from the first 230 samples of the EDC. Curves (c) and (d) depict the EDC of the overall transfer function for the G-MWF-1 and G-MWF-2, respectively. Compared with (a) and (b), a reduced decay time is observed due to the coherent combining of the phase terms. As a result, the direct components of the ATF are enhanced, which leads to an improvement in speech quality of the overall system.

For the classroom scenario, the resulting EDCs are shown in Fig. 3. Due to the longer reverberation time, compared with the car environment, the resulting EDCs show a different behavior. Curves (e) and (f) depict the EDCs of the resulting transfer function for the S-MWF and the P-MWF, respectively. Curves (g) and (h) depict the EDCs of the overall transfer functions for the G-MWF-1 and the G-MWF-2. Compared to (e), it can be observed in (f) that the direct signal component for the first few samples is augmented, due to the partial equalization, but that the decay time is increased. While (h) still shows a slightly better performance than (g) for the first 7000 samples, the decay time is increased by a small amount compared with (h) during the samples 7000–10,000. However, the reverberation energy for the G-MWF-1 and G-MWF-2 in (g) and (h) is noticeably reduced compared with (e) and (f).

As a measure of reverberation, the direct-to-reverberation ratio (DRR) can be calculated from the resulting overall transfer functions $\tilde{H}_d$. The DRR is defined as [39]

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_d} h_d^2(n)}{\sum_{n=n_d+1}^{\infty} h_d^2(n)}\right)\mathrm{dB}, \tag{38}$$

where h_d is the impulse response of the overall transfer function $\tilde{H}_d$ in the time domain and n_d are the samples of the direct path. For n_d, we considered a time interval of 8 ms after the first arrival of the direct sound. In Table 1, the DRR values for the different overall transfer functions $\tilde{H}_d$ are presented. From the table, it can be seen that the G-MWF approaches improve the DRR in both scenarios compared with the S-MWF and P-MWF.

Both versions of the G-MWF result in similar overall transfer functions. This can be observed in Figs. 4 and 5. Figure 4 presents the magnitude response of the ATFs of the car environment for both microphones as well as the overall transfer function of G-MWF-2 for frequencies between 2600 and 4000 Hz. Clearly, the resulting partial equalization of the G-MWF-2 can be seen. Figure 5 depicts the overall transfer function of both G-MWF versions for the same frequency section. It is shown that the magnitude response of both approaches looks quite similar.

Finally, we consider a noisy car scenario. The noise was recorded at a driving speed of 100 km/h with the same microphone setup as specified above. For $\mu > 0$, the MWF performs an adaptive noise reduction and therefore the resulting overall transfer function is time varying. As a result, signal-based performance measures for the noise reduction and dereverberation performance need to be used. For the dereverberation performance, the signal-to-reverberation ratio (SRR) after [39] is used, i.e.,

$$\mathrm{SRR} = 10\log_{10}\left(\frac{\mathbb{E}\left\{|s_d(k)|^2\right\}}{\mathbb{E}\left\{|\hat{s}(k) - s_d(k)|^2\right\}}\right)\mathrm{dB}, \tag{39}$$

where $s_d(k)$ is the direct path signal component of the first microphone and $\hat{s}(k)$ is the output signal of the beamformer in the time domain. It should be noted that this measure is only valid for signal segments, where speech activity is detected.

Table 2 presents the results for the SRR and the broadband output SNR for two settings of the trade-off parameter μ, where a larger value of μ results in more noise

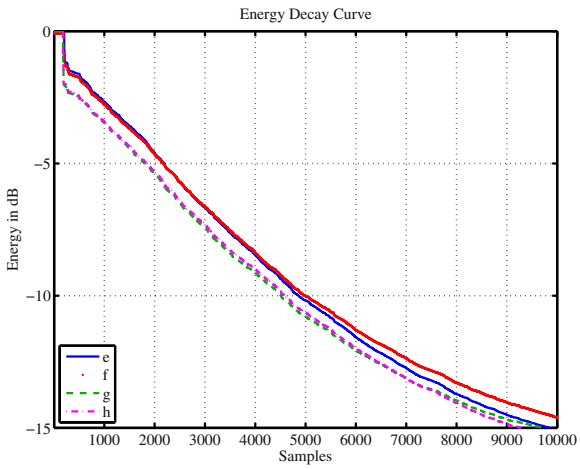

Fig. 3 EDC of the resulting acoustic transfer functions of the classroom environment: (*e*) ATF from the speech signal source to microphone 1 (S-MWF), (*f*) overall transfer function of P-MWF with phase reference of microphone 1, (*g*) overall transfer function of G-MWF-1, (*h*) overall transfer function of G-MWF-2

