Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 169853, 15 pages
doi:10.1155/2012/169853

Research Article

Blind-Matched Filtering for Speech Enhancement with Distributed Microphones

Sebastian Stenzel and Jürgen Freudenberger

Institute for System Dynamics, HTWG Konstanz, Brauneggerstrasse 55, 78462 Konstanz, Germany

Correspondence should be addressed to Sebastian Stenzel, sebastian.stenzel@htwg-konstanz.de

Received 4 May 2012; Revised 18 July 2012; Accepted 2 August 2012

Academic Editor: Sriram Srinivasan

A multichannel noise reduction and equalization approach for distributed microphones is presented. The speech enhancement is based on a blind-matched filtering algorithm that combines the microphone signals such that the output SNR is maximized. The algorithm is developed for spatially uncorrelated but nonuniform noise fields, that is, the noise signals at the different microphones are uncorrelated, but the noise power spectral densities can vary. However, no assumptions on the array geometry are made. The proposed method will be compared to the speech distortion-weighted multichannel Wiener filter (SDW-MWF). Similar to the SDW-MWF, the new algorithm requires only estimates of the input signal to noise ratios and the input cross-correlations. Hence, no explicit channel knowledge is necessary. A new version of the SDW-MWF for spatially uncorrelated noise is developed which has a reduced computational complexity, because matrix inversions can be omitted. The presented blind-matched filtering approach is similar to this SDW-MWF for spatially uncorrelated noise but additionally achieves some improvements in the speech quality due to a partial equalization of the acoustic system.

1. Introduction

In many speech communication systems, like hands-free car kits, teleconferencing systems, and speech recognition systems, the desired speech signal is linearly distorted by the room acoustics and also corrupted by undesired background noise. Therefore, efficient speech processing techniques are required to enhance the speech signal under the constraint of a small speech distortion. The use of multiple microphones can improve the performance compared to single microphone systems [1]. The most common way to place the microphones is beamformer arrays with designed array geometry. Beamforming algorithms exploit the spatial directivity effects by a proper combining, like the Frost beamformer [2] or the generalized sidelobe canceler (GSC) [3]. Usually, the microphones are located in close proximity and the same signal conditions at the microphone positions are assumed.

Alternatively, multimicrophone setups have been proposed that combine the processed signals of two or more distributed microphones. The microphones are positioned separately in order to ensure incoherent recording of noise [4–6]. Basically, all these approaches exploit the fact that speech components in the microphone signals are strongly correlated while the noise components are only weakly correlated if the distance between the microphones is sufficiently large.

For immersive communication, future communication devices must be able to collect the desired speech signal as naturally as possible. But the speech signal quality depends on the speaker's distance to the microphone (array). Therefore, we propose the use of a setup with distributed microphones, where the user can place the microphones arbitrarily. Hence, the array geometry is arbitrary and not a priori known.

In this paper, we discuss schemes for an optimal speech signal combining in real-world acoustic scenarios with distributed microphones. With distributed arrays, the transfer functions to the different microphones vary and these variations have to be taking into account providing an optimal signal combining. Often when the room acoustic is taken into account in a beamformer design, one microphone is taken as a reference channel, for example, the speech distortion weighted-multichannel Wiener filter (SDW-MWF) [7, 8] or the general transfer function GSC (TF-GSC) [9, 10]. For microphone arrays with close proximities and similar transfer functions, this, is a suitable solution. However, for distributed microphones, the a priori chosen reference channel is not necessarily the ideal choice. Moreover, possible equalization capabilities are often neglected.

The matched filter (MF) [11] and the special case of the MF, the minimum variance distortionless response (MVDR) beamformer, provide a signal combining that maximizes the signal-to-noise ratio (SNR) in the presence of additive noise. A direct implementation of matched filtering requires knowledge of the acoustic transfer functions. With perfect channel knowledge, the MVDR beamformer also provides perfect equalization. However, with speech applications, the acoustic transfer functions are unknown and we have no means to directly measure the room impulse responses. There exist several blind approaches to estimate the acoustic transfer functions (see, e.g., [12–14]) which were successfully applied to dereverberation. However, the proposed estimation methods are computationally demanding. In [15], an iterative procedure was proposed where the matched filter was utilized in combination with a least-mean-squares (LMSs) adaptive algorithm for blind identification of the required room impulse responses.

In general, a signal combining for distributed microphones is desirable which does not require explicit knowledge of the channel characteristics. In a previous work, we have developed a matched-filter approach under the assumption of a uniform incoherent noise field [16]. The optimal weighting of the matched filter can be estimated by an approximation of the input SNR values and a phase estimate. Similarly, a scaled version of the MVDR-beamformer coefficients can be found by maximizing the beamformer output SNR [17]. In the frequency domain, these coefficients can be obtained by estimating the dominating generalized eigenvector (GEV) of the noise covariance matrix and the input signal covariance matrix. For instance, an adaptive variant for estimating a GEV was proposed by Doclo and Moonen [18], and later by Warsitz and Haeb-Umbach [19]. Furthermore, it can be shown that the SDW-MWF also provides an optimal signal combining that maximizes the signal-to-noise ratio [20]. The SDW-MWF requires only estimates of the input and noise correlation matrices. Hence, no explicit channel knowledge is required. However, the SDW-MWF does not equalize the speech signal.

In this work, we consider speech enhancement with distributed microphones. In Section 3, we present some measurement results that motivate a distributed microphone array. In particular, we consider two different acoustic situations: a conference room where the noise level is typically low, but the speech signal is distorted due to reverberation, and a car environment where the reverberation time is low, but the strong background noise occurs.

The basic idea of the presented approach is to apply the well-known matched-filter technique for a blind equalization of the acoustic system in the presence of additive background noise. This concept is strongly related to the SDW-MWF. Therefore, we discuss different matched-filtering techniques and their relation to the multichannel Wiener filter in Section 4.

In many speech applications, a diffuse noise field can be assumed [21]. With a diffuse noise field, the correlation of the noise signals depends on the frequency and the distance between the microphones. Typically, for small microphone distances, the low-frequency band is highly correlated whereas the correlation is low for higher frequencies. With a larger microphone spacing, the noise correlation is further decreased and the noise components can be assumed to be spatially uncorrelated. In Sections 5 and 6, we demonstrate that this fact can be exploited to reduce the complexity of the SDW-MWF algorithm as well as to improve the equalization capabilities.

The calculation of the MWF requires the inversion of the correlation matrix of the input signals. This is a computationally demanding and also numerically sensitive task. In Section 5, we show that for a scenario with a single speech source and with spatially uncorrelated noise the matrix inversion can be omitted. Using the matrix inversion lemma [22] the equation of the MWF filter weights can be rewritten to an equation that only depends on the correlations of the input signals and the input noise power spectral densities at the different microphones.

In Section 7, we present a blind-matched filtering approach for speech recording in spatially uncorrelated noise where no assumption on the geometrical adjustment of the microphones is made. The approach presented in [16] is limited to uncorrelated noise signals where the noise power spectral densities are equal for all microphone inputs. In this work, we extend these results to situations where the noise signals are spatially uncorrelated, but the noise power spectral densities can vary. Furthermore, we show that combined with a single channel Wiener filter, this new structure is equivalent with the SDW-MWF with respect to noise suppression. However, the new approach provides a partial equalization of the acoustic transfer functions between the local speaker and the microphone positions.

Finally, we demonstrate in Section 8 that the presented filter structure can be utilized for blind system identification. For equal noise power spectral densities at all microphone inputs, the matched filter is equal to the vector of transfer function coefficients up to a common factor. Hence, by estimating the ideal matched filter, we estimate the linear acoustic system up to a common filter. Note that many known approaches for blind system identification can only infer the different channels up to a common filter [15]. Similarly, with the proposed system, all filters are biased. We derive the transfer function of the common filter and demonstrate that the biased acoustic transfer functions can be reliably estimated even in the presence of strong background noise.

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 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 ith microphone signal. The subsampled time index and the frequency bin index are denoted by κ and ν, respectively. In the following, the dependencies on κ and ν are often omitted for lucidity. Hence, we can define the M-dimensional vectors $\mathbf{S}$, $\mathbf{N}$, and $\mathbf{Y}$, in which the signals are stacked as follows:

$$\begin{gathered}\mathbf{S} = \begin{bmatrix} S_1 & S_2 & \cdots & S_M \end{bmatrix}^T, \\ \mathbf{N} = \begin{bmatrix} N_1 & N_2 & \cdots & N_M \end{bmatrix}^T, \\ \mathbf{Y} = \mathbf{S} + \mathbf{N}. \end{gathered} \tag{2}$$

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:

$$\mathbf{H} = \begin{bmatrix} H_1 & H_2 & \cdots & H_M \end{bmatrix}^T. \tag{3}$$

In the following, we assume that the noise signals are zero-mean random processes with the variances $\sigma^2_{N_1}, \ldots, \sigma^2_{N_M}$. We denote the signal-to-noise ratio (SNR) at the microphone i by

$$\gamma_i = \frac{\sigma_X^2 |H_i|^2}{\sigma^2_{N_i}}, \tag{4}$$

where σ_X^2 is the speech power at the speaker's mouth.

3. Measurement and Simulation Setup

Throughout the paper, we will illustrate the proposed method with measurement and simulation results for two different acoustic situations: a conference room where the noise level is typically low, but the speech signal is distorted due to reverberation, and a car environment where the reverberation time is low, but the strong background noise may lead to very low input SNR values. In this section, we first present some measurement results that motivate a distributed microphone array. Then, we describe the setup for the simulations.

The basic idea of the presented approach is to apply the well-known matched filter technique for a blind equalization of the acoustic system in the presence of additive background noise. We first discuss some measurement results obtained in a conference room with a size of 4.7 × 4.8 × 3.0 m. For these measurements, we used three omnidirectional microphones which are placed on a table in the conference room as shown in Figure 1. The microphone distance was chosen 1.2 m for mic. 1 and mic. 3, and 1 m between the other microphone pairs (see Figure 1). This results in distances in the range from 0.5 m to 1.3 m between the local speakers and the microphones.

