Multi-Microphone Speech Dereverberation and Noise Reduction Using Relative Early Transfer Functions

O. Schwartz, S. Gannot, *Senior Member, IEEE,* and Emanuël A.P. Habets, *Senior Member, IEEE*

***Abstract*—In speech communication systems the microphone signals are degraded by reverberation and ambient noise. The reverberant speech can be separated into two components, namely, an early speech component that includes the direct path and some early reflections, and a late reverberant component that includes all the late reflections. In this paper, a novel algorithm to simultaneously suppress early reflections, late reverberation and ambient noise is presented. A multi-microphone minimum mean square error estimator is used to obtain a spatially filtered version of the early speech component. The estimator constructed as a minimum variance distortionless response (MVDR) beamformer (BF) followed by a postfilter (PF). Three unique design features characterize the proposed method. First, the MVDR BF is implemented in a special structure, named the nonorthogonal generalized sidelobe canceller (NO-GSC). Compared with the more conventional orthogonal GSC structure, the new structure allows for a simpler implementation of the GSC blocks for various MVDR constraints. Second, In contrast to earlier works, relative early transfer functions (RETFs) are used in the MVDR criterion rather than either the entire relative transfer functions (RTFs) or only the direct-path of the desired speech signal. An estimator of the RETFs is proposed as well. Third, the late reverberation and noise are processed by both the beamforming stage and the PF stage. Since the relative power of the noise and the late reverberation varies with the frame index, a computationally efficient method for the required matrix inversion is proposed to circumvent the cumbersome mathematical operation. The algorithm was evaluated and compared with two alternative multichannel algorithms and one single-channel algorithm using simulated data and data recorded in a room with a reverberation time of 0.5 s for various source-microphone array distances (1-4 m) and several signal-to-noise levels. The processed signals were tested using two commonly used objective measures, namely perceptual evaluation of speech quality and log-spectral distance. As an additional objective measure, the improvement in word accuracy percentage of an automatic speech recognition system is also demonstrated.**

Ofer Schwartz and Sharon Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel (e-mail: ofer.shwartz@live.biu.ac.il, Sharon.Gannot@biu.ac.il).

E. A. P. Habets is with the International Audio Laboratories Erlangen (a joint institution between the University of Erlangen-Nuremberg and Fraunhofer IIS), Erlangen, Germany (e-mail: emanuel.habets@audiolabs-erlangen.de).

I. Introduction

Dereverberation aims at the reduction of reverberation that is caused by a multitude of reflections from walls and other objects and has become a major research subject in the last decade due to theoretical advances in understanding the reverberation phenomenon and available computational power. Highly reverberant speech can be difficult to understand for both humans and machines, and can lead to listening fatigue. Existing dereverberation methods can be divided into two categories: reverberation cancelation and reverberation suppression [1]. Reverberation cancelation can be accomplished directly by inverting the acoustic system, or indirectly by first identifying and then equalizing the acoustic system. Since the clean speech is unobservable, these algorithms need to blindly estimate the acoustic system or its inverse directly. Reverberation suppression algorithms circumvent the cumbersome task of blind identification of the acoustic system and instead apply spectral enhancement procedures.

In the category of reverberation cancelation, multichannel linear prediction techniques were used to blindly equalize the acoustic impulse response (AIR) without the need to first identify the AIRs [2], [3]. In [4], a dual-channel reconstruction method was presented based on cepstrum techniques. A single-channel dereverberation method was presented in [5] based on the harmonic structure of the anechoic speech signal. The direct path was approximated by extracting its harmonic parameters from the reverberant signal and then the AIR was estimated by a division in the frequency domain. In [6], a two stage multichannel dereverberation method was proposed. In the first stage, the AIRs were extracted from the null subspace of the data matrix. In the second stage, these estimates were used to equalize the microphone signals using the classical multichannel inverse theorem (MINT) method [7]. More recently, researchers proposed to apply channel shortening techniques to compute the inverse of the AIRs [8]–[11].

Various technique fall into the category of reverberation suppression. Polack, in [12], formulated the AIR as an independent and identically distributed white Gaussian noise with an exponential decaying variance. This property was first utilized in [13] to show that the late reverberant power spectral density (PSD) can be expressed as a delayed and attenuated version of the instantaneous reverberant PSD. In [13], [14], a spectral subtraction algorithm was used to obtain an estimate of the early speech component using Polack's statistical model. This method was extended to the multi-microphone case in [15], by employing the single-channel spectral subtraction algorithm to the output of a delay and sum (DS) beamformer (BF). In this case, the late reverberant PSD was estimated by computing the spatial expectation.

In [16], [17], both reverberation and noise were considered by adding the late reverberation and noise PSDs. Since Polack's model does not take the direct-path into account, the reverberant PSD is overestimated when the direct-to-reverberation ratio (DRR) is larger than 0 dB. A model and PSD estimator that takes the DRR into account was proposed in [1], [18].

In [19], a single-channel estimate maximize (EM)-based algorithm for speech dereverberation and noise reduction was presented. The acoustic path was modeled as an auto regressive (AR) system, and the anechoic speech PSD was given an all-pole model. In the E-step, the reverberant speech is estimated (without the noise component), and in the M-step the acoustic path and the speech parameters are estimated (the noise parameters are assumed to bo known). Dereverberation is obtained by applying a multichannel Wiener filter. The EM algorithm is also used in [20]. The acoustic path of the late reverberation is modelled as an AR system. In the E-step, the Kalman smoother is applied to estimate the reverberant speech while in the M-step the AR coefficients of the acoustic path are estimated.

The minimum variance distortionless response (MVDR) BF, usually implemented using a generalized sidelobe canceller (GSC) structure [21], is a popular noise reduction algorithm [22], [23] that can be also useful for dereverberation. In [24] the AIRs are modelled as convolutive transfer function (CTF) to circumvent the requirement for very long processing frames in high reverberation levels. Similar to [23], the algorithm in [24] focuses on noise reduction and yields reverberant outputs. In [25] the fixed BF block in the GSC structure is replaced by a simple DS BF, while other blocks remain intact. It is interesting to note that the branches of the resulting GSC are not orthogonal anymore. In the current contribution we further elaborate on this issue.

In [26]–[29], a structure comprised of an MVDR BF and a postfilter (PF) was proposed. The MVDR was designed to suppress the ambient noise and the AIRs were modeled as delayed versions of the anechoic speech. The late reverberation was only suppressed by the PF stage, using a late reverberation level estimate based on Polack's model. In [29], the spatial coherence matrix of the noise is either estimated from the noisy data or alternatively, if an insufficient number of noise-only frames is available, set as a noise matrix of ideal diffused sound field with diagonal loading. In [30], a two stage approach was presented to jointly suppress reverberation and noise. In the first stage, a super-directive beamformer (SDBF) is used to generate a reference signal consisting of a dereverberated speech signal and residual noise. In the second stage, the microphone signals were utilized to obtain an estimate of the dereverberated speech signal. Various spatial filter structures and estimators were considered.

In the current contribution, we propose a multichannel minimum mean square error (MMSE) estimator (i.e. the multichannel Wiener filter (MWF)) to jointly reduce reverberation and noise. The estimator can be decomposed into a MVDR BF followed by a single-channel Wiener filter [31], [32]. In an ideal diffuse sound field the MVDR BF [33], [34] attains maximum directivity [35]. Therefore, by adopting the well-established modeling of the late reverberation as a diffuse sound field [36], [37], the MVDR BF is a natural choice for reverberation reduction.

The AIR is modeled by two components (that are assumed to be uncorrelated), namely the early reverberation (including the direct path and some early reflections) and the late reverberation [14], [36], [38]. The early reverberation is characterized by discrete reflections of sound waves on the walls and other rigid objects. In the short-time Fourier transform (STFT) domain, the early speech component can be modeled as a multiplication of the transformed signal frame and the frequency response of the early component of the AIR. The late reflections are usually dense, since they are a summation of many reflections arriving from all directions. Therefore, the late reverberation and ideal diffuse sounds field have very similar spatial properties. In the STFT domain the late reverberation can be modelled as a diffuse sound field with a time-varying level.

The MVDR BF is conveniently implemented in a GSC structure, comprised of three blocks in two branches. The fixed beamformer (FBF), which constitutes the upper branch, is responsible for maintaining a desired response towards the signal of interest. The lower branch, responsible of interference reduction, is comprised of the other two blocks: The blocking matrix (BM) blocks the signal of interest while the noise canceller (NC) cancels the interference. The two branches are usually orthogonal.