Table 1 DRR of the overall transfer function for choosing a different phase and magnitude reference

	S-MWF	P-MWF	G-MWF1	G-MWF2
Car scenario	12.6 dB	9.3 dB	14.7 dB	14.3 dB
Classroom scenario	−3.8 dB	−3.7 dB	−1.4 dB	−1.7 dB

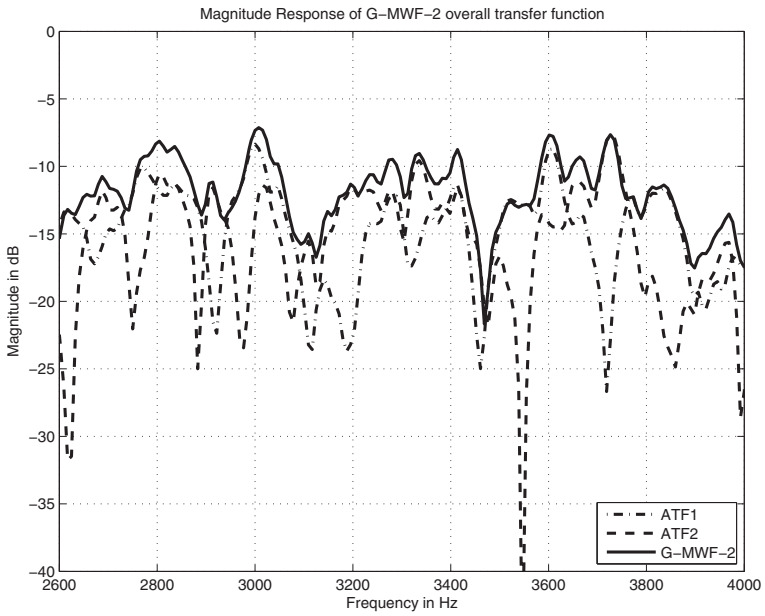

Fig. 4 Magnitude response of G-MWF2

reduction. The SRR was measured in time frames where speech was present. The performance of both G-MWF approaches are compared with the S-MWF and P-MWF. It can be observed that both G-MWF approaches outperform the S-MWF in terms of SRR and SNR. G-MWF-1 outperforms the P-MWF in terms of SRR and SNR, whereas G-MWF-2 improves the SRR compared to G-MWF-1 at the expense of a small SNR loss.

8 Conclusions

For the multichannel Wiener filter, the influence of the phase reference is often neglected, because it has no impact on the narrow-band output SNR. In this work, we have shown that the phase reference influences the overall transfer function. Moreover, the overall transfer function determines the speech distortion and impacts the broad-band output SNR. We have proposed two generalized