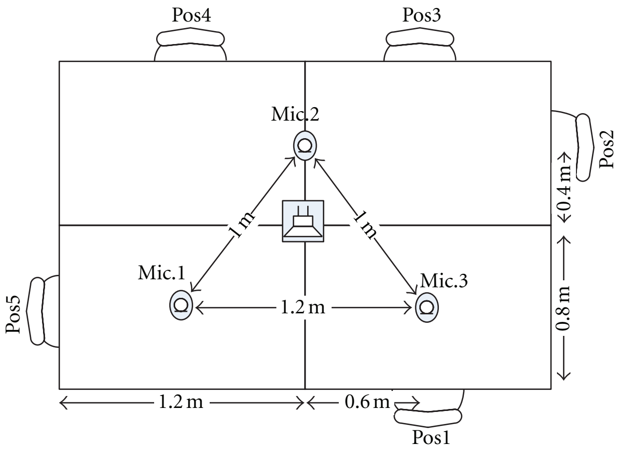


Figure 1: Measurement setup for the conference scenario.

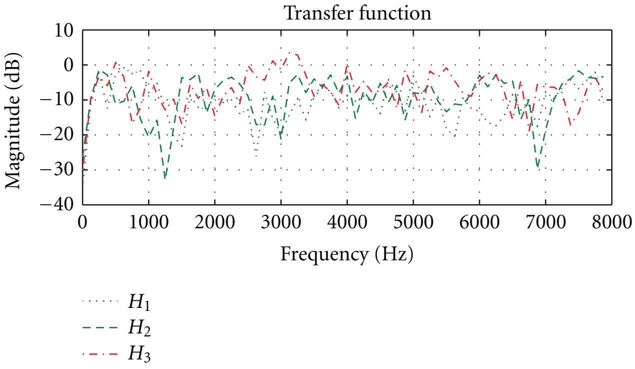


Figure 2: Magnitudes of the transfer functions for speaker three.

With an artificial head, we measured the room impulse responses for five local teleconference participants. For this scenario, Figure 2 shows the magnitudes of the acoustic transfer functions. The influence of the room acoustic is clearly visible. For some frequencies, the magnitudes of the acoustic transfer functions show differences of more than 20 dB. It can also be stated that the microphone with the best transfer function is not obvious, because for some frequencies $H_1(\nu)$, $H_2(\nu)$, and for others $H_3(\nu)$ has less attenuation.

Figure 3 depicts the SNR versus frequency for a situation with background noise which arises from a fan of a video projector. From this figure, we observe that the SNR values for frequencies above 1.5 kHz are quite distinct for these three microphone positions with differences of up to 10 dB depending on the particular frequency. Again, the best microphone position is not obvious in this case, because the SNR curves across several times.

Theoretically, if we assume spatially uncorrelated noise signals, a matched filter combining these input signals would

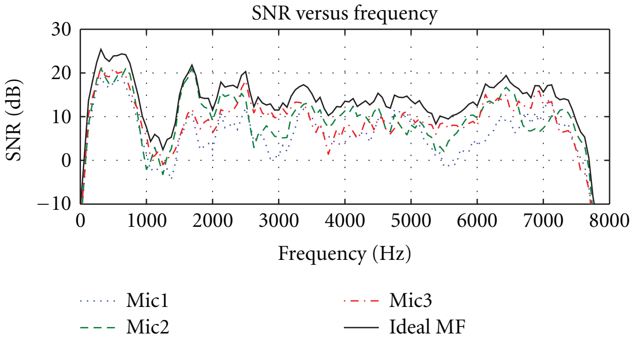


Figure 3: Input SNR for a conference scenario with background noise.

result in an output SNR equal to the sum of the input SNR values. With three inputs, a matched filter array achieves a maximum gain of 4.8 dB for equal input SNR values. In the case of varying input SNR values, the sum is dominated by the maximum value. Hence, for the curves in Figure 3, the output SNR would essentially be the envelope of the three curves. This is also shown in Figure 3, where the output SNR of an optimal signal combining is plotted (solid line).

For the simulation results presented throughout the paper the following setup was used: all processed signals are sampled at a sampling rate of $f_s = 16000$ Hz. For the short-time Fourier transform (STFT), we used a length of $L = 512$ and an overlap of $K = 384$ samples, while an overlap-add processing for the signal synthesis was performed. As clean speech signals we used two male and two female speech samples, each of a length of 8 seconds. Therefore, we took the German-speaking test samples from the recommendation P.501 of the International Telecommunication Union (ITU) [23]. To generate the speech signals s_i at the microphones, the clean speech was convolved with the corresponding room impulse responses. The reverberation time for the conference scenario was $T_{60} = 0.25$ s. We also show results for a conference scenario with $T_{60} = 0.5$ s, but for this scenario the impulse responses are generated using the image method [24]. Most presented algorithms require estimates of the noise power spectral density (PSD) and a voice activity detection (VAD), here we used the methods described in [16] throughout the paper.

For the measurements in the car environment, one microphone was installed close to the inside mirror, while the second microphone was mounted at the A-pillar (the A-pillar of a vehicle is the first pillar of the passenger compartment, usually surrounding the windscreen). This microphone setup leads to a distance of 0.6 m between the two microphones. We consider three different background noise situations for noise recordings: driving noise at 100 km/h and 140 km/h, and the noise arising from an electric fan (defroster, car in standstill). For a comparison with typical beamformer constellations, we installed a second pair of microphones at the inside mirror, such that the microphone distance between these two microphone was 0.04 m. This microphone setup is evaluated later in Section 5. Figure 4 shows the measured setup for the in-car environment. For all measurements in the car, we used cardioid microphones, which are often used for automotive applications. For the measurements of the room impulse responses also the artificial head was used. The reverberation time for this scenario was $T_{60} = 0.05$ s.

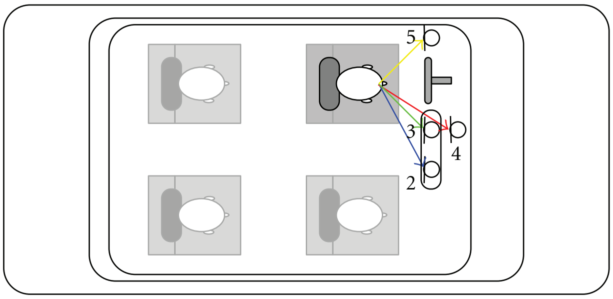


Figure 4: Illustration of the measured in-car scenario.

4. Optimal Signal Combining

In this section, we discuss combining schemes which are optimal in a certain manner. For such a combining of the microphone signals, each input Y_i is processed by a linear filter G_i before the signals are added. Stacking these filter functions into the vector $\mathbf{G}$, we have

$$\mathbf{G} = \begin{bmatrix} G_1 & G_2 & \cdots & G_M \end{bmatrix}^T. \tag{5}$$

Therefore, the processed signal at the output of the combining system can be expressed as follows:

$$\begin{aligned} \hat{X} &= \mathbf{G}^\dagger \mathbf{Y}, \\ \hat{X} &= \mathbf{G}^\dagger \mathbf{H} X + \mathbf{G}^\dagger \mathbf{N}. \end{aligned} \tag{6}$$

The SNR at the system output is defined as the ratio:

$$\begin{aligned} \gamma &= \frac{\mathbb{E}\left\{ \left| \mathbf{G}^\dagger \mathbf{H} X \right|^2 \right\}}{\mathbb{E}\left\{ \left| \mathbf{G}^\dagger \mathbf{N} \right|^2 \right\}}, \\ &= \frac{\sigma_X^2 \mathbf{G}^\dagger \mathbf{H} \mathbf{H}^\dagger \mathbf{G}}{\mathbf{G}^\dagger \mathbf{R}_N \mathbf{G}}, \end{aligned} \tag{7}$$

where

$$\mathbf{R}_N = \mathbb{E}\left\{ \mathbf{N} \mathbf{N}^\dagger \right\} \tag{8}$$

is the correlation matrix of the noise signals.

4.1. Maximization of the SNR. Our aim is now to find the filter functions $\mathbf{G}$, which are optimal in the sense of a maximal output SNR of the combining system. Hence, the maximization problem can be stated as

$$\mathbf{G}^{\mathrm{MF}} = \arg\max_{\mathbf{G}} \left\{ \frac{\sigma_X^2 \mathbf{G}^\dagger \mathbf{H} \mathbf{H}^\dagger \mathbf{G}}{\mathbf{G}^\dagger \mathbf{R}_N \mathbf{G}} \right\}. \tag{9}$$

This maximization problem leads to an eigenvalue problem with the matched filter (MF) solution [25]:

$$\mathbf{G}^{\mathrm{MF}} = c \cdot \mathbf{R}_N^{-1} \mathbf{H}, \tag{10}$$

where c is a nonzero constant value. Hence, one has to weight the input signals according to the acoustic transfer functions and the inverse of the noise correlation matrix. This weighting is also known as maximum SNR (MSNR) beamformer.

Applying a constant factor to $\mathbf{R}_N^{-1}\mathbf{H}$ does not affect the SNR at the filter output. Therefore, the matched filter can also be utilized for equalization to get a flat frequency response according to the source speech position ($\mathbf{G}^{\mathrm{MF}\dagger}\mathbf{H} = 1$). In this case, the following form is used:

$$\mathbf{G}^{\mathrm{MVDR}} = \frac{\mathbf{R}_N^{-1}\mathbf{H}}{\mathbf{H}^\dagger\mathbf{R}_N^{-1}\mathbf{H}}. \tag{11}$$

This algorithm is also called the minimum variance distortionless response (MVDR) beamformer and was described by Cox et al. [26]. For this technique, knowledge about the room impulse responses and the noise correlation matrix is needed. For an estimation of the noise power density and the cross-power density, several approaches exist in the literature [21, 27–30]. But the estimation of the room impulse responses is a blind estimation problem [12–14]. A reliable estimation of the room impulse responses in realtime is still an open issue. Often only a linear-phase compensation is done, by applying a sufficient time delay for the signals, thus the transfer functions are replaced by the steering vector:

$$\mathbf{d}(\nu) = \left[1, e^{-j\Delta_2(\nu)}, \dots, e^{-j\Delta_M(\nu)}\right]^T, \tag{12}$$

where $\Delta_i(\nu)$ denotes the phase difference between the first and ith microphone. This corresponds to the classical Frost beamformer [2]. Thus, an estimate of the time delay of arrival (TDOA) is required. Note that for a known array geometry this information is equivalent to the direction of arrival (DOA).