In our work, we have taken a unique design approach for the blocks of the GSC. First, we have adopted a GSC structure in which the two branches are nonorthogonal. Similarity, to [25], the FBF is implemented as a DS BF in order to enhance the direct arrival while incoherently adding the early reflections. In contrast to many earlier works, the proposed lower branch is not orthogonal to the upper branch. The BM is constructed such that the entire early speech component is blocked. For that, estimates of the relative early transfer functions (RETFs) are required. We propose to use the least squares (LS) estimator for this purpose. Since the interference signal in our case is highly non-stationary (since reverberation is actually a speech signal), implementing an adaptive solution is a cumbersome task. We are therefore proposing to implement the NC in a closed-form multichannel MMSE structure, utilizing the late reverberation level obtained by previously developed estimators. An efficient implementation of this block is then derived.

This paper is organized as follows. In Section II we formulate the joint dereverberation and noise reduction problem. In Section III the optimal MMSE multichannel solution is presented. In Section IV the MVDR component of the multichannel MMSE BF is implemented in a nonorthogonal generalized sidelobe canceller (NO-GSC) structure. In Section V estimation procedures for the various parameters of the system, namely the RETF and the interference PSD matrix, are presented. In Section VI, the performance of the proposed algorithm is evaluated. Section VIII is dedicated to concluding remarks.

II. Problem Formulation

We formulate the problem in the STFT domain where m denotes the time frame index and k denotes the frequency index. The late reverberation and the ambient noise are modeled as additive interference such that the ith microphone signal can be expressed as:

$$Y_i(m,k) = X_{\mathrm{e},i}(m,k) + R_i(m,k) + V_i(m,k), \tag{1}$$

where $R_i(m,k)$ and $V_i(m,k)$ denote the additive late reverberation and ambient noise received by the ith microphone, respectively. The early speech component of the observed signal microphone signal is denoted by $X_{\mathrm{e},i}(m,k)$. We further assume that the various components $X_{\mathrm{e},i}(m,k)$, $R_i(m,k)$ and $V_i(m,k)$ are mutually uncorrelated.

We also assume that the observed early speech component at the ith microphone can be approximated in the STFT domain as a multiplication of an anechoic speech signal $S(m,k)$ and the slowly time-varying early transfer function (ETF) $G_{\mathrm{e},i}(k)$, that models the direct path and some early reflections from the source to the ith microphone:

$$X_{\mathrm{e},i}(m,k) = G_{\mathrm{e},i}(k)S(m,k). \tag{2}$$

The N microphone signals can be stacked in a vector form:

$$\begin{aligned}\mathbf{y}(m,k) &= \begin{bmatrix} Y_1(m,k) & Y_2(m,k) & \dots & Y_N(m,k) \end{bmatrix}^T \\ &= \mathbf{x}_{\mathrm{e}}(m,k) + \mathbf{r}(m,k) + \mathbf{v}(m,k) \\ &= \mathbf{g}_{\mathrm{e}}(k)S(m,k) + \mathbf{r}(m,k) + \mathbf{v}(m,k), \end{aligned} \tag{3}$$

where

$$\begin{aligned}\mathbf{x}_{\mathrm{e}}(m,k) &= \begin{bmatrix} X_{\mathrm{e},1}(m,k) & X_{\mathrm{e},2}(m,k) & \dots & X_{\mathrm{e},N}(m,k) \end{bmatrix}^T \\ \mathbf{r}(m,k) &= \begin{bmatrix} Y_1(m,k) & Y_2(m,k) & \dots & Y_N(m,k) \end{bmatrix}^T \\ \mathbf{v}(m,k) &= \begin{bmatrix} V_1(m,k) & V_2(m,k) & \dots & V_N(m,k) \end{bmatrix}^T \\ \mathbf{g}_{\mathrm{e}}(k) &= \begin{bmatrix} G_{\mathrm{e},1}(k) & G_{\mathrm{e},2}(k) & \dots & G_{\mathrm{e},N}(k) \end{bmatrix}^T. \end{aligned}$$

It should be noted that the signal model in (3) differs from the model proposed in [23], in which the microphone signal was defined as the sum of i) the entire reverberant signal that consists of the direct-path, early reflections, and late reverberation, and ii) the noise signal.

The probability density function (p.d.f.) of the observed data given the anechoic speech and the p.d.f. of the anechoic speech are, respectively, modelled as a complex Gaussian probability functions:

$$\begin{aligned} & f(\mathbf{y}(m,k)|S(m,k);\mathbf{g}_{\mathrm{e}}(k),\mathbf{\Phi}(m,k)) \\ & \qquad = \mathcal{N}_{\mathrm{C}}\left(\mathbf{y}(m,k);\mathbf{g}_{\mathrm{e}}(k)S(m,k),\mathbf{\Phi}(m,k)\right) \end{aligned} \tag{4}$$

$$f(S(m,k);\phi_S(m,k)) = \mathcal{N}_{\mathrm{C}}\left(S(m,k);0,\phi_S(m,k)\right) \tag{5}$$

where $\mathbf{\Phi}(m,k)$ is the PSD matrix of the late reverberation plus ambient noise:

$$\mathbf{\Phi}(m,k) = \mathbf{\Phi}_{\mathbf{r}}(m,k) + \mathbf{\Phi}_{\mathbf{v}}(m,k), \tag{6}$$

with $\mathbf{\Phi}_{\mathbf{r}}(m,k) = E\{\mathbf{r}(m,k)\mathbf{r}^H(m,k)\}$ and $\mathbf{\Phi}_{\mathbf{v}}(m,k) =$

$E\{\mathbf{v}(m,k)\mathbf{v}^H(m,k)\}$ the PSD matrices of the late reverberation and the ambient noise, respectively. The PSD of the anechoic speech is denoted $\phi_S(m,k) = E\{|S(m,k)|^2\}$. The conjugate-transpose of $\mathbf{a}$ is denoted $\mathbf{a}^H$ and $E\{A\}$ denotes the mathematical expectation of the random variable A.

The aim of this work is to provide an optimal multichannel estimate of a *filtered version* of the source signal that is given by

$$S_F(m,k) = F(k)S(m,k) \tag{7}$$

where $F(k)$ denotes the transfer function of a filter. The MMSE estimate of $S_F(m,k)$ is then given by

$$\widehat{S}_F(m,k) = E\left\{S_F(m,k)|\mathbf{y}(m,k)\right\}. \tag{8}$$

In the following section the MMSE estimator of $S_F(m,k)$ is implemented as a concatenation of the MVDR BF and a single-channel Wiener filter.

III. Optimal Multichannel Dereverberation and Noise Reduction

In this section we first describe the optimal MMSE estimator of the filtered signal $S_F(m,k)$ and then discuss various possible choices for $F(k)$. To simplify the derivation, we first rewrite the received signal model (3) in terms of the filtered signal:

$$\begin{aligned}\mathbf{y}(m,k) &= \frac{\mathbf{g}_\text{e}(k)}{F(k)}\,F(k)S(m,k) + \mathbf{r}(m,k) + \mathbf{v}(m,k)\\ &= \tilde{\mathbf{g}}_\text{e}(k)\;S_F(m,k) + \mathbf{r}(m,k) + \mathbf{v}(m,k)\end{aligned} \tag{9}$$

where $\tilde{\mathbf{g}}_\text{e}(k) \triangleq \mathbf{g}_\text{e}(k)/F(k)$. Define also, $\phi_{S_F}(m,k)$ as the PSD of the filtered version of the source signal.