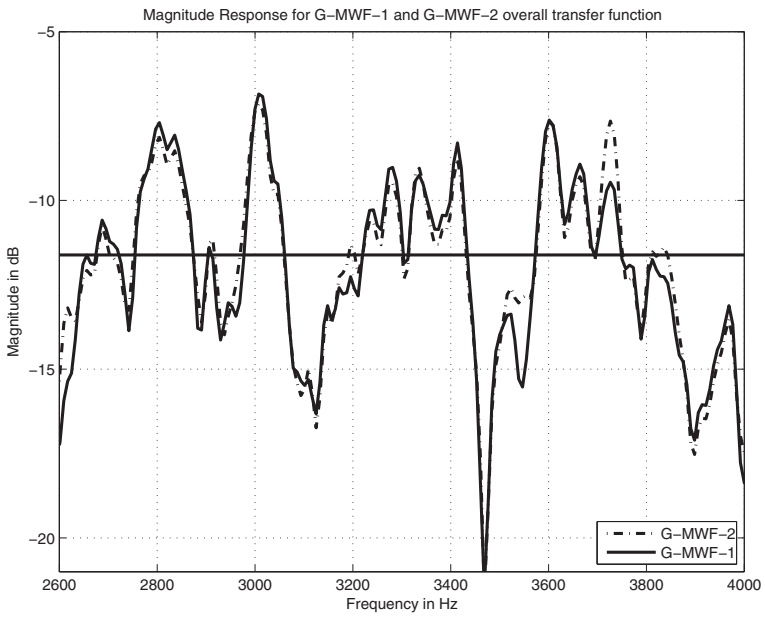

Fig. 5 Comparison of G-MWF1 and G-MWF2

Table 2 SRR and SNR comparison for different MWF formulations

$\mu = 0$	SNR	SRR
S-MWF	−1.94 dB	2.87 dB
P-MWF	−0.86 dB	2.29 dB
G-MWF1	−0.72 dB	4.69 dB
G-MWF2	−1.33 dB	5.86 dB
$\mu = 30$	SNR	SRR
S-MWF	2.82 dB	1.66 dB
P-MWF	4.35 dB	1.81 dB
G-MWF1	4.90 dB	3.49 dB
G-MWF2	4.25 dB	5.08 dB

formulations for the MWF where the phase reference is based on the phase of a delay-and-sum beamformer. The proposed G-MWF technique requires an estimate of the time-difference-of-arrival, which can be acquired from the estimates of the speech and noise correlation matrices. Thus, the G-MWF requires only information about the second order statistics of the signals. The presented simulation results indicate that both G-MWF versions can achieve a better signal-to-reverberation ratio and an improvement in broadband output SNR compared to previously known MWF formulations.

Competing interests
The authors declare that they have no competing interest.

Author details
[1]HTWG Konstanz, University of Applied Sciences Institute for System Dynamics—Signal Processing Group, 78462 Konstanz, Germany. [2]University of Oldenburg Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, 26111 Oldenburg, Germany.

Received: 16 December 2015 Accepted: 20 June 2016
Published online: 07 July 2016

References

1. S Wehr, I Kozintsev, R Lienhart, W Kellermann, in *Proceedings of IEEE Sixth International Symposium on Multimedia Software Engineering*. Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation (IEEE, 2004), pp. 18–25
2. S Doclo, M Moonen, T Van den Bogaert, J Wouters, Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids. IEEE Transac. Audio, Speech, and Language Processing. **17**(1), 38–51 (2009)
3. TC Lawin-Ore, S Doclo, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Analysis of rate constraints for MWF-based noise reduction in acoustic sensor networks (IEEE, 2011), pp. 269–272
4. S Stenzel, J Freudenberger. Blind matched filtering for speech enhancement with distributed microphones, vol. 2012, (2012), p. 15. Article ID 169853
5. S Stenzel, TC Lawin-Ore, J Freudenberger, S Doclo, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. A multichannel Wiener filter with partial equalization for distributed microphones, (Mohonk Mountain House, New Paltz, NY, 2013)
6. TC Lawin-Ore, S Stenzel, J Freudenberger, S Doclo, in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*. Alternative formulation and robustness analysis of the multichannel Wiener filter for spatially distributed microphones, (Antibes, France, 2014), pp. 208–212
7. TC Lawin-Ore, S Stenzel, J Freudenberger, S Doclo, in *Proc. ITG Conference on Speech Communication*. Generalized multichannel Wiener filter for spatially distributed microphones, (Erlangen, Germany, 2014), pp. 1–4
8. TC Lawin-Ore, S Doclo, Analysis of the average performance of the multichannel Wiener filter based noise reduction using statistical room acoustics. Signal Process. **107**, 96–108 (2015)
9. S Markovich-Golan, A Bertrand, M Moonen, S Gannot, Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks. Signal Process. **107**, 4–20 (2015)
10. J Schmalenstroeer, P Jebramcik, R Haeb-Umbach, A combined hardware-software approach for acoustic sensor network synchronization. Signal Process. **107**, 171–184 (2015)
11. S Miyabe, N Ono, S Makino, Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation. Signal Process. **107**, 185–196 (2015)
12. L Wang, S Doclo, Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation. IEEE/ACM Trans. Audio, Speech Lang. Process. **24**, 571–582 (2016)
13. S Gannot, D Burshtein, E Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Transac. Signal Process. **49**(8), 1614–1626 (2001)
14. EAP Habets, J Benesty, I Cohen, S Gannot, J Dmochowski, New insights into the MVDR beamformer in room acoustics. IEEE Transactions on Audio, Speech, Language Process. **18**(1), 158–170 (2010)
15. J Chen, J Benesty, Y Huang, S Doclo, New insights into the noise reduction Wiener filter. IEEE Transac. Audio, Speech Lang. Process. **14**(4), 1218–1234 (2006)
16. S Doclo, A Spriet, J Wouters, M Moonen, Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction. Speech Commun. **49**(7–8), 636–656 (2007)
17. TC Lawin-Ore, S Doclo, in *Proceedings of 10. ITG Symposium on Speech Communication*. Reference microphone selection for MWF-based noise reduction using distributed microphone arrays (VDE, Braunschweig, 2012), pp. 31–34
18. JB Allen, DA Berlkey, J Blauert, Multimicrophone signal-processing technique to remove room reverberation from speech signals. J. Acoust. Soc. Am. **62**(4), 912–915 (1977)
19. Q-G Liu, B Champagne, P Kaba, Room speech dereverberation via minimum-phase and all-pass component processing of multi-microphone signals. IEEE Pacific Rim Conf. Commun. Comput. Signal Process, 571–574 (1995)
20. EAP Benesty, J Habets, A two-stage beamforming approach for noise reduction and dereverberation. IEEE Transac. Audio, Speech, Language Process. **21**(5), 945–958 (2013)
21. S Grimm, J Freudenberger, in *Jahrestagung für Akustik (DEGA)*. A phase reference for a multichannel Wiener filter by a delay and sum beamformer, (Nürnberg, 2015), pp. 208–212
22. I Kodrasi, S Doclo, Joint dereverberation and noise reduction based on acoustic multichannel equalization. IEEE Transac. Audio, Speech, Lang. Process. **24**(4), 680–9693 (2016)
23. C Knapp, G Carter, The generalized correlation method for estimation of time delay. IEEE Transac. Acoust. Speech Signal Process. **24**(4), 320–327 (1976)
24. GC Carter, *Coherence and time delay estimation: an applied tutorial for research, development, test and evaluation engineers*. (IEEE Press, 1993)
25. S Doclo, M Noonen, Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. EURASIP J. Appl Signal Process. **11**, 1110–1124 (2003)
26. MS Brandstein, HF Silverman, A robust method for speech signal time-delay estimation in reverberant rooms. Proc. IEEE Int. Conf. Acoust. Speech Signal Process. **1**, 375–378 (1997)
27. TG Dvorkind, S Gannot, Time difference of arrival estimation of speech source in a noisy and reverberant environment. Signal Process. **85**(1), 177–204 (2005)
28. J Chen, J Benesty, Y (Arden) Huang, Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments. EURASIP J. Appl. Signal Process. **2005**(1), 25–36 (2005)