4.2. Multichannel MMSE Criterion. Another related criterion is the minimization of the mean square error (MSE) between the output signal and a reference signal. This can also be used to find an optimal combining strategy for the microphone signals. To calculate the minimum mean squared error (MMSE) estimate of the clean speech signal X at the speaker's mouth, one has to minimize the following cost function:

$$\mathbf{G}^{\mathrm{MWF}} = \arg\min_{\mathbf{G}} \mathbb{E}\left\{\left|\mathbf{G}^\dagger\mathbf{Y} - X\right|^2\right\}. \tag{13}$$

By setting the complex derivative with respect to $\mathbf{G}^*$ to zero, one obtains the solution of this minimization problem as

$$\mathbf{G}^{\mathrm{MWF}} = \mathbf{R}_Y{}^{-1}\mathbf{R}_{Yx}, \tag{14}$$

where $\mathbf{R}_{Yx} = \mathbb{E}\{\mathbf{Y}X^*\}$ is the cross-correlation vector between the clean speech and the microphone input signal and $\mathbf{R}_Y = \mathbb{E}\{\mathbf{Y}\mathbf{Y}^\dagger\}$ is the correlation matrix of the microphone input signal, respectively. In the literature, this is often referred as the multichannel Wiener filter (MWF), which can be used for signal combining and noise reduction.

To overcome the problem of the required but unavailable cross-correlation vector $\mathbf{R}_{Yx}$ in the definition of the MWF, cf. (14), one can define an MWF that minimizes the mean squared error with respect to the speech signal of a reference microphone signal S_{ref}. In [7], the SDW-MWF was proposed, while a tradeoff parameter was introduced to the MWF. With this parameter, it is possible to adjust the noise reduction capabilities of the MWF with respect to the speech signal distortion at the output. Thus, the signal distortion is taken into account in the optimization. Therefore, the distortion is measured as the distance between the speech component of the output signal and the speech component of an input channel. This reference channel is selected arbitrarily in advance.

The error signal ε for the minimization is then defined as the difference between the output signal $\mathbf{G}^\dagger\mathbf{Y}$ and the speech component of the signal Y_{ref}:

$$\begin{aligned} \varepsilon &= \mathbf{G}^\dagger\mathbf{Y} - \mathbf{u}^T\mathbf{S} \\ &= \left(\mathbf{G}^\dagger - \mathbf{u}^T\right)\mathbf{S} + \mathbf{G}^\dagger\mathbf{N} \\ &= \varepsilon_s + \varepsilon_n. \end{aligned} \tag{15}$$

The column vector $\mathbf{u}$ selects the reference channel, that is, the corresponding entry is set to one and the others are set to zero. Using the two MSE cost functions:

$$\begin{aligned} J_n(\mathbf{G}) &= \mathbb{E}\left\{|\varepsilon_n|^2\right\}, \\ J_s(\mathbf{G}) &= \mathbb{E}\left\{|\varepsilon_s|^2\right\}, \end{aligned} \tag{16}$$

the unconstraint minimization criterion for the SDW-MWF is defined by

$$\mathbf{G}^{\mathrm{SDW}} = \arg\min_{\mathbf{G}} J_n(\mathbf{G}) + \frac{1}{\mu_W}J_s(\mathbf{G}), \tag{17}$$

where $1/\mu_W$ is a Lagrange multiplier. This results in the solution:

$$\mathbf{G}^{\mathrm{SDW}} = (\mathbf{R}_S + \mu_W\mathbf{R}_N)^{-1}\mathbf{R}_S\mathbf{u}, \tag{18}$$

where $\mathbf{R}_S$ is the speech correlation matrix and μ_W is a parameter which allows a trade-off between speech distortion and noise reduction (for details cf. [7]).

For further analyses, we assume that the single speaker speech signal is a zero-mean random proces with the PSD σ_X^2 and a time-invariant acoustic system. The correlation matrix of the speech signal can be written as

$$\begin{aligned} \mathbf{R}_S &= \mathbb{E}\left\{\mathbf{S}\mathbf{S}^\dagger\right\} \\ &= \sigma_X^2\mathbf{H}\mathbf{H}^\dagger. \end{aligned} \tag{19}$$

Using the matrix inversion lemma [22], the SDW-MWF can be decomposed as

$$\begin{aligned} \mathbf{G}^{\mathrm{SDW}} &= \frac{\sigma_X^2}{\sigma_X^2 + \mu_W\left(\mathbf{H}^\dagger\mathbf{R}_N^{-1}\mathbf{H}\right)^{-1}} \frac{\mathbf{R}_N^{-1}\mathbf{H}}{\mathbf{H}^\dagger\mathbf{R}_N^{-1}\mathbf{H}}\mathbf{H}^\dagger\mathbf{u}, \\ \mathbf{G}^{\mathrm{SDW}} &= G^{\mathrm{WF}}\mathbf{G}^{\mathrm{MVDR}}H_{\mathrm{ref}}^*, \end{aligned} \tag{20}$$

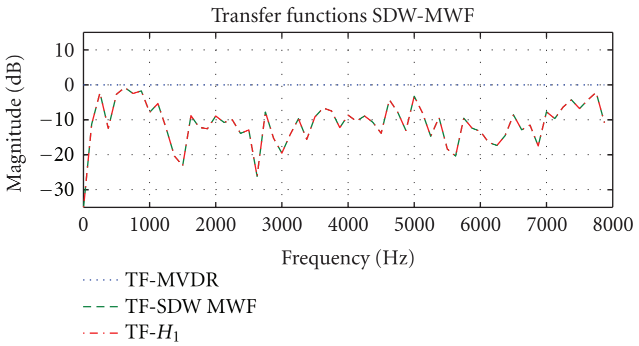


Figure 5: Comparison of the overall transfer function of the SDW-MWF and the ideal MVDR beamformer.

where $(\mathbf{H}^\dagger \mathbf{R}_N^{-1} \mathbf{H})^{-1}$ is the noise variance at the output of $\mathbf{G}^{\mathrm{MVDR}}$:

$$\left(\mathbf{H}^\dagger \mathbf{R}_N^{-1} \mathbf{H}\right)^{-1} = \mathbf{G}^{\mathrm{MVDR}\dagger} \mathbf{R}_N \mathbf{G}^{\mathrm{MVDR}}, \\ = \sigma^2{}_{N_{\mathrm{MVDR}}}. \tag{21}$$

Appendix A provides a derivation of this decomposition.

Thus the SDW-MWF is decomposed in an MVDR beamformer and a filter that is equal to the acoustic transfer function of the reference channel ($H^*_{\mathrm{ref}} = \mathbf{H}^\dagger \mathbf{u}$). Furthermore, the noise reduction is achieved by a single channel Wiener filter

$$G^{\mathrm{WF}} = \frac{\sigma_X^2}{\sigma_X^2 + \mu_W \sigma^2{}_{N_{\mathrm{MVDR}}}}, \tag{22}$$

where μ_W can be interpreted as a noise overestimation factor [31].

From this decomposition, it can be seen that the SDW-MWF provides an optimal signal combining with respect to the output SNR. Yet, it is not able to equalize the acoustic transfer functions. This is also obvious in Figure 5, where the overall system transfer function is depicted (dashed line). Here, the first microphone with the transfer function $H_1(\nu)$ was used as reference channel. For this plot, the Wiener filter part $G^{\mathrm{WF}}(\nu)$ of the transfer function was neglected. We observe that the overall transfer function of the SDW-MWF (dashed line) is equivalent to the transfer function of the reference channel (semidashed line). Note that we measured the transfer function between the speaker's mouth and the output of the SDW-MWF. Also, the flat transfer function of the MVDR beamformer is plotted for a comparison (dotted line).

5. The SDW-MWF for Spatially Uncorrelated Noise

The calculation of the MWF in (18) requires the inversion of the correlation matrix of the input signals. This is a computationally demanding and also numerically sensitive task. In this section, we show that for a scenario with a single speech source and with spatially uncorrelated noise, the matrix inversion can be omitted. Using the matrix inversion lemma [22], the equation of the MWF filter weights can be rewritten to an equation that only depends on the correlations of the input signals and the input noise PSDs at the different microphones.

Consider the speech-distortion-weighted multichannel Wiener filter according to (18). We can rewrite the inverse $(\mathbf{R}_S + \mu_W \mathbf{R}_N)^{-1}$ using the result in (A.1):

$$(\mu_W \mathbf{R}_N + \mathbf{R}_S)^{-1} = \frac{1}{\mu_W} \mathbf{R}_N^{-1} - \frac{\mathbf{R}_N^{-1} \mathbf{R}_S \mathbf{R}_N^{-1}}{\mu_W^2 (1 + \mu_W^{-1} \sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1} \mathbf{H})}. \tag{23}$$

Furthermore, using (21), we have

$$\sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1} \mathbf{H} = \frac{\sigma_X^2}{\sigma^2{}_{N_{\mathrm{MVDR}}}} \\ = \gamma, \tag{24}$$

which is the signal-to-noise ratio at the output of the MVDR beamformer.

Using the inverse of $(\mu_W \mathbf{R}_N + \mathbf{R}_S)$ from (23) and the definition of the SDW-MWF in (18), we have

$$\begin{aligned} \mathbf{G}^{\mathrm{SDW}} &= \left(\frac{1}{\mu_W} \mathbf{R}_N^{-1} - \frac{\mathbf{R}_N^{-1} \mathbf{R}_S \mathbf{R}_N^{-1}}{\mu_W^2 (1 + \mu_W^{-1} \sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1} \mathbf{H})} \right) \mathbf{R}_S \mathbf{u} \\ &= \left(1 - \frac{\gamma}{\mu_W + \gamma} \right) \frac{1}{\mu_W} \mathbf{R}_N^{-1} \mathbf{R}_S \mathbf{u} \\ &= \frac{\mathbf{R}_N^{-1} \mathbf{R}_S \mathbf{u}}{\mu_W + \gamma}. \end{aligned} \tag{25}$$

Because speech and noise are independent, we can estimate $\mathbf{R}_S$ by $\mathbf{R}_S = \mathbf{R}_Y - \mathbf{R}_N$. Therefore, we obtain

$$\begin{aligned} \mathbf{G}^{\mathrm{SDW}} &= \frac{\mathbf{R}_N^{-1} (\mathbf{R}_Y - \mathbf{R}_N) \mathbf{u}}{\mu_W + \gamma} \\ &= \frac{(\mathbf{R}_N^{-1} \mathbf{R}_Y - \mathbf{I}) \mathbf{u}}{\mu_W + \gamma} \\ &= \frac{\mathbf{R}_N^{-1} \mathbf{R}_Y \mathbf{u} - \mathbf{u}}{\mu_W + \gamma}. \end{aligned} \tag{26}$$

Note that the column vector $\mathbf{u}$ selects the reference channel, that is, the corresponding entry is set to one and the others are set to zero.