A. MMSE Estimator

Since $S_F(m,k)$ and $\mathbf{y}(m,k)$ are assumed to be zero-mean complex Gaussian random variables, the MMSE estimator of $S_F(m,k)$ can be calculated using:

$$\begin{aligned}\widehat{S}_F(m,k) &= E\{S_F(m,k)\mathbf{y}^H(m,k)\} \qquad (10)\\ &\quad\times E\{\mathbf{y}(m,k)\mathbf{y}^H(m,k)\}^{-1}\,\mathbf{y}(m,k)\\ &= \phi_{S_F}(m,k)\tilde{\mathbf{g}}_\text{e}^H(k)\\ &\quad\times\left[\phi_{S_F}(m,k)\tilde{\mathbf{g}}_\text{e}(k)\tilde{\mathbf{g}}_\text{e}^H(k) + \boldsymbol{\Phi}(m,k)\right]^{-1}\mathbf{y}(m,k). \qquad (11)\end{aligned}$$

Using the Woodbury identity [39] and some straightforward algebraic steps, $\widehat{S}_F(m,k)$ can be expressed as

$$\begin{aligned}\widehat{S}_F(m,k) = &\underbrace{\frac{\phi_{S_F}(m,k)}{\phi_{S_F}(m,k) + [\tilde{\mathbf{g}}_\text{e}^H(k)\boldsymbol{\Phi}^{-1}(m,k)\tilde{\mathbf{g}}_\text{e}(k)]^{-1}}}_{H_\text{W}(m,k)}\\ &\times\underbrace{\frac{\tilde{\mathbf{g}}_\text{e}^H(k)\boldsymbol{\Phi}^{-1}(m,k)}{\tilde{\mathbf{g}}_\text{e}^H(k)\boldsymbol{\Phi}^{-1}(m,k)\tilde{\mathbf{g}}_\text{e}(k)}}_{\mathbf{h}^H_\text{MVDR}(m,k)}\mathbf{y}(m,k).\end{aligned} \tag{12}$$

The MVDR BF, $\mathbf{h}_\text{MVDR}(m,k)$, is the well-known solution of the following optimization criterion:

$$\begin{aligned}\mathbf{h}_\text{MVDR}(m,k) = \operatorname*{argmin}_{\mathbf{h}}\,\mathbf{h}^H\boldsymbol{\Phi}(m,k)\mathbf{h}&\\ \text{subject to } \mathbf{h}^H\tilde{\mathbf{g}}_\text{e}(k) = 1&\end{aligned} \tag{13}$$

and $H_\text{W}(m,k)$ is the single-channel Wiener filter at the output of $\mathbf{h}_\text{MVDR}(m,k)$.

B. Alternative Constraints

Different choices for $F(k)$ yield different array manifold vectors, $\tilde{\mathbf{g}}_\text{e}(k)$. Some array manifold vectors can be more easily estimated than the others.

With $F(k) = 1$ we aim at estimating the anechoic speech $S(m,k)$, requiring an estimate of $\mathbf{g}_\text{e}(k)$. Although this would be a very attractive choice, it remains a major challenge to blindly estimate the ETFs $\mathbf{g}_\text{e}(k)$.

Adopting the idea presented in [23], we can aim instead at estimating the early speech component, as received by the first microphone by using $F(k) = G_{\text{e},1}(k)$. As a result, we require an estimate of the RETF $\tilde{\mathbf{g}}_\text{e}(k)$ that can be estimated more easily compared to the ETF $\mathbf{g}_\text{e}(k)$.

Here we aim at a more general filtered version of the early speech component. We can either apply a single-channel filter to the anechoic speech, as proposed in [40], or a multichannel filter, as proposed in [25], [30], [41]. Here, the latter idea is adopted such that

$$F(k) = \mathbf{h}_\text{d}^H(k)\;\mathbf{g}_\text{e}(k). \tag{14}$$

where $\mathbf{h}_\text{d}(k)$ denotes a filter vector of a signal-independent BF. In contrast to the aforementioned works, we aim here at finding a spatially filtered version of the ETFs, $\mathbf{h}_\text{d}^H(k)\;\mathbf{g}_\text{e}$, rather than the acoustic transfer functions (ATFs), given by $\mathbf{h}_\text{d}^H(k)\;\mathbf{g}$.

IV. A NONORTHOGONAL GSC AND POSTFILTERING

The MVDR BF is conveniently implemented in a GSC structure [23], [42]:

$$\mathbf{h}_{\text{MVDR}}(m,k) = \mathbf{h}_0(k) - \mathbf{B}(k)\mathbf{h}_{\text{NC}}(m,k) \qquad (15)$$

where $\mathbf{h}_0(k)$ is the FBF satisfying $\mathbf{h}_0^H(k)\mathbf{g}_{\text{e}}(k) = F(k)$, $\mathbf{B}(k)$ is the BM satisfying $\mathbf{B}^H(k)\mathbf{g}_{\text{e}}(k) = \mathbf{0}$, and $\mathbf{h}_{\text{NC}}(m,k)$ is the NC that is responsible of mitigating the residual reverberation and ambient noise at the output of the FBF.

In previous works the MVDR filter vector is commonly decomposed into two *orthogonal* filter vectors. In this work we propose to decompose the MVDR filter vector into two *nonorthogonal* filter vectors, as illustrated in Fig. 1.

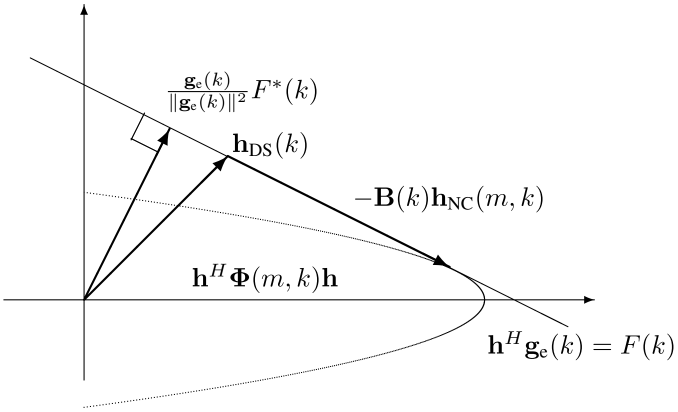

$\frac{\mathbf{g}_{\text{e}}(k)}{\|\mathbf{g}_{\text{e}}(k)\|^2}F^*(k)$
$\mathbf{h}_{\text{DS}}(k)$
$-\mathbf{B}(k)\mathbf{h}_{\text{NC}}(m,k)$
$\mathbf{h}^H\mathbf{\Phi}(m,k)\mathbf{h}$
$\mathbf{h}^H\mathbf{g}_{\text{e}}(k) = F(k)$

Fig. 1. Geometric interpretation of the nonorthogonal GSC components.

A. Fixed Beamformer

Following the orthogonal decomposition, the fixed BF is given by [23]

$$\mathbf{h}_0(k) = \frac{\mathbf{g}_{\text{e}}(k)}{\|\mathbf{g}_{\text{e}}(k)\|^2}F^*(k) = \frac{\tilde{\mathbf{g}}_{\text{e}}(k)}{\|\tilde{\mathbf{g}}_{\text{e}}(k)\|^2}, \qquad (16)$$

which can be slowly time-varying and difficult to implement. Specifically, it can be easily verified that this FBF is a noncausal filter and might be relatively long.

Using the constraint in (14), the FBF is given by

$$\mathbf{h}_0(k) = \mathbf{h}_{\text{d}}(k). \qquad (17)$$

It can be easily verified that

$$\mathbf{h}_0^H(k)\mathbf{y}(m,k) = \underbrace{\mathbf{h}_{\text{d}}^H(k)\mathbf{g}_{\text{e}}(k)}_{F(k)}S(m,k) + \underbrace{\mathbf{h}_{\text{d}}^H(k)\left[\mathbf{r}(m,k)+\mathbf{v}(m,k)\right]}_{\text{residual reverberation and noise}}. \qquad (18)$$

The latter choice for the FBF is independent of $\mathbf{g}_{\text{e}}(k)$, and therefore no transfer functions have to be estimated to construct the FBF.

By selecting $\mathbf{h}_{\text{d}}(k) = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T$ it follows that $F(k) = G_{\text{e},1}(k)$. We, however, are using instead the well-known DS BF given by:

$$\mathbf{h}_{\text{d}}(k) = \frac{1}{N}\begin{bmatrix} 1 & \exp(jk\tau_2) & \dots & \exp(jk\tau_N) \end{bmatrix}^T, \qquad (19)$$

where τ_i is the time difference of arrival (TDOA) between the ith and the 1st microphone. Apart from the direct path, all of the other early reflections are assumed to be incoherent, such that the DS BF enhances the direct path and suppress the early reflections. Alternatively, we can use a super-directive BF, as proposed in [30]. While the SDBF achieves the highest directivity index, we have chosen the DS BF since it is simpler to implement and since it achieves the highest white noise gain (WNG), hence exihibiting higher robustness to sensor gain and phase mismatches compared with the SDBF.