29. J Chen, J Benesty, Y Huang, Time delay estimation in room acoustic environments: an overview. EURASIP J. Appl. Signal Process. **2006**, 1–19 (2006)
30. TG Manickam, RJ Vaccaro, DW Tufts, A least-squares algorithm for multipath time-delay estimation. IEEE Transac. Signal Process. **42**(11), 3229–3233 (1994)
31. S Doclo, A Spriet, M Moonen, J Wouters, in *Speech Enhancement*. Speech distortion weighted multichannel Wiener filtering techniques for noise reduction, chapter 9 (Springer, Berlin/Heidelberg, 2005)
32. J Freudenberger, S Stenzel, in *IEEE Workshop on Statistical Sig. Proc. (SSP), Nice*. Time-frequency dependent voice activity detection based on a simple threshold test (IEEE, Nice, 2011)
33. I Cohen, Relative transfer function identification using speech signals. Speech Audio Process. IEEE Transac. **12**(5), 451–459 (2004)
34. S Markovich-Golan, S Gannot, I Cohen, Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Transac. Audio, Speech, Lang. Process. **17**(6), 1071–1086 (2009)
35. S Markovich-Golan, S Gannot, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method (IEEE, South Brisbane, 2015), pp. 544–548
36. M Schwab, P Noll, T Sikora, in *European Signal Processing Conference (EUSIPCO)*. Noise robust relative transfer function estimation, vol. 2 (IEEE, Florence, 2006), pp. 1–5
37. R Stewart, M Sandler, in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Database of omnidirectional and B-format room impulse responses (IEEE, Dallas, 2010), pp. 165–168
38. MR Schroeder, Frequency correlation functions of frequency responses in rooms. J. Acoust. Soc. Am. **34**(2), 1819–1823 (1962)
39. PA Naylor, ND Gaubitch, *Speech Dereverberation*, 1st edn. (Springer, London, 2010)