Because we assume that the noise signals at the different microphones are uncorrelated, $\mathbf{R}_N$ is a diagonal matrix, and the elements of the main diagonal are the noise variances $\sigma_{N_1}^2, \ldots, \sigma_{N_M}^2$. Therefore, we obtain the inverse

$$\mathbf{R}_N^{-1} = \begin{pmatrix} \frac{1}{\sigma_{N_1}^2} & 0 & \cdots & 0 \\ & \ddots & & \\ 0 & 0 & \cdots & \frac{1}{\sigma_{N_M}^2} \end{pmatrix}. \tag{27}$$

Let ref be the index of the one in the vector $\mathbf{u}$. $\mathbf{R}_N^{-1}\mathbf{R}_Y\mathbf{u}$ results in the column vector:

$$\mathbf{w}_{\text{ref}} = \mathbf{R}_N^{-1}\mathbf{R}_Y\mathbf{u} = \left[\frac{\mathbb{E}\{Y_1 Y_{\text{ref}}^*\}}{\sigma_{N_1}^2}, \ldots, \frac{\mathbb{E}\{Y_M Y_{\text{ref}}^*\}}{\sigma_{N_M}^2}\right]^T. \tag{28}$$

Therefore, we obtain the following expression for the SDW-MWF for spatially uncorrelated noise signals:

$$\mathbf{G}^{\text{SDW}} = \frac{\mathbf{w}_{\text{ref}} - \mathbf{u}}{\mu_W + \gamma}. \tag{29}$$

This representation of the speech distortion weighted multichannel Wiener filter omits the inversion of the matrix $(\mathbf{R}_S + \mu_W \mathbf{R}_N)$. For spatially uncorrelated noise signals, the SNR γ can be calculated as the sum of the input signal-to-noise ratios:

$$\gamma = \sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H} = \sum_{i=1}^{M} \frac{\sigma_X^2 |H_i|^2}{\sigma_{N_i}^2}. \tag{30}$$

6. The SDW-MWF for Diffuse Noise

In the literature, many investigations on the spatial correlation properties of noise fields have been made. The assumption of spatially uncorrelated noise is rarely fulfilled in real-world scenarios, but it has been found, for example, by Martin and Vary in [21], that many noise fields can be assumed to be diffuse, like the noise in a car environment [32], office noise [21], or, for example, babble noise [33]. For diffuse noise, the spatial correlation depends on the intermicrophone distance and is dominant especially in the lower-frequency bands. Typically, the low-frequency band is highly correlated whereas the correlation is low for higher frequencies. This fact can be exploited by omitting the matrix inversion for higher frequencies.

To evaluate the correlation between the noise signals at different positions, the coherence function of the noise signals from different intermicrophone distances can be computed. The magnitude squared coherence (MSC) between two signals n_i and n_j is defined as follows:

$$C_{N_iN_j}(\nu) = \frac{\left|\sigma_{N_iN_j}(\nu)\right|^2}{\sigma_{N_i}^2(\nu)\sigma_{N_j}^2(\nu)}, \tag{31}$$

where $\sigma_{N_iN_j}(\nu)$ and $\sigma_{N_i}^2$ are the cross-power spectral density (CPSD) and the power spectral densities (PSDs) of the signals n_i and n_j, respectively. The values of the coherence function are between 0 and 1, where 0 means no correlation between the two signals at that frequency point. For highly correlated signals, the MSC will become close to 1 for all frequencies.

In [34], Armbrüster et al. have shown that the coherence of an ideal diffuse sound field recorded with omnidirectional microphones can be computed as follows:

$$C_{\text{theo}}(\nu) = \frac{\sin^2(2\pi\nu d_{\text{mic}} f_s/(Lc))}{(2\pi\nu d_{\text{mic}}/c)^2}, \tag{32}$$

where L denotes the length of the short-time Fourier transform (STFT) and f_s is the sampling frequency. The speed of sound is denoted by c, and d_{mic} represents the microphone distance. The zeros of the theoretical coherence function in (32) can by calculated by the following expression:

$$\nu_{\text{zero},m} = \frac{mLc}{2d_{\text{mic}} f_s}, \quad m = 1, 2, 3, \ldots. \tag{33}$$

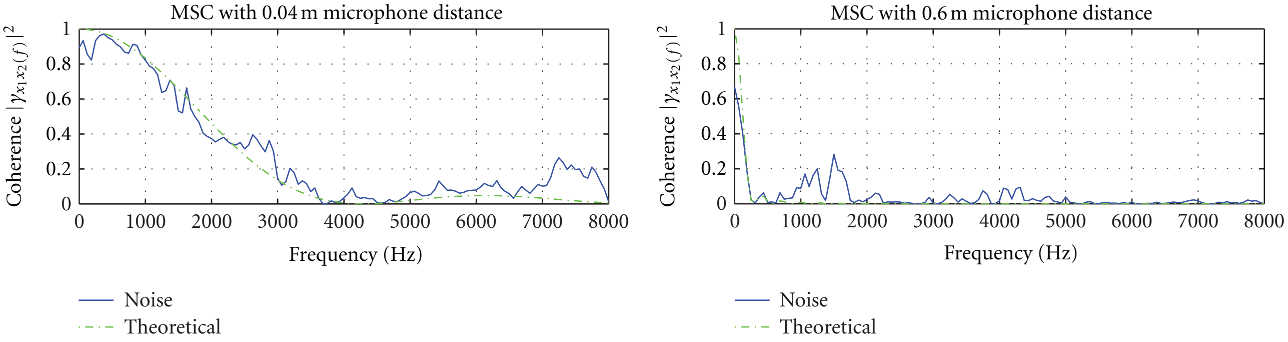
In the following, we consider the coherence for the noise signals of the in-car scenario for the driving situation at 100 km/h (see Section 3). Figure 6(a) shows the coherence functions of the noise for the microphone pair with an intermicrophone spacing of $d_{\text{mic}} = 0.04$ m. Also, the theoretical coherence function computed according to (32) is shown. Obviously, there is a high correlation of the noise signals at frequencies below 2 kHz. Note that the coherence of the noise signals is closely approximated by the theoretical coherence function C_{theo}, although cardioid microphones were used for this measurement. In Figure 6(b), the coherence function of the noise for the microphone pair with a 0.6 m spacing is depicted. In this constellation, the noise signals at the two microphones are highly correlated for frequencies below 150 Hz only.

From (33) and Figure 6, it is obvious that the correlation of the diffuse noise signals depends on the intermicrophone distance. Therefore, the noise has only a high correlation at low frequencies and especially the high frequencies are only weakly correlated. Thus, the assumption of spatially uncorrelated noise is fulfilled for the higher frequency bands. Therefore, we propose to calculate the filter weights depending on the theoretical coherence $C_{\text{theo}}(\nu)$; for frequencies with a high coherence, we calculate the filter weights using the matrix inversion (see (18)), while for frequencies with a low coherence, we assume uncorrelated noise and thus the weights are computed according to (29). Hence, the filter function is calculated according to

$$\mathbf{G}^{\text{SDW}}(\nu) = \begin{cases} (\mathbf{R}_S(\nu) + \mu_W \mathbf{R}_N(\nu))^{-1}\mathbf{R}_S(\nu)\mathbf{u}, & C_{\text{theo}}(\nu) \geq C_{\text{lim}}, \\ \dfrac{\mathbf{w}_{\text{ref}}(\nu) - \mathbf{u}}{\mu_W + \gamma(\nu)}, & \text{otherwise}, \end{cases} \tag{34}$$

where C_{lim} is a parameter that allows a trade-off between accuracy and computing time.

The simulation results for the in-car microphone scenario with the two different microphone setups are given in Table 1. Each scenario (microphone setup and noise condition) was simulated twice. The first time it was simulated using the fullband matrix inversion according to (18) (denoted by *fullband MWF*). These results can be seen as an upper bound for the performance evaluation of the proposed method. The second time we used the proposed approach of the SDW-MWF with the partially inversion of the correlation matrix (*partial MWF*). Therefore, the inversion was omitted for all frequency bins with a theoretical coherence C_{theo} less than 0.7. This leads in our simulation setup to the threshold frequencies $f_{\text{lim}} = 1500$ Hz for the closed spaced microphone



(a) MSC for intermicrophone spacing of 0.04 m

(b) MSC for intermicrophone spacing of 0.6 m

Figure 6: Comparison of the magnitude squared coherence.

Table 1: Comparison of the fullband (according to (18)) and partial MWF (according to (34)).

	d_{mic} [m]	0.04			0.6		
		100 km/h	140 km/h	defrost	100 km/h	140 km/h	defrost
	SSNR ch.1 [dB]	−2.7	−7.1	−0.8	−2.7	−7.7	−0.2
	SSNR ch.2 [dB]	−3.8	−8.2	−1.3	−1.1	−5.8	0.6
Fullband MWF	SSNR [dB]	4.4	−0.8	5.3	5.4	−0.1	7.6
	S-MOS	4.6	4.7	4.7	4.7	4.7	4.8
	N-MOS	3.2	2.6	3.0	3.1	2.6	3.0
	G-MOS	3.8	3.5	3.8	3.8	3.5	3.8
Partial MWF	SSNR [dB]	4.6	−0.6	4.9	5.4	−0.1	7.3
	S-MOS	4.6	4.7	4.7	4.6	4.7	4.7
	N-MOS	3.2	2.6	3.0	3.1	2.6	3.0
	G-MOS	3.8	3.5	3.8	3.7	3.5	3.7

pair and $f_{\text{lim}} = 100$ Hz for the setup with the microphone spacing of 0.6 m. As an objective evaluation criterion, we calculated the segmental signal-to-noise ratio (SSNR) of the output signal. Therefore, a voice activity detection according to the ITU P.56 was used [35]. Furthermore, we show results from an instrumental quality analysis in Table 1. The speech quality and noise reduction were evaluated according to the ETSI standard EG 202 396-3 [36]. This algorithm calculates three objective quality measures (according to the mean opinion score (MOS) scale): Speech-MOS (S-MOS), Noise-MOS (N-MOS), and Global-MOS (G-MOS). From these results, we observe that the *partial MWF* algorithm obtains nearly the same performance as the fullband SDW-MWF.