B. Blocking Matrix

The purpose of the BM is to block the early speech components and to provide a good reference for the late reverberation (and ambient noise). It is very important to avoid leakage of early speech components to the BM output to mitigate the *self-cancellation* phenomenon, which usually results in a severe speech distortion. Note that the BM does not dependent on $F(k)$.

As mentioned above, the blocking matrix should satisfy $\mathbf{B}^H(k)\mathbf{g}_e(k) = 0$, and therefore should also satisfy $\mathbf{B}^H(k)\tilde{\mathbf{g}}_e(k) = 0$. Provided that the multiplicative transfer function (MTF) assumption holds, the sparse blocking matrix [23] based on the RETFs is given by

$$\mathbf{B}(k) = \begin{bmatrix} -\tilde{G}^*_{e,2}(k) & -\tilde{G}^*_{e,3}(k) & \dots & -\tilde{G}^*_{e,N}(k) \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix}, \qquad (20)$$

where A^* denotes the conjugate of A, $\tilde{G}_{e,i}(m,k)$ is the RETF, defined as the ratio of the ETF from the source to the ith microphone and the ETF from the source to the first microphone (arbitrarily chosen here as the reference microphone), i.e.,

$$\tilde{G}_{\text{e},i}(k) = \frac{G_{\text{e},i}(k)}{G_{\text{e},1}(k)}. \qquad (21)$$

C. Noise Canceller

The NC is obtained by minimizing the interference power at the output of the MVDR BF. Since the FBF satisfies the constraint, and the blocking matrix ensures that all reference signals are uncorrelated with respect to $S(m,k)$, the NC filters can be obtained by solving the following unconstraint minimization problem:

$$\mathbf{h}_{\text{NC}}(m,k) = \underset{\mathbf{h}}{\operatorname{argmin}} \left(\mathbf{h}_0(k) - \mathbf{B}(k)\mathbf{h}\right)^H \boldsymbol{\Phi}(m,k) \left(\mathbf{h}_0(k) - \mathbf{B}(k)\mathbf{h}\right). \tag{22}$$

It is easily verified that $\mathbf{h}(m,k)$ is a multichannel Wiener filter, where the input signals are the BM output signals and the FBF output signal is the desired signal. If the BM outputs are indeed free from the early speech components, the Wiener filter tends only to reduce the residual late reverberation and ambient noise at the FBF output. Otherwise, cancellation of the early speech components at the MVDR BF output might occur. The closed-form solution of (22) is given by:

$$\mathbf{h}_{\text{NC}}(m,k) = \left(\mathbf{B}^H(k)\boldsymbol{\Phi}(m,k)\mathbf{B}(k)\right)^{-1} \times \mathbf{B}^H(k)\boldsymbol{\Phi}(m,k)\mathbf{h}_0(k). \tag{23}$$

D. Postfilter

As shown in [32], [43] and (12), the multichannel Wiener filter can be decomposed into an MVDR BF and a single-channel Wiener filter $H_{\text{W}}(m,k)$. The single-channel Wiener filter can be written as

$$H_{\text{W}}(m,k) = \frac{\xi(m,k)}{\xi(m,k)+1}, \tag{24}$$

where $\xi(m,k)$ is the *a priori* signal-to-reverberation plus noise ratio (SRNR) that is usually unobservable. The *a priori* SRNR can be recursively estimated by the posterior SRNR as proposed in [14]:

$$\xi(m,k) = \beta_r |H_{\text{W}}(m-1,k)|^2 \gamma(m-1,k) + (1-\beta_r) \max\left\{\gamma(m,k)-1, 0\right\}, \tag{25}$$

where β_r is a weighting factor and $\gamma(m,k)$ is the *a posteriori* SRNR at the MVDR output given by:

$$\gamma(m,k) = \frac{|\mathbf{w}_{\text{MVDR}}^H(m,k)\,\mathbf{y}(m,k)|^2}{\tilde{\phi}_R(m,k)+\tilde{\phi}_V(m,k)}. \tag{26}$$

The residual reverberation $\tilde{\phi}_R(m,k)$ and the residual noise $\tilde{\phi}_V(m,k)$ at the output of the MVDR stage are given by $\mathbf{w}_{\text{MVDR}}^H \boldsymbol{\Phi}_{\mathbf{r}} \mathbf{w}_{\text{MVDR}}$ and $\mathbf{w}_{\text{MVDR}}^H \boldsymbol{\Phi}_{\mathbf{v}} \mathbf{w}_{\text{MVDR}}$, respectively.

To minimize speech distortion and musical noise, the Wiener filter is lower-bounded by a time and frequency dependent gain $H_{\text{min}}(m,k)$. In this work, $H_{\text{min}}(m,k)$ is chosen such that a weighted minimum attenuation of the noise and the reverberation is obtained. The lower bound is then given by

$$H_{\text{min}}(m,k) = \frac{H_{\text{min},R}\,\tilde{\phi}_R(m,k) + H_{\text{min},V}\,\tilde{\phi}_V(m,k)}{\tilde{\phi}_R(m,k)+\tilde{\phi}_V(m,k)}, \tag{27}$$

where $H_{\text{min},R}$ and $H_{\text{min},V}$ are used to control the maximum amount of reverberation and noise reduction, respectively.

V. Parameters Estimation

The MMSE estimator requires an estimate of two parameters, namely the PSD matrix of the interference and the RETFs. The PSD matrix of the late reverberation is modeled as a diffuse sound field and the reverberation level is estimated by Polack's model [38]. To estimate the RETFs, we first estimate the early speech components $\widehat{X}_{\text{e},1}(m,k),\ldots,\widehat{X}_{\text{e},N}(m,k)$ utilizing N single-channel dereverberation filters. Then, the RETFs are identified using N independent LS estimators.

A. Interference PSD Matrix Estimation

First, we derive an estimator for the interference PSD matrix. Since the late reverberation and the noise are assumed to be uncorrelated, the estimation of their PSD matrices can be made separately. Here we model the late reverberation as a diffuse sound field with time-varying level.

An estimate of the PSD level at each microphone, $\phi_{R_i}(m,k)$, can be obtained using Polack's model [12] (c.f. [14], [26], [38], [44], [45]) after compensating for the noise level $\widehat{\phi}_{V,i}(m-L,k)$ at each microphone:

$$\widehat{\phi}_{R,i}(m,k) = \exp(-2\alpha RL) \times \left[\widehat{\phi}_{Y,i}(m-L,k) - \widehat{\phi}_{V,i}(m-L,k)\right], \tag{28}$$

where $\alpha = \frac{3\log(10)}{T_{60} f_s}$, L is the time in frames (measured with respect to the arrival time of the direct sound) indicating the beginning of late reverberation, R is the number of samples between two subsequent STFT frames, T_{60} is the reverberation time, and f_s is the sampling frequency in Hz.

The PSD of $Y_i(m,k)$ can be directly estimated from the microphone signals using

$$\widehat{\phi}_{Y,i}(m,k) = \beta_y \widehat{\phi}_{Y,i}(m-1,k) + (1-\beta_y)|Y_i(m,k)|^2 \tag{29}$$

where β_y is a forgetting factor. We assume that the source is sufficiently far from the microphones and that the late reverberant sound field is homogeneous such that the reverberation level is approximately equal for all microphones, i.e., $\phi_{R,i}(m,k) \equiv \phi_R(m,k)$ for all $i \in \{1,2,\dots,N\}$. This assumption might be violated if the distance between the speaker and microphones is small. When this assumption holds, an estimate of the reverberation level is obtained by averaging the PSD estimates across all channels [26], [28]:

$$\widehat{\phi}_R(m,k) = \frac{1}{N}\sum_{i=1}^{N}\widehat{\phi}_{R,i}(m,k). \tag{30}$$

By modeling the late reverberation as an ideal spherical diffuse sound field the interference PSD matrix is given by:

$$\mathbf{\Phi}(m,k) = \phi_R(m,k)\mathbf{\Gamma}(k) + \mathbf{\Phi}_{\mathbf{v}}(m,k) \tag{31}$$

where $\mathbf{\Gamma}(k)$ is the spatial coherence matrix of the spherical diffuse sound field [46], [47]

$$\mathbf{\Gamma}(k) = \begin{bmatrix} \mathrm{sinc}\left(\frac{f_s k d_{1,1}}{Kc}\right) & \dots & \mathrm{sinc}\left(\frac{f_s k d_{1,N}}{Kc}\right) \\ \vdots & \ddots & \vdots \\ \mathrm{sinc}\left(\frac{f_s k d_{N,1}}{Kc}\right) & \dots & \mathrm{sinc}\left(\frac{f_s k d_{N,N}}{Kc}\right) \end{bmatrix}, \tag{32}$$

where $\mathrm{sinc}(x) = \sin(x)/x$, K is the number of frequency bins, $d_{i,j}$ is the inter-distance between microphones i and j, and c is the sound velocity.

The noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(m,k)$ can be estimated during speech-absence by using an estimate of the speech presence probability (c.f. [48]–[51]). Estimating the noise PSD matrix is beyond the scope of this contribution.

B. Relative Early Transfer Function Estimation

According to (21), it can be easily verified that

$$X_{\mathrm{e},i}(m,k) = \tilde{G}_{\mathrm{e},i}(k)X_{\mathrm{e},1}(m,k);\ i = 2,\dots,N. \tag{33}$$

Therefore, an estimate of the RETF $\tilde{G}_{\mathrm{e},i}(k)$ can be obtained using an estimate of the early speech components $\widehat{X}_{\mathrm{e},1}(m,k)$ and $\widehat{X}_{\mathrm{e},i}(m,k)$.

Here we propose to use a single-channel dereverberation filter to estimate the ith early speech component in the MMSE sense. An estimate of the early speech component is given by

$$\widehat{X}_{\mathrm{e},i}(m,k) = H_{\mathrm{e},i}(m,k)Y_i(m,k), \tag{34}$$

where $H_{\mathrm{e},i}(m,k)$ denotes the single-channel Wiener filter applied to the ith microphone that is given by:

$$H_{\mathrm{e},i}(m,k) = \frac{\xi_i(m,k)}{1+\xi_i(m,k)}. \tag{35}$$

The *a priori* SRNR $\xi_i(m,k)$ can be recursively estimated similarly to (25), with the *a posteriori* SRNR being:

$$\gamma_i(m,k) = \frac{|Y_i(m,k)|^2}{\widehat{\phi}_{R,i}(m,k)+\widehat{\phi}_{V,i}(m,k)}. \tag{36}$$

Here $\widehat{\phi}_{R,i}$ can be replaced by $\widehat{\phi}_R$ using (30). Finally, a lower bound is applied to the Wiener filter to control the maximum attenuation of the noise and reverberation and to limit the distortion of the early speech component:

$$H_{\min,i}(m,k) = \frac{H_{\min,r}\,\widehat{\phi}_{R,i}(m,k) + H_{\min,v}\,\widehat{\phi}_{V,i}(m,k)}{\widehat{\phi}_{R,i}(m,k)+\widehat{\phi}_{V,i}(m,k)}. \tag{37}$$

Multiplying both sides in (34) by $X^*_{e,1}(m,k)$ and taking the expectation, we find that

$$\phi_{X_{\mathrm{e},i},X_{\mathrm{e},1}}(m,k) = \tilde{G}_{\mathrm{e},i}(k)\phi_{X_{\mathrm{e},1},X_{\mathrm{e},1}}(m,k), \tag{38}$$

which can be used to formulate a LS optimization criterion for the estimation of the RETF. Assuming that the RETF are slowly time varying, and hence mat be considered time-invariant during the latest M time frames, an LS estimate of $\tilde{G}_{\mathrm{e},i}(m,k)$ is given by[1]:

$$\widehat{\tilde{G}}_{\mathrm{e},i}(m,k) = \frac{\sum_{m'=m-M+1}^{m}\phi_{X_{\mathrm{e},i},X_{\mathrm{e},1}}(m',k)\,\phi_{X_{\mathrm{e},1},X_{\mathrm{e},1}}(m',k)}{\sum_{m'=m-M+1}^{m}\phi^2_{X_{\mathrm{e},1},X_{\mathrm{e},1}}(m',k)}. \tag{39}$$

The auto- and cross-PSDs are, respectively, recursively estimated using:

$$\widehat{\phi}_{X_{\mathrm{e},1},X_{\mathrm{e},1}}(m,k) = \beta_e\widehat{\phi}_{X_{\mathrm{e},1},X_{\mathrm{e},1}}(m-1,k) + (1-\beta_e)|\widehat{X}_{\mathrm{e},1}(m,k)|^2 \tag{40}$$

and

$$\widehat{\phi}_{X_{\mathrm{e},i},X_{\mathrm{e},1}}(m,k) = \beta_e\widehat{\phi}_{X_{\mathrm{e},i},X_{\mathrm{e},1}}(m-1,k) + (1-\beta_e)\widehat{X}_{\mathrm{e},i}(m,k)\widehat{X}^*_{\mathrm{e},1}(m,k). \tag{41}$$

The procedure for estimating the RETFs is summarized in Algorithm 1. Finally, note that the early speech components that are estimated using the single-channel Wiener filter are only used to estimate the RETFs. The final estimate of the early speech component is obtained in the MMSE sense using all microphone signals.

[1] Since $X_{\mathrm{e},i}(m,k)$ is estimated using the single-channel Wiener filter, the obtained phase of $\widehat{X}_{\mathrm{e},i}(m,k)$ is equal to the phase of $Y_i(m,k)$. Therefore, there is an inherent inaccuracy in the estimation of $X_{\mathrm{e},i}(m,k)$. If the phase error has zero-mean, the LS estimate minimizes this inaccuracy.

C. Reducing the Computational Complexity

It should be noted that the late reverberant signal component is highly time-varying and therefore the NC given by (23) needs to be calculated for every time frame m and frequency bin k, which results in a high computational burden. If the RETFs and the noise PSD matrix are slowly time-varying, it can be deduced that $\mathbf{B}$, $\mathbf{h}_0$ and $\mathbf{\Phi_v}$ are also slowly time-varying. Since the spatial properties of the late reverberation are assumed to be time-invariant, the respective spatial coherence matrix $\mathbf{\Gamma}$ is also time-invariant. The only parameter that is (highly) time-dependent is the late reverberation level, $\phi_R(m,k)$.

In a noiseless or, at least, high signal to noise ratio (SNR) scenarios, the noise PSD $\phi_V(m,k)$ can be neglected, and hence the late reverberation PSD $\phi_R(m,k)$ in (23) cancels out[2] yielding:

$$\mathbf{h}_{\text{NC}} = (\mathbf{B}^H\mathbf{\Gamma}\mathbf{B})^{-1}\mathbf{B}^H\mathbf{\Gamma}\mathbf{h}_0. \tag{42}$$

This entails significant computational efficiency, since *all* the GSC components can be calculated in advance.

However, when the SNR is low, the MVDR should also suppress the noise. In this case, the PSD matrix $\mathbf{\Phi}(m,k)$ of the late reverberation plus noise is dependent on the instantaneous reverberation-to-noise ratio and is therefore time-varying. This requires the calculation of $\mathbf{h}_{\text{NC}}$ for each time frame and frequency bin. The most expensive operation in this calculation is the matrix inversion $(\mathbf{B}^H\mathbf{\Phi}\mathbf{B})^{-1}$. In the following, we show how to reduce the number of calculations by following some algebraic steps.