7. Matched Filtering for Spatially Uncorrelated Noise

We have seen in Section 4.2 that the SDW-MWF provides an optimal signal combining with respect to the output SNR, where the SDW-MWF does not require explicit channel knowledge to obtain this result. For spatially uncorrelated noise the SDW-MWF according to (29) requires only estimates of the input SNR values and the input cross-correlation with respect to the reference channel. However, in contrast to the MVDR beamformer, the SDW-MWF does not equalize the acoustic system.

In the following, we show that knowledge of the input SNR values and the input cross-correlation with respect to the reference channel is sufficient to provide at least a partial channel equalization. We consider the matched filter for spatially uncorrelated noise signals. If we assume that the noise signals at the different microphones are uncorrelated, $\mathbf{R}_N$ is a diagonal matrix, and the elements of the main diagonal are the noise variances $\sigma_{N_1}^2, \ldots, \sigma_{N_M}^2$. Therefore, we obtain the inverse $\mathbf{R}_N^{-1}$ as in (27). In this case, the filter coefficients of the matched filter can be determined independently and we obtain

$$G_i^{\text{MF}} = \frac{H_i}{\sigma_{N_i}^2} \tag{35}$$

as ith coefficient of the matched filter according to (10) and

$$G_i^{\text{MVDR}} = \frac{H_i}{\sigma_{N_i}^2 \left(|H_1|^2/\sigma_{N_1}^2 + |H_2|^2/\sigma_{N_2}^2 + \cdots \right)} \tag{36}$$

according to (11).

7.1. Filter Design. In [16], we have demonstrated that under the assumption of a uniform and spatially uncorrelated noise field, this optimal MF weighting can be obtained by the following filter:

$$G_i = \sqrt{\frac{\gamma_i}{\gamma}} = \sqrt{\frac{\gamma_i}{\gamma_1 + \gamma_2 + \cdots + \gamma_M}} \qquad (37)$$

and an additional phase synchronization. γ denotes the sum of all input SNR values. Hence, this filter requires only estimates of the input SNRs. In the following, we extend this concept to nonuniform noise fields. In this case, the optimal weighting depends also on the noise power densities $\sigma_{N_i}^2$. Consider now the following filter:

$$G_i = \sqrt{\frac{\gamma_i \widetilde{\sigma}_N^2}{\sigma_{N_i}^2 \gamma}} = \sqrt{\frac{\gamma_i \widetilde{\sigma}_N^2}{\sigma_{N_i}^2 (\gamma_1 + \gamma_2 + \cdots + \gamma_M)}}, \qquad (38)$$

where $\widetilde{\sigma}_N^2$ is the mean of the noise power spectral densities at the different microphones, defined by

$$\widetilde{\sigma}_N^2 = \frac{1}{M} \sum_{i=1}^{M} \sigma_{N_i}^2. \qquad (39)$$

This filter depends on the noise power density $\sigma_{N_i}^2$ and all input SNR values. Using (4), we obtain

$$\begin{aligned} G_i &= \sqrt{\frac{|H_i|^2 \widetilde{\sigma}_N^2}{\left(\sigma_{N_i}^2\right)^2 \left(|H_1|^2/\sigma_{N_1}^2 + |H_2|^2/\sigma_{N_2}^2 + \cdots\right)}} \\ &= \frac{|H_i|}{\sigma_{N_i}^2 \sqrt{\left(1/\widetilde{\sigma}_N^2\right)\left(|H_1|^2/\sigma_{N_1}^2 + |H_2|^2/\sigma_{N_2}^2 + \cdots\right)}}. \end{aligned} \qquad (40)$$

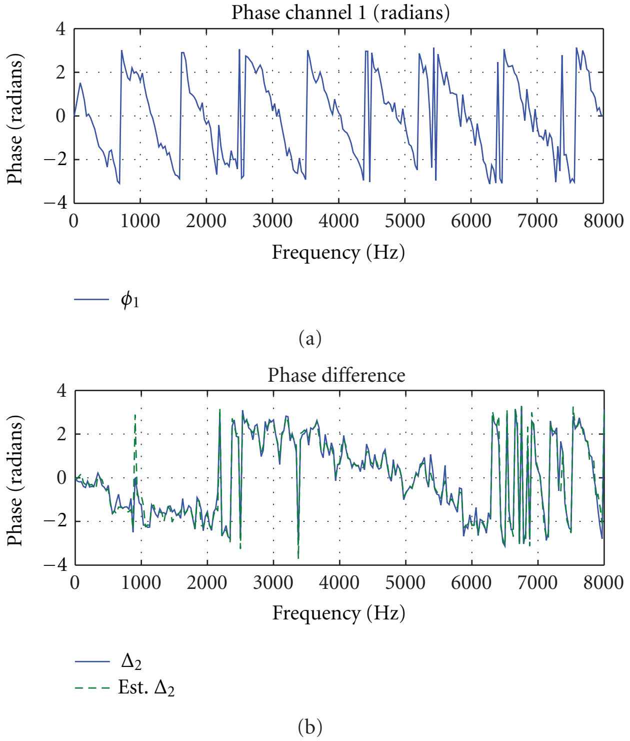
Note that the term $(1/\widetilde{\sigma}_N^2)(|H_1|^2/\sigma_{N_1}^2 + |H_2|^2/\sigma_{N_2}^2 + \cdots)$ is common to all filter coefficients. Hence, the filter according to (38) is proportional to the magnitude of the matched filter according to (35).

The proposed filter in (38) is real valued. To ensure cophasal signal combining, we require some additional system components for phase estimation.

7.2. Phase Estimation. For a coherent combining of the speech signals, we have to compensate the phase difference between the speech signals at each microphone. Therefore it is sufficient to estimate the phase differences to a reference microphone. Let $\phi_i(\nu)$ be the phase of the complex channel coefficient $H_i(\nu)$. We consider the phase differences to the a reference microphone $\Delta_i(\nu) = \phi_{\text{ref}}(\nu) - \phi_i(\nu)$, for all $i \neq \text{ref}$ and $\Delta_{\text{ref}}(\nu) = 0$. Cophasal addition is then achieved by

$$\hat{X} = \sum_{i=1}^{M} G_i e^{j\Delta_i} Y_i. \qquad (41)$$

For multimicrophone systems with spatially separated microphones a reliable phase estimation is a challenging task. A coarse estimate of the phase difference can also be obtained from the time-shift τ_i between the speech components in the microphone signals, for example, using the generalized correlation method [37]. However, for distributed microphone arrays in reverberant environments this phase compensation leads to a poor estimate of the actual phase differences. This can be observed in Figure 7 which depicts the phase $\phi_1(\nu)$ of the reference channel for the in car scenario with intermicrophone spacing of 0.6 m (see Section 3). In an anechoic environment, the phase of the reference channel as well as the phase difference Δ_2 for the second microphone would be linear functions of the frequency. Hence, we could expect ideal sawtooth functions if we consider the phase in the interval $[-\pi, \pi]$. From Figure 7, we observe that this is only a rough estimate of the actual phase values.



(a)

(b)

Figure 7: Actual phase of the reference channel (a) determined from the impulse response, actual phase difference and estimated phase difference for channel 2.

In order to ensure a cophasal addition of the signals, we employ a phase estimation similar to the approach presented in [16]. We use a frequency domain least-squares (FLMS) algorithm to estimate the required phase difference. Using Y_{ref} as reference signal, the filter $G_i^{\text{ANG}}(\kappa, \nu)$ is updated according to

$$G_i^{\text{ANG}}(\kappa + 1, \nu) = G_i^{\text{ANG}}(\kappa, \nu) + \mu Y_i^*(\kappa, \nu) E_i(\kappa, \nu), \qquad (42)$$

with

$$E_i(\kappa, \nu) = Y_{\text{ref}}(\kappa, \nu) - Y_i(\kappa, \nu) G_i^{\text{ANG}}(\kappa, \nu). \qquad (43)$$

Note that the filter is only adapted if voice activity is detected, where we used the VAD method described in [16]. The FLMS algorithm minimizes the expected value:

$$\mathbb{E}\left\{ \left| Y_{\text{ref}}(\kappa, \nu) - Y_i(\kappa, \nu) G_i^{\text{ANG}}(\kappa, \nu) \right|^2 \right\}. \qquad (44)$$

For stationary signals the adaptation converts to a filter transfer function:

$$G_i^{\text{ANG}} = \frac{\mathbb{E}\{Y_i^* Y_{\text{ref}}\}}{\mathbb{E}\left\{|Y_i|^2\right\}}, \qquad (45)$$

where $\mathbb{E}\{Y_i^* Y_{\text{ref}}\}$ is the cross-power spectrum of the two microphone signals and $\mathbb{E}\{|Y_i|^2\}$ is the power spectrum of the ith microphone signal. Assuming that the speech signal and the noise signals are spatially uncorrelated, (45) can be written as

$$G_i^{\text{ANG}} = \frac{\mathbb{E}\left\{|X|^2\right\} H_i^* H_{\text{ref}} + \mathbb{E}\left\{N_i^* N_j\right\}}{\mathbb{E}\left\{|X|^2\right\} |H_i|^2 + \mathbb{E}\left\{|N_i|^2\right\}}. \qquad (46)$$

For frequencies where the noise components are uncorrelated, that is, $\mathbb{E}\{N_i^* N_1\} = 0$, this formula is reduced to

$$G_i^{\text{ANG}} = \frac{\mathbb{E}\left\{|X|^2\right\} H_i^* H_{\text{ref}}}{\mathbb{E}\left\{|X|^2\right\} |H_i|^2 + \mathbb{E}\left\{|N_i|^2\right\}}. \qquad (47)$$

The phase of the filter G_i^{ANG} is determined by the two complex channel coefficients H_i and H_{ref} where the product $H_i^* H_{\text{ref}}$ has the sought phase $\Delta_i(\nu) = \phi_{\text{ref}}(\nu) - \phi_i(\nu)$. Hence, for the coherent signal combining, we use the phase of the filter G_i^{ANG}:

$$\hat{\Delta}_i(\kappa, \nu) = \arg\left(G_i^{\text{ANG}}(\kappa, \nu)\right). \qquad (48)$$

According to (28) and (29), the phase of the filter $\mathbf{G}^{\text{SDW}}$ is determined by the cross-correlation of the input signals. Comparing (28) and (45), we note that the proposed approach leads to the same phase compensation as with the SDW-MWF. Note that the output signal of the SDW-MWF is computed as $\hat{X} = \mathbf{G}^{\text{SDW}\dagger}\mathbf{Y}$.

With the estimated phase, we can now express the complex filter as

$$\widetilde{G}_i = \sqrt{\frac{\gamma_i \widetilde{\sigma}_N^2}{\sigma_{N_i}^2 \gamma}} e^{j\hat{\Delta}_i}. \qquad (49)$$

Figure 7 presents simulation results for this phase estimation, where Δ_2 denotes the actual phase difference computed from the measured impulse responses and *est*. Δ_2 is the estimated phase difference. The presented results correspond to the driving situation with a car speed of 140 km/h and an intermicrophone distance of 0.6 m, as described in Section 3.

7.3. Residual Transfer Function. Next, we derive the residual transfer function of the proposed signal combining. Using (40), the complex filter transfer function can be expressed as

$$\widetilde{G}_i = \frac{|H_i| e^{j\hat{\Delta}_i}}{\sigma_{N_i}^2 \sqrt{\left(1/\widetilde{\sigma}_N^2\right)\left(|H_1|^2/\sigma_{N_1}^2 + |H_2|^2/\sigma_{N_2}^2 + \cdots\right)}}. \qquad (50)$$

Assuming ideal knowledge of the SNR values and a perfect phase estimation, we can derive the overall transfer function. Comparing the MVDR beamformer in (36) with (50), we observe that the proposed system has a resulting transfer function:

$$\widetilde{H} = e^{j\phi_{\text{ref}}} \sqrt{\widetilde{\sigma}_N^2 \left(\frac{|H_1|^2}{\sigma_{N_1}^2} + \frac{|H_2|^2}{\sigma_{N_2}^2} + \cdots\right)}. \qquad (51)$$

That is,

$$\widetilde{G}_i = G_i^{\text{MVDR}} \widetilde{H}. \qquad (52)$$

Hence, the proposed system does not provide perfect equalization. However, the filter provides partial dereverberation, where the dips of the acoustic transfer functions are smoothed if the dips occur not in all transfer functions. Moreover, if the noise is uniform and stationary, the number of channels M is sufficiently high and in case of spatially uncorrelated channel coefficients, the sum $(|H_1|^2/\sigma_{N_1}^2 + |H_2|^2/\sigma_{N_2}^2 + \cdots)$ tends to a constant value independent of the frequency (cf. [15]).

7.4. Noise Reduction. As shown by the decomposition of the speech-distortion-weighted multichannel Wiener filter in (20), the noise reduction of the MWF is achieved by a single-channel Wiener filter. Therefore, we combine the proposed matched filter approach with a single channel Wiener filter. In the reminder of this section, we discuss the integration of the Wiener postfilter to the blind-matched filter approach. The considered system is shown in Figure 8.

The single channel Wiener filter in (20) can be rewritten to an equation which only depends on the output SNR γ:

$$G^{\text{WF}} = \frac{\gamma}{\gamma + \mu_W}. \qquad (53)$$

It is possible to integrate the Wiener filter function from (53) in the filter functions of the proposed blind-matched filter (38). This leads to the filter function G_i^{MFWF}, which consists of a blind matched filter (MF) with a single-channel Wiener postfilter (WF):

$$\begin{aligned} G_i^{\text{MFWF}} &= G^{\text{WF}} \widetilde{G}_i \\ &= \frac{\gamma}{\gamma + \mu_W} \sqrt{\frac{\gamma_i \widetilde{\sigma}_N^2}{\sigma_{N_i}^2 \gamma}} e^{j\hat{\Delta}_i} \\ &= \sqrt{\frac{\gamma_i \gamma \widetilde{\sigma}_N^2}{\sigma_{N_i}^2}} \frac{e^{j\hat{\Delta}_i}}{\gamma + \mu_W}, \end{aligned} \qquad (54)$$

thus the MSE with respect to the speech component of the combined signal is minimized.

7.5. Simulation Results. In this section, we present some simulation results for the proposed combining system with additional noise suppression. Therefore, we used the simulation setup described in Section 3. Table 2 presents the results for the simulated in-car environment using the configuration with the intermicrophone distance of 0.6 m. As an objective

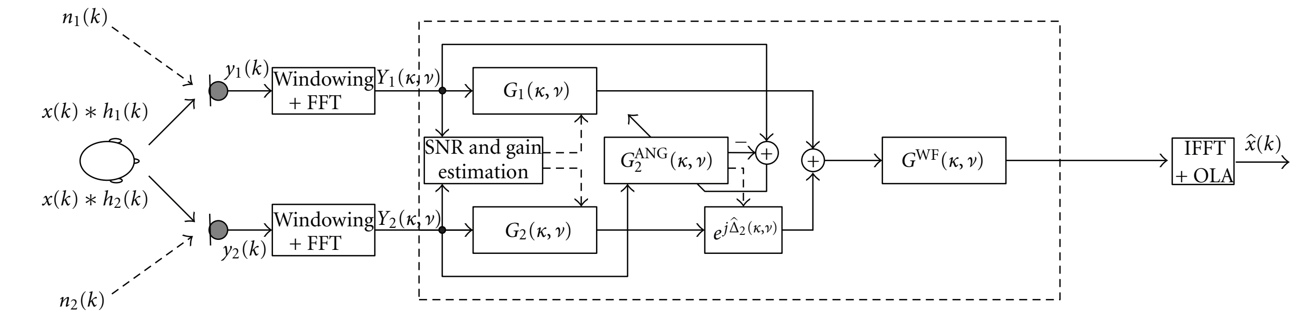


FIGURE 8: System structure of the blind-matched filtering with noise reduction for two channels.

TABLE 2: Simulation results for the proposed system in the car environment, cf. results on the right of Table 1.

d_{mic} [m]		0.6	
	100 km/h	140 km/h	defrost
SSNR ch.1 [dB]	−2.7	−7.7	−0.2
SSNR ch.2 [dB]	−1.1	−5.8	−0.6
SSNR [dB]	10.3	4.8	12.1
S-MOS	4.6	4.5	4.7
N-MOS	3.8	3.2	3.8
G-MOS	4.0	3.7	4.0

TABLE 3: Simulation results for the system in the conference environment with with $T_{60} = 0.25$ s.

Speaker pos.	1	2	3	4	5
SSNR ch.1 [dB]	8.8	8.3	8.5	10.1	13.2
SSNR ch.2 [dB]	8.3	10.1	10.2	9.9	7.8
SSNR ch.3 [dB]	13.6	10.1	9.9	8.0	9.3
$\mathbf{G}^{\mathrm{MFWF}}$ SSNR [dB]	21.3	19.8	19.8	19.6	20.8
S-MOS	4.7	4.7	4.7	4.8	4.8
N-MOS	4.6	4.5	4.5	4.6	4.6
G-MOS	4.3	4.3	4.3	4.4	4.4
$\mathbf{G}^{\mathrm{SDW}}$ SSNR [dB]	16.8	15.9	15.7	16.6	18.8
S-MOS	4.8	4.8	4.8	4.8	4.8
N-MOS	3.9	3.9	4.9	4.1	4.4
G-MOS	4.2	4.1	4.1	4.2	4.4

evaluation criterion, we calculated also the segmental SNR of the output signal. Also the input signal-to-noise ratios are shown in Table 2. Furthermore, we show results from an instrumental quality analysis in Table 1. Comparing these values with the results shown in Table 1, we observe that the proposed algorithm outperforms the SDW-MWF for this scenario. For this simulation, a higher noise overestimation factor μ_W can be used, while the S-MOS results are nearly the same in comparison to the results of SDW-MWF. This is because the proposed combining system partial equalizes the acoustic system and, therefore, the speech signal components are equalized with respect to the speech signal X at the speaker's mouth. Thus, the system can achieve a higher output SNR at the same level of speech distortion.

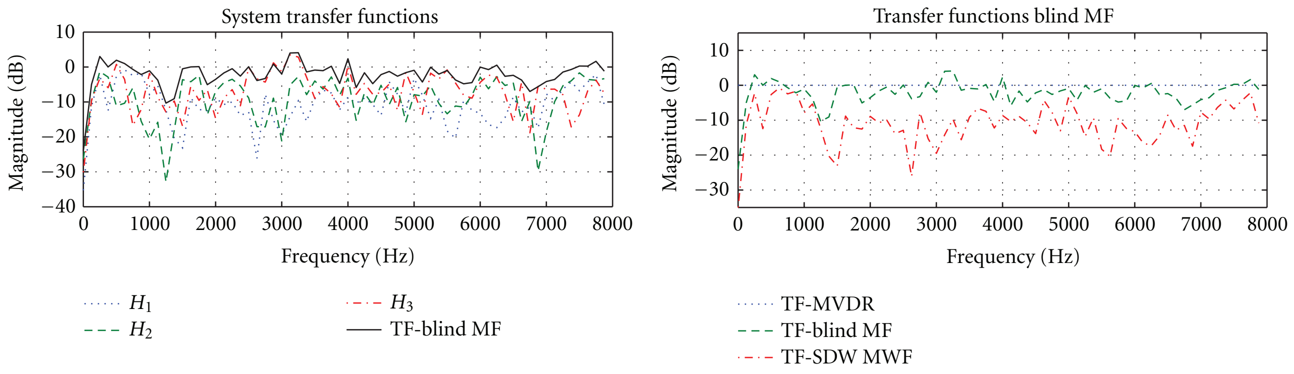
For the conference scenario, Table 3 shows the results of the performed simulations. It can be seen that the output SNR of the system as well as the MOS values is nearly the same for all speaker positions. This is a result of the proposed combining scheme.

The effect of the partial equalization is obvious by a comparison of the individual acoustic transfer functions with the overall system transfer function $\widetilde{H}(\nu)$. This is shown in Figure 9(a) for the speaker at position three, where the transfer functions between the speaker's mouth and the microphones are plotted as dashed and dotted lines. The overall system transfer function (including the system and the acoustic signal path) is plotted as solid line. It can be seen that the deep dips of the individual transfer functions are equalized. The overall transfer function follows the envelope of all transfer functions (which may include also the microphone characteristic). In Figure 9(b), the transfer functions of the SDW-MWF without the Wiener filter part and of the MVDR beamformer are plotted in comparison with the proposed blind-matched filter approach.