First, define the eigenvalue decomposition (EVD) of the matrix $\mathbf{B}^H\mathbf{\Gamma}\mathbf{B}$:

$$\mathbf{B}^H\mathbf{\Gamma}\mathbf{B} = \mathbf{Q}\mathbf{D}\mathbf{Q}^H \tag{43}$$

from which it can be deduced that

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{Q}^H(\mathbf{B}^H\mathbf{\Gamma}\mathbf{B})\mathbf{Q}\mathbf{D}^{-\frac{1}{2}} = \mathbf{R}^H(\mathbf{B}^H\mathbf{\Gamma}\mathbf{B})\mathbf{R} = \mathbf{I} \tag{44}$$

where $\mathbf{R} = \mathbf{Q}\mathbf{D}^{-\frac{1}{2}}$ is an invertible matrix. Now, the matrix inversion in (23), $(\mathbf{B}^H\mathbf{\Phi}\mathbf{B})^{-1}$, can be expressed as:

$$\begin{aligned}(\mathbf{B}^H\mathbf{\Phi}\mathbf{B})^{-1} &= \mathbf{R}\mathbf{R}^{-1}(\mathbf{B}^H\mathbf{\Phi}\mathbf{B})^{-1}\mathbf{R}^{-H}\mathbf{R}^H \\ &= \mathbf{R}\left(\mathbf{R}^H\mathbf{B}^H\mathbf{\Phi}\mathbf{B}\mathbf{R}\right)^{-1}\mathbf{R}^H \\ &= \mathbf{R}\left(\mathbf{R}^H\mathbf{B}^H(\phi_R\mathbf{\Gamma}+\mathbf{\Phi_v})\mathbf{B}\mathbf{R}\right)^{-1}\mathbf{R}^H \end{aligned} \tag{45}$$

where the last transition is due to (31). Using (44), we obtain

$$(\mathbf{B}^H\mathbf{\Phi}\mathbf{B})^{-1} = \mathbf{R}\left(\phi_R\mathbf{I} + \mathbf{R}^H\mathbf{B}^H\mathbf{\Phi_v}\mathbf{B}\mathbf{R}\right)^{-1}\mathbf{R}^H. \tag{46}$$

Algorithm 1: RETF estimation.

```
for $i = 1, \dots, N$ do
    Estimate $\widehat{\phi}_{R,i}(m,k)$ using (28).
    Calculate $\gamma_i(m,k)$ using (36) and $H_{\text{e},i}(m,k)$
    using (35).
    Estimate $\widehat{X}_{\text{e},i}(m,k)$ using (34).
    Estimate $\widehat{\phi}_{X_{\text{e},1},X_{\text{e},1}}(m,k), \widehat{\phi}_{X_{\text{e},i},X_{\text{e},1}}(m,k)$ using (40)
    and (41).
    Estimate $\widehat{\widehat{G}}_{\text{e},i}(k)$ using (39).
end
```

We can now further decompose $\mathbf{R}^H\mathbf{B}^H\mathbf{\Phi_v}\mathbf{B}\mathbf{R}$ using EVD once again:

$$\mathbf{R}^H\mathbf{B}^H\mathbf{\Phi_v}\mathbf{B}\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H. \tag{47}$$

By substituting (47) in (46) and by using the orthonormality of $\mathbf{V}$, namely $\mathbf{V}\mathbf{V}^H = \mathbf{I}$, the desired matrix inversion can be rewritten as

$$\begin{aligned}(\mathbf{B}^H\mathbf{\Phi}\mathbf{B})^{-1} &= \mathbf{R}\left(\phi_R\mathbf{V}\mathbf{V}^H + \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H\right)^{-1}\mathbf{R}^H \\ &= \mathbf{R}\mathbf{V}\left(\phi_R\mathbf{I}+\mathbf{\Lambda}\right)^{-1}(\mathbf{R}\mathbf{V})^H. \end{aligned} \tag{48}$$

Finally, substituting $(\mathbf{B}^H\mathbf{\Phi}\mathbf{B})^{-1}$ in (23), $\mathbf{h}_{\text{NC}}$ can be expressed as:

$$\mathbf{h}_{\text{NC}} = \mathbf{R}\mathbf{V}\left(\phi_R\mathbf{I}+\mathbf{\Lambda}\right)^{-1}(\mathbf{R}\mathbf{V})^H\mathbf{B}^H\mathbf{\Phi}\mathbf{h}_0. \tag{49}$$

The matrix $\mathbf{R}$ depends on the spatial sound field of the reverberant signal, namely the RETF and the diffused sound coherence function, and is hence time-invariant in static scenarios. Assuming the ambient noise is stationary (or at least slowly time-varying), the matrices $\mathbf{V}$ and $\mathbf{\Lambda}$ are also time-invariant. Comparing (23) and (49) we conclude that the $(M-1)\times(M-1)$ matrix inversion ($\mathcal{O}(M^3)$ operations) is substituted by a simpler procedure, involving the multiplication of the $(M-1)\times M$ matrix $\mathbf{R}\,\mathbf{V}$ (that can be calculated in advance) with the inverse of a *diagonal* matrix $\phi_R\mathbf{I}+\mathbf{\Lambda}$ and then with the matrix $(\mathbf{R}\,\mathbf{V})^H$. The latter procedure requires only $\mathcal{O}(M^2)$ operations, implying a significant reduction in the computational burden, especially for large number of microphones.

The overall dereverberation and noise reduction algorithm is summarized in Algorithm 2 and its block diagram is depicted in Fig. 2.

VI. Performance Evaluation

The performance of the proposed algorithm is evaluated in terms of two objectives measures that are commonly used in the speech enhancement community, namely perceptual

[2] The time and frequency indices are omitted for brevity.

Algorithm 2: Multi-microphone speech dereverberation and noise reduction.

```
for all time frames $m$ do
    Estimate $\widehat{\phi}_R(m,k)$ using (30).
    Calculate the MVDR
    $\mathbf{h}_{\text{MVDR}}(m,k) = \mathbf{h}_0(k) - \mathbf{B}(m,k)\,\mathbf{h}_{\text{NC}}(m,k)$ with:
    if SNR is low then
        Calculate $\mathbf{h}_{\text{NC}}(m,k)$ using (49)
    else
        Calculate $\mathbf{h}_{\text{NC}}(k)$ using (42)
    end
    Calculate $\gamma(m,k)$ using (26) and $\xi(m,k)$ using (25).
    Calculate the PF $H_{\text{W}}(m,k)$ using (24).
    Calculate the MMSE estimate
    $\widehat{S}_F(m,k) = H_{\text{W}}(m,k)\,\mathbf{h}^H_{\text{MVDR}}(m,k)\,\mathbf{y}(m,k)$.
end
```

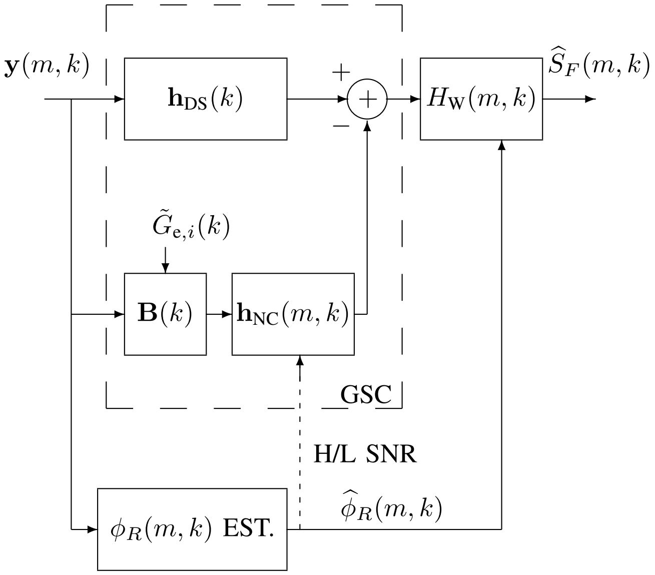


Fig. 2. Block diagram of the proposed algorithm.

evaluation of speech quality (PESQ) [52] and log-spectral distance (LSD), and the word accuracy of an automatic speech recognition (ASR) system. The following scenarios were considered: 1) simulated reverberant signals without additive noise; 2) simulated reverberant signals with spatially white Gaussian noise; 3) recorded reverberant signals (high SNR); 4) recorded reverberant signals plus recorded air-condition noise.

A. Setup

In all our experiments, a loudspeaker (Fostex 6301BX) was positioned in various distances (i.e., 0.15, 1, 2, 3, 4 m) in front of a four microphone array, such that no delay compensation is required in the FBF. The FBF was therefore set to $\mathbf{h}_0 = \mathbf{h}_{\text{d}} = \frac{1}{N}\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$. The inter-distances between the microphones were $[3,\ 4,\ 3]$ cm. An illustration of the setup with the various loudspeaker positions is depicted in Fig. 3.

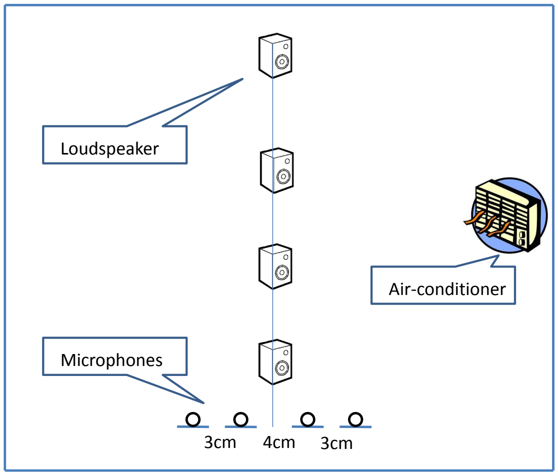


Fig. 3. Illustration of microphones and speaker setup.

The sampling frequency was 16 kHz, the frame length of the STFT was 32 ms with 8 ms between successive time frame (i.e., 25 % overlap). For measuring the speech quality we set the following values for the parameters. The forgetting factor β_{e} for the PSD estimation was set to 0.7 and the weighting factor β_{r} for the decision directed *a priori* SRNR estimator in (25) was set to 0.9. The lower bounds $H_{\min,R}$ and $H_{\min,V}$ were set to -18 dB and -15 dB, respectively. We assumed that the late reverberation starts 32 ms after the arrival of the direct-path by using $L = 4$.

An ASR system, known as PocketSphinx [53], was used with 39 Mel frequency cepstral coefficient (MFCC) features including delta and delta-delta features. The acoustic model consisted of a hidden Markov model with 5000 states. Each state observation was modeled with a Gaussian mixture model with 16 mixture components. The acoustic model was trained using the Wall Street Journal (WSJ) database [54] and recorded speech with a source-microphone distance of 15 cm (i.e., only close talking). Finally, the 20,000-word vocabulary language model was trained using WSJ as well. For testing the ASR performance we have used a database comprising 5 female and 5 male speakers, each uttering approximately 150 English sentences. The utterances were provided by Samsung

Electronics and were taken from different speech databases[3] (samples from the WSJ were not included). Overall, the test database consists of 1497 sentences, 2–4 sec long (2–8 words). The recorded speech was then played from the loudspeaker in our lab, as explained above.

The performance of the proposed algorithm was compared with the following algorithms: 1) single-channel derverberation algorithm based on spectral substraction and Polack's model [13], [14]; 2) the proposed NO-GSC algorithm without the PF; 3) the proposed NO-GSC algorithm (with the same PF) with some modifications. The noise matrix Φ in (23) was substituted by the spatial coherence matrix in (32) with diagonal loading. The role of the diagonal loading is to reduce the sensitivity of the BF by limiting its white noise gain. In the noisy case a value of 10^{-3} was used while in the noiseless case a value of 10^{-5} was chosen. Due to this modification the matrix Φ becomes time-invariant. Similarly to the proposed method, the PF was calculated at the output of the BF. The GSC was designed to satisfy a constraint on the direct component of the acoustic path rather than a constraint on the RETF as in the proposed method. Henceforth this reference algorithm will be referred to as SDBF with PF.

VII. Performance Measures

The speech quality was evaluated by computing the PESQ score and LSD. Both the PESQ and the LSD were measured by comparing $\widehat{S}_F(m,k)$ with $S_F(m,k)$, where $S_F(m,k)$ was obtained by filtering the anechoic speech $S(m,k)$ with the average early transfer function $1/N\sum_{i=1}^{N} G_{\mathrm{e},i}(k)$. For the simulated reverberant signals, the first 32 ms (measured from the arrival time of the direct-path) of the AIRs were convolved with the anechoic signal to create the reference signal. For the recorded signals scenario, the early transfer functions are not available, and are therefore first identified using a supervised system identification method. The average LSD between $\widehat{S}_F(m,k)$ and $S_F(m,k)$ is given by

$$\mathrm{LSD} = \frac{1}{M}\sum_m \sqrt{\frac{1}{K}\sum_k \left[20\log_{10}\left(\frac{\max\{|S_F(m,k)|,\epsilon\}}{\max\{|\widehat{S}_F(m,k)|,\hat{\epsilon}\}}\right)\right]^2} \tag{50}$$

where

$$\begin{aligned}\epsilon &= 10^{-A_{\mathrm{dB}}/10}\max_{m,k}\{|S_F(m,k)|\}\\ \hat{\epsilon} &= 10^{-A_{\mathrm{dB}}/10}\max_{m,k}\{|\widehat{S}_F(m,k)|\}.\end{aligned} \tag{51}$$

The parameter A_{dB} is set to the desired dynamic range, which in our case is set to 60 dB. For comparison, we also evaluated the performance of the single-channel dereverberation algorithm proposed in [14]. The PESQ scores and LSD measure were computed by averaging the results obtained using 295 sentences, 146 uttered by female and 149 by male speakers, drawn from the same database, used for the ASR experiments.

A. Simulated Data

The AIRs were computed using an efficient implementation of the image method [55], [56]. Room dimensions were set to $[6.1,\ 5.3,\ 2.7]$ m and the reverberation time was set to $T_{60} = 0.5$ s. Sampling rate for simulating the AIRs was set to 16 kHz. Finally, the AIRs were truncated to $12\cdot 10^3$ coefficients. Four source-microphone distances were tested. The obtained PESQ scores, LSD, and word accuracy ($\mathrm{W}_{\mathrm{Acc}}$) are summarized in Table I. The best results are highlighted in boldface. The results show that for all methods the processed signals attain a higher PESQ score, lower LSD, and higher word accuracy compared with the unprocessed signal. The proposed NO-GSC multichannel dereverberation algorithm *without* the PF exhibits inferior performance measures, emphasizing the importance of the single-channel postfiltering stage. The proposed method with PF is slightly superior to the SDBF with PF, in all cases but for the 1 m distance. These results can be interpreted as follows. In the noiseless case the interference PSD matrix, used in the BF design, is equivalent in both methods. Moreover, in the 1 m case, the difference between the direct-path and the early reflections is not pronounced due to high DRR, therefore resulting in an advantage for the simpler, direct-path only model. In the following experiment, spatially-white noise was added to the simulated reverberant signals to obtain various SNR levels. The noise PSD, $\phi_V(m,k)$, was estimated from one of the microphones during speech absence (assuming an ideal voice activity detector). Finally, the spatial noise PSD matrix was set to $\mathbf{\Phi}_{\mathbf{v}}(m,k) = \phi_v(m,k)\,\mathbf{I}$, where $\mathbf{I}$ is an $M\times M$ identity matrix. In Table II the results for several SNR levels and a speaker-array distance of 3 m are depicted. We observe that the performance gain between the unprocessed and processed signals monotonically increases with decreasing SNR. More importantly, the proposed multichannel algorithm evidently outperforms the SDBF with PF. This can be attributed to the better modelling of the interference PSD matrix, taking into account the time-varying nature of the reverberation-to-noise ratio.

[3] Some examples are given at http://www.eng.biu.ac.il/gannot/speech-enhancement.

PESQ	15cm	1m	2m	3m	4m
Unprocessed	3.29	2.34	2.24	2.05	2.1
Single-channel derev.		2.67	2.59	2.30	2.39
Proposed NO-GSC w.o. PF		2.54	2.48	2.45	2.43
SDBF w. PF		**2.84**	2.66	2.4	2.51
Proposed NO-GSC w. PF		2.82	**2.83**	**2.64**	**2.69**

LSD	15cm	1m	2m	3m	4m
Unprocessed	1.77	4.26	4.21	4.57	4.89
Single-channel derev.		3.48	3.59	3.67	4.04
Proposed NO-GSC w.o. PF		3.45	3.57	3.58	4.04
SDBF w. PF		**3.17**	**3.39**	3.47	**3.84**
Proposed NO-GSC w. PF		3.24	3.45	**3.40**	**3.84**

$\mathrm{W}_{\mathrm{Acc}}$	15cm	1m	2m	3m	4m
Unprocessed	78	29.9	31.5	22.3	21.6
Single-channel derev.		54.8	54.8	42.6	41.3
Proposed NO-GSC w.o. PF		43.3	46	40.4	37.7
SDBF w. PF		**66.3**	64.3	59.3	57.8
Proposed NO-GSC w. PF		65.2	**65.4**	**60.9**	**59.4**

TABLE I
SIMULATED ENVIRONMENT (NOISELESS). FOR CLEAN UTTERANCES $\mathrm{W}_{\mathrm{ACC}}$=84.2%.