To show the applicability of the proposed system also in more reverberant environments, we used a simulated conference scenario with a reverberation time $T_{60} = 0.5$ s. For the generation of the room impulse responses, we used the image method described by Allen and Berkley in [24]. The results are presented in Table 4, again the output SNRs for the different speaker positions are in the same range. Also, simulation results for the SDW-MWF are given for a comparison of these two techniques.

8. Blind System Identification

In order to demonstrate that the filter $\widetilde{G}_i$ approximates the matched filter, we show that the structure in Figure 10 can be used for blind system identification. The SNR values for speech signals are fast time-varying. Hence, we use again FLMS filters G_i^{LMS} to estimate the average filter transfer functions. Note that if we have equal noise power spectral densities at all microphone inputs, the matched filter $\mathbf{G}^{\mathrm{MF}}$ is equal to the vector of transfer function coefficients $\mathbf{H}$ up to a common factor. This factor can vary with the frequency. Hence, by estimating the ideal matched filter, we estimate the linear acoustic system up to a common filter. Furthermore, note that many known approaches for blind



(a) Comparison of the blind MF with the transfer functions to the microphones

(b) Comparison of the blind MF with the ideal MVDR beamformer and the SDW-MWF

Figure 9: Comparison of the system transfer functions.

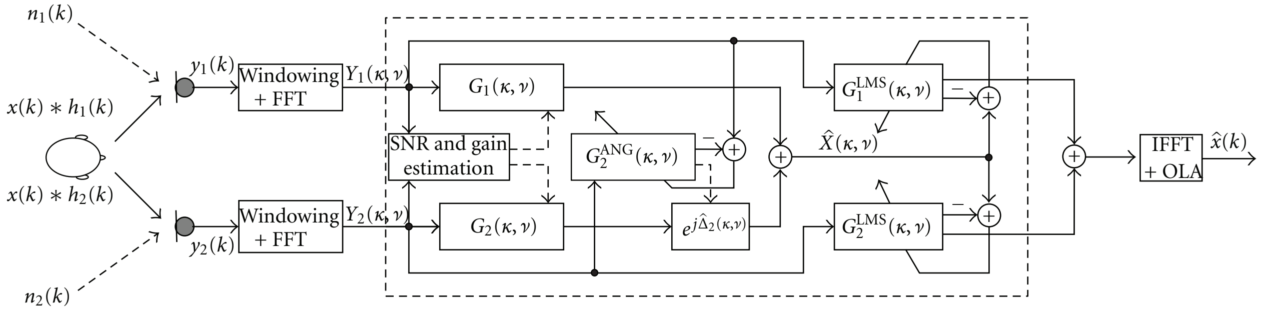


Figure 10: Basic system structure for the system identification approach for two channels.

Table 4: Simulation results for the system in a simulated conference environment with $T_{60} = 0.5$ s.

Speaker pos.	1	2	3	4	5
SSNR ch.1 [dB]	9.8	8.8	9.7	9.1	12.0
SSNR ch.2 [dB]	10.1	11.2	12.2	12.2	11.2
SSNR ch.3 [dB]	12.4	9.9	8.3	8.8	8.8
$\mathbf{G}^{\mathrm{MFWF}}$ SSNR [dB]	21.5	21.0	21.5	20.8	21.6
S-MOS	4.7	4.7	4.7	4.7	4.7
N-MOS	4.5	4.4	4.5	4.5	4.6
G-MOS	4.3	4.3	4.3	4.3	4.3
$\mathbf{G}^{\mathrm{SDW}}$ SSNR [dB]	17.8	17.0	17.1	16.9	19.0
S-MOS	4.8	4.7	4.7	4.8	4.8
N-MOS	3.8	3.7	3.7	3.7	4.1
G-MOS	4.1	4.0	4.1	4.1	4.2

system identification can only infer the M different channels up to a common filter [15]. Similarly, with the proposed system, all filters G_i are biased by a common factor. For equal noise power spectral densities, this common filter has the transfer function:

$$\tilde{H} = e^{j\phi_{\mathrm{ref}}}\sqrt{|H_1|^2 + |H_2|^2 + \cdots + |H_M|^2} \tag{55}$$

and the LMS filters should converge to

$$G_i^{\mathrm{LMS}} = \frac{H_i^*}{\tilde{H}}. \tag{56}$$

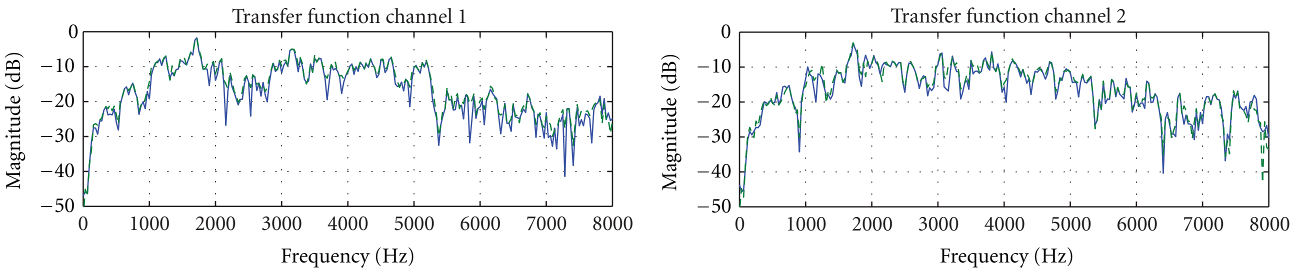
For simulations, we use the two microphone in-car setup with an intermicrophone distance of 0.6 m. We consider the driving situation with a car speed of 140 km/h. The magnitude of the actual transfer functions H_i and the magnitude of the corrected filter transfer function $G_i^{\mathrm{LMS}}\tilde{H}$ are depicted in Figure 11. We observe that the transfer functions are well approximated. As a quality measure, we use the distance

$$D_i = 10\log_{10}\left(\frac{\sum_\nu \left|G_i^{\mathrm{LMS}}\tilde{H} - H_i^*\right|^2}{\sum_\nu |H_i|^2}\right) \tag{57}$$

and obtain values of $D_1 = -16.4$ dB and $D_2 = -11$ dB after 5 seconds of speech activity. For a driving situation with a car speed of 100 km/h, we obtain $D_1 = -17.9$ dB and $D_2 = -11.9$ dB, respectively.

9. Conclusions

In this paper, we have presented a speech enhancement system with distributed microphones, where the array geometry is arbitrary and not a priori known. The system is based



(a)

(b)

FIGURE 11: Estimated (dashed line) and actual transfer functions for the two channels.

on a blind-matched filtering approach where the filter coefficients depend only on the input signal-to-noise ratios and the correlation between the input signals. For spatially uncorrelated but not necessarily uniform noise, the system provides an optimal signal combining that maximizes the output SNR.

Moreover, the presented approach achieves a partial equalization of the acoustic system up to a common filter. To demonstrate that the ideal filter coefficients can be reliably estimated, we have presented an application for blind system identification. The system is able to identify the *M* different channels up to a common filter. The presented simulation results indicate that this identification is robust against background noise. To provide a perfect equalization, the remaining filter ambiguity needs to be resolved separately. However, the presented system could also be combined with other speech dereverberation algorithms, for example, the single channel reverberation suppression algorithms presented in [38, 39]. The system assumes a single speech source, but a situation with more than one active speaker cannot be avoided in real conference scenarios, so further investigations are needed to evaluate the concept for such scenarios.

Appendix

A. Decomposition of the SDW-MWF

Let $\mathbf{A}$ and $\mathbf{B}$ be two square matrices, where $\mathbf{A}$ has full rank and $\mathbf{B}$ has rank one. In this case, we can rewrite the inverse of the sum $\mathbf{A}+\mathbf{B}$ using of the matrix inversion lemma [22]:

$$(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} - \frac{1}{1+g}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}, \tag{A.1}$$

where g is the trace of $\mathbf{A}^{-1}\mathbf{B}$.

Now, consider the term $(\mathbf{R}_S + \mu_W \mathbf{R}_N)^{-1}$ in the definition of the SDW-MWF in (18). $\mathbf{R}_S = \sigma_X^2 \mathbf{H}\mathbf{H}^\dagger$ has rank one due to the structure of the matrix $\mathbf{H}\mathbf{H}^\dagger$ (see, e.g., [40]), and if we assume that the noise variances at all microphones are nonzero, $\mathbf{R}_N$ will have full rank. Using (A.1) the inverse $(\mu_W \mathbf{R}_N + \mathbf{R}_S)^{-1}$ in (18) can be rewritten as

$$(\mu_W \mathbf{R}_N + \mathbf{R}_S)^{-1} = \frac{1}{\mu_W}\mathbf{R}_N^{-1} - \frac{\mathbf{R}_N^{-1}\mathbf{R}_S\mathbf{R}_N^{-1}}{\mu_W^2 (1+\mu_W^{-1}\sigma_x^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H})}. \tag{A.2}$$

Hence, we can decompose the SDW-MWF as follows [20]:

$$\begin{aligned}
\mathbf{G}^{\text{SDW}} &= \left(\frac{1}{\mu_W}\mathbf{R}_N^{-1} - \frac{\mathbf{R}_N^{-1}\mathbf{R}_S\mathbf{R}_N^{-1}}{\mu_W^2 (1+\mu_W^{-1}\sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H})}\right)\mathbf{R}_S\mathbf{u} \\
&= \frac{1}{\mu_W}\mathbf{R}_N^{-1}\left(\mathbf{I} - \frac{\sigma_X^2 \mathbf{H}\mathbf{H}^\dagger \mathbf{R}_N^{-1}}{\mu_W + \sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}\right)\sigma_X^2 \mathbf{H}\mathbf{H}^\dagger \mathbf{u} \\
&= \frac{1}{\mu_W}\mathbf{R}_N^{-1}\left(\mathbf{H} - \frac{\sigma_X^2 \mathbf{H}\mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}{\mu_W + \sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}\right)\sigma_X^2 \mathbf{H}^\dagger \mathbf{u} \\
&= \frac{1}{\mu_W}\mathbf{R}_N^{-1}\left(1 - \frac{\sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}{\mu_W + \sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}\right)\sigma_X^2 \mathbf{H}\mathbf{H}^\dagger \mathbf{u} \\
&= \left(1 - \frac{\sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}{\mu_W + \sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}\right)\frac{1}{\mu_W}\mathbf{R}_N^{-1}\sigma_X^2 \mathbf{H}\mathbf{H}^\dagger \mathbf{u} \\
&= \frac{\sigma_X^2}{\mu_W + \sigma_X^2 \mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}\mathbf{R}_N^{-1}\mathbf{H}\mathbf{H}^\dagger \mathbf{u} \\
&= \frac{\sigma_X^2}{\sigma_X^2 + \mu_W (\mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H})^{-1}}\frac{\mathbf{R}_N^{-1}\mathbf{H}}{\mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}\mathbf{H}^\dagger \mathbf{u} \\
&= \frac{\sigma_X^2}{\sigma_X^2 + \mu_W \sigma^2{}_{N_{\text{MVDR}}}}\frac{\mathbf{R}_N^{-1}\mathbf{H}}{\mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}\mathbf{H}^\dagger \mathbf{u} \\
&= \frac{\gamma}{\gamma + \mu_W}\frac{\mathbf{R}_N^{-1}\mathbf{H}}{\mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}\mathbf{H}^\dagger \mathbf{u} \\
&= G^{\text{WF}}\mathbf{G}^{\text{MVDR}}H_{\text{ref}}^*,
\end{aligned} \tag{A.3}$$

where $\sigma^2{}_{N_{\text{MVDR}}}$ is the noise variance at the output of $\mathbf{G}^{\text{MVDR}}$

$$\begin{aligned}
\sigma^2{}_{N_{\text{MVDR}}} &= \mathbf{G}^{\text{MVDR}\dagger}\mathbf{R}_N\mathbf{G}^{\text{MVDR}} \\
&= \frac{\mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}}{(\mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H})^2} \\
&= \left(\mathbf{H}^\dagger \mathbf{R}_N^{-1}\mathbf{H}\right)^{-1}.
\end{aligned} \tag{A.4}$$

Acknowledgments

The Research for this article was supported by DFG (FR 2673/2-1). Last but not the least, the authors would like to thank the reviewers for their constructive comments and suggestions which greatly improve the quality of this manuscript.

References

[1] E. Hänsler and G. Schmidt, *Speech and Audio Processing in Adverse Environments: Signals and Communication Technologie*, Springer, 2008.

[2] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[4] A. Guérin, R. Le Bouquin-Jeannès, and G. Faucon, "A two-sensor noise reduction system: applications for hands-free car kit," *Eurasip Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1125–1134, 2003.

[5] T. Gerkmann and R. Martin, "Soft decision combining for dual channel noise reduction," in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH '06-ICSLP*, pp. 2134–2137, Pittsburgh, Pa, USA, September 2006.

[6] J. Freudenberger, S. Stenzel, and B. Venditti, "Spectral combining for microphone diversity systems," in *Proceedings of the European Signal Processing Conference (EUSIPCO '09)*, pp. 854–858, Glasgow, Scotland, UK, 2009.

[7] S. Doclo, A. Spriet, M. Moonen, and J. Wouters, "Speech distortion weighted multichannel wiener filtering techniques for noise reduction," in *Speech Enhancement*, chapter 9, Springer, 2005.

[8] 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, 2007.

[9] 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.

[10] 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, no. 6, pp. 561–571, 2004.

[11] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, 1997.

[12] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *Eurasip Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[13] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1791–1801, 2007.

[14] I. Ram, E. Habets, Y. Avargel, and I. Cohen, "Multi-microphone speech dereverberation using LIME and least squares filtering," in *Proceedings of the European Signal Processing Conference (EUSIPCO '08)*, Lausanne, Switzerland, Augast 2008.

[15] D. Schmid and G. Enzner, "Robust subsystems for iterative multichannel blind system identification and equalization," in *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM '09)*, pp. 889–893, August 2009.

[16] J. Freudenberger, S. Stenzel, and B. Venditti, "Microphone diversity combining for In-car applications," *Eurasip Journal on Advances in Signal Processing*, vol. 2010, Article ID 509541, 2010.

[17] S. Applebaum, "Adaptive arrays," *IEEE Transactions on Antennas and Propagation*, vol. 24, no. 5, pp. 585–598, 1976.

[18] S. Doclo and M. Moonen, "Robust time-delay estimation in highly adverse acoustic environments," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 59–62, October 2001.

[19] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[20] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, chapter 9, Wiley & Sons, 2009.

[21] R. Martin and P. Vary, "Combined acoustic echo cancellation, dereverberation and noise reduction: a two microphone approach," *Annales Des Télécommunications*, vol. 49, no. 7-8, pp. 429–438, 1994.

[22] K. S. Miller, "On the inverse of the sum of matrices," *Mathematics Magazine*, vol. 54, no. 2, pp. 67–72, 1981.

[23] ITU-T, *Test Signals for Use in Telephonometry*, Recommendation ITU-T P.501, International Telecommunication Union, Geneva, Switzerland, 2007.

[24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustic Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[25] B. Holter and G. E. Oien, "The optimal weights of a maximum ratio combiner using an eigenfilter approach," in *Proceedings of the 5th IEEE Nordic Signal Processing Symposium (NORSIG '02)*, Hurtigruten, Norway, 2002.

[26] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[27] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH '95)*, vol. 2, pp. 1513–1516, Madrid, Spain, 1995.

[28] I. Cohen, "On speech enhancement under signal presence uncertainty," in *Proceedings of the Interntional Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 1, pp. 661–664, May 2001.

[29] M. Rahmani, A. Akbari, and B. Ayad, "An iterative noise cross-PSD estimation for two-microphone speech enhancement," *Applied Acoustics*, vol. 70, no. 3, pp. 514–521, 2009.

[30] R. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, p. 223, 2011.

[31] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '79)*, pp. 208–211, 1979.

[32] N. Dal Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Processing*, vol. 15, no. 1, pp. 43–56, 1988.

[33] M. Doerbecker, "Sind kohärenzbasierte Störgeräuschreduktionsverfahren für elektrische Hörhilfen geeignet?—Modelle zur Beschreibung der Kohärenzeigenschaften kopfbezogener Mikrofonsignale ," in *ITG-Fachtagung Sprachkommunikation und 9. Konferenz Elektronische Sprachsignalverarbeitung*, pp. 53–56, Dresden, Germany, 1998.

[34] W. Armbrüster, R. Czarnach, and P. Vary, "Adaptive noise cancellation with reference input," in *Signal Processing III*, pp. 391–394, Elsevier, 1986.

[35] ITU-T, *Objective Measurement of Active Speech Level*, Recommendation ITU-T P.56, International Telecommunication Union, Geneva, Switzerland, 1996.

[36] J. Reimes, H. W. Gierlich, F. Kettler, and S. Poschen, "Ein neues Verfahren zur objektiven Beurteilung der Sprachqualität bei Hintergrundger äuschen : 3QUEST," in *8. ITG-Fachtagung Sprachkommunikation*, 2008.

[37] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[38] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proceedings of the 15th Annual Workshop on Circuits*, vol. 10, pp. 250–254, 2004.

[39] H. W. Löllmann and P. Vary, "Low Delay Noise Reduction and Dereverberation for Hearing Aids," *Eurasip Journal on Advances in Signal Processing*, vol. 2009, Article ID 437807, 2009.

[40] G. Strang, *Linear Algebra and Its Applications*, Harcourt Brace Jovanovich, Lisbon, Portugal, 1988.

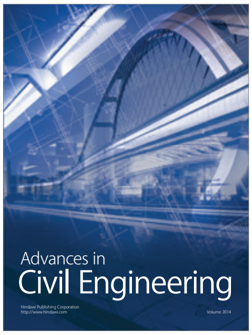
Journal of Engineering

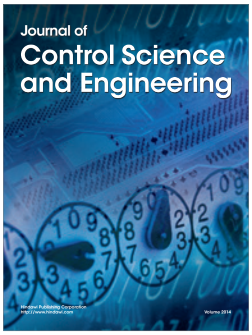
The Scientific World Journal

International Journal of Rotating Machinery

Journal of Sensors

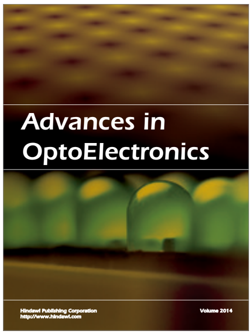
International Journal of Distributed Sensor Networks


Advances in Civil Engineering


Journal of Control Science and Engineering

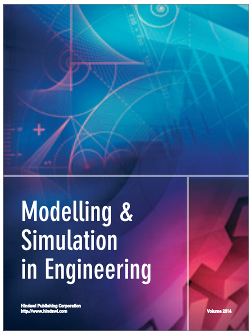
Journal of Robotics

Journal of Electrical and Computer Engineering

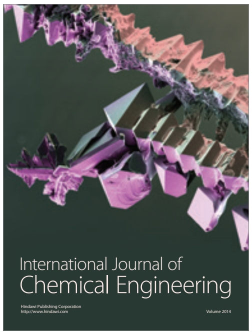

Advances in OptoElectronics

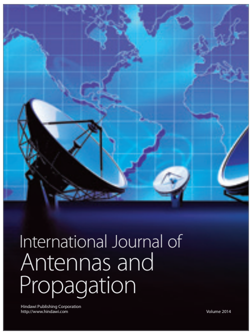
VLSI Design

International Journal of Navigation and Observation

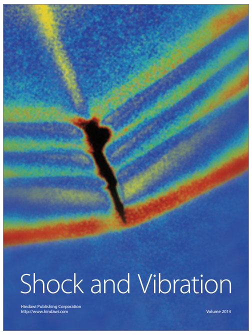

Modelling & Simulation in Engineering

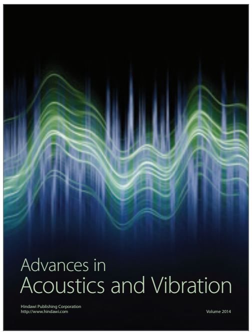
International Journal of Aerospace Engineering


International Journal of Chemical Engineering


International Journal of Antennas and Propagation

Active and Passive Electronic Components


Shock and Vibration


Advances in Acoustics and Vibration