PESQ	10 dB	20 dB	30 dB
Unprocessed	1.535	1.88	2.01
Single-channel derev.	1.85	2.17	2.26
Proposed NO-GSC w.o. PF	1.96	2.22	2.32
SDBF w. PF	1.68	2.21	2.40
Proposed NO-GSC w. PF	**2.25**	**2.46**	**2.55**

LSD	10 dB	20 dB	30 dB
Unprocessed	16.34	9.4	5.23
Single-channel derev.	8.85	4.87	3.82
Proposed NO-GSC w.o. PF	8.61	4.47	3.83
SDBF w. PF	7.79	4.62	3.54
Proposed NO-GSC w. PF	**4.81**	**3.64**	**3.35**

$\mathrm{W}_{\mathrm{Acc}}$	10 dB	20 dB	30 dB
Unprocessed	2.8	15.5	23.5
Single-channel derev.	4.1	33.1	42.9
Proposed NO-GSC w.o. PF	7.4	26.1	36.6
SDBF w. PF	10.3	21.9	56.2
Proposed NO-GSC w. PF	**20.9**	**48.5**	**56.5**

TABLE II
SIMULATED REVERBERANT SIGNALS PLUS SPATIALLY-WHITE NOISE FOR A SPEAKER-ARRAY DISTANCE OF 3 M.

PESQ	15cm	1m	2m	3m	4m
Unprocessed	3.68	2.74	2.35	2.26	2.27
Single-channel derev.		3.16	2.64	2.6	2.54
Proposed NO-GSC w.o. PF		2.82	2.44	2.37	2.38
SDBF w. PF		3.24	2.67	2.61	2.51
Proposed NO-GSC w. PF		**3.27**	**2.75**	**2.73**	**2.64**

LSD	15cm	1m	2m	3m	4m
Unprocessed	1.38	2.96	3.89	4.29	4.47
Single-channel derev.		3.12	3.43	3.69	3.69
Proposed NO-GSC w.o PF		2.83	3.67	4	4.17
SDBF w. PF		2.92	3.37	3.69	3.74
Proposed NO-GSC w. PF		**2.80**	**3.19**	**3.44**	**3.56**

$\mathrm{W}_{\mathrm{Acc}}$	15cm	1m	2m	3m	4m
Unprocessed	81.63	59	36.9	31.5	31.1
Single-channel derev.		71.9	57.6	53.5	48.6
Proposed NO-GSC w.o. PF		66	45.5	39.7	38.9
SDBF w. PF		76.3	60.8	57.3	50.3
Proposed NO-GSC w. PF		**76.7**	**64.2**	**61.4**	**56.6**

TABLE III
RECORDED REVERBERANT SIGNALS WITH HIGH SNR. FOR CLEAN UTTERANCES$\mathrm{W}_{\mathrm{ACC}}$=84.2%.

B. Recorded Data

For the following experiment reverberant signals with and without air-conditioning noise were recorded in the var-echoic acoustic laboratory at Bar-Ilan University, Israel. The speech utterances were played in the room using a Fostex 6301BX loudspeaker and were recorded by four AKG CK32 omnidirectional microphones, mounted on a metal ruler. The room dimensions are $[6,\ 6,\ 2.4]$ m. Reverberation time was set by adjusting the room panels, and was measured to be approximately $T_{60} = 0.5$ s.

The results without the air-conditioning are summarized in Table III. The average SNR in this case was approximately 40 dB. We observe that the PESQ and LSD scores as well as the ASR results for the recorded reverberant signals are better compared to the results in simulated environment, presented above. Moreover, for all performance measures the proposed multichannel algorithm outperforms the competing algorithms. The performance gain, also for the 1 m case, can be attributed to the small noise level in the real recordings.

For the last experiment, real air-conditioner noise was recorded and added to the recorded reverberant speech signals with several SNR levels. The spatial PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(m,k)$, which in this case is non-diagonal, was estimated using inactive speech periods. The results of this experiment in various noise levels are given in Table IV. Although lower performance measures are demonstrated in comparison with

PESQ	10 dB	20 dB	30 dB
Reverberant, noisy signals	1.75	2.10	2.22
Single-channel derev.	2.07	2.41	2.55
Proposed NO-GSC w.o. PF	1.97	2.25	2.34
SDBF w. PF	2.17	2.47	2.57
Proposed NO-GSC w. PF	**2.25**	**2.59**	**2.71**

LSD	10 dB	20 dB	30 dB
Unprocessed	9.52	6.31	4.78
Single-channel derev.	5.92	4.20	3.70
Proposed NO-GSC w.o. PF	7.14	5.14	4.28
SDBF w. PF	5.55	4.34	3.92
Proposed NO-GSC w. PF	**5.40**	**4.00**	**3.50**

$\mathrm{W}_{\mathrm{Acc}}$	10 dB	20 dB	30 dB
Reverberant, noisy signals	22.9	36.7	35
Single-channel derev.	40.3	56.4	59.2
Proposed NO-GSC w.o. PF	22.3	35.3	36.4
SDBF w. PF	40	56	58.1
Proposed NO-GSC w. PF	**45.8**	**59.7**	**61.7**

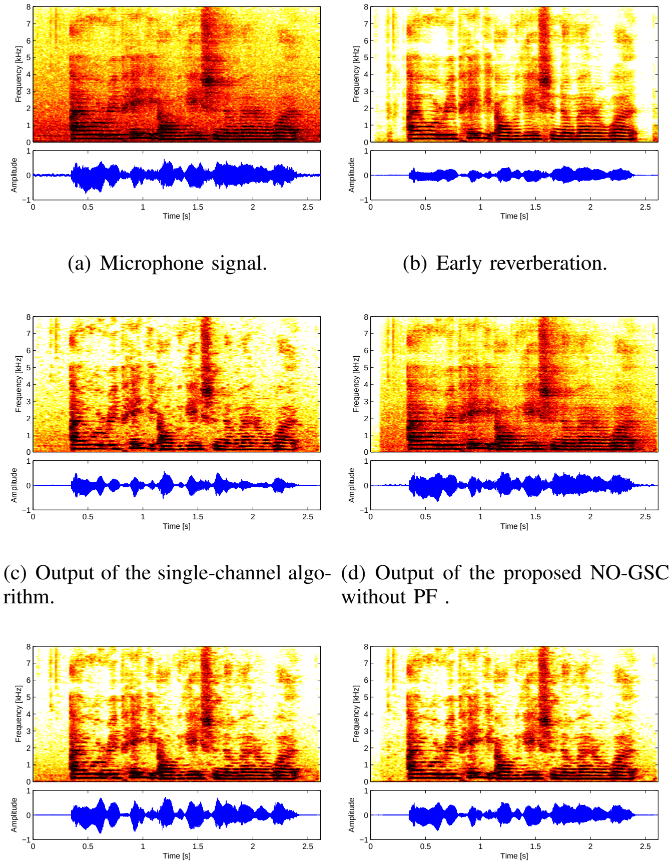
TABLE IV
RECORDED REVERBERANT SIGNALS AT A SOURCE-ARRAY DISTANCE OF 3 M AND WITH ADDITIVE AIR-CONDITIONING NOISE.

the high SNR case, the difference between the algorithms is emphasized. Again, this performance advantage can be attributed to the better modelling of both the early reflections and the time-varying sPSD matrix of the interference signals (noise and reverberation). The performance gain of the proposed algorithm with respect to the SDBF with PF is more pronounced in the lower SNR values. The significant contribution of the postfiltering stage is also evident.

The results for recorded signals with moderate noise level can also be verified by assessing the speech sonograms as depicted in Fig. 4. It can be clearly deduced that both the proposed algorithm and the SDBF with PF indeed reduces both noise and reverberation while maintaining low distortion and significantly outperforms the single-channel algorithm[4]. Informal listening tests verify that the proposed method outperforms the SDBF with PF.

VIII. CONCLUSIONS

In this contribution we have derived a multi-microphone MMSE estimator implemented as an MVDR BF followed by a PF. The aim was to obtain an estimate of a spatially filtered version of the early speech component. The MVDR BF was implemented in a NO-GSC structure. We chose the DS BF as the FBF block in order to reduce early reflections.



(a) Microphone signal.
(b) Early reverberation.
(c) Output of the single-channel algorithm.
(d) Output of the proposed NO-GSC without PF .
(e) Output of the SDBF with PF.
(f) Output proposed NO-GSC with PF.

Fig. 4. Sonograms of a real recording with a segmental SNR of 20 dB and source-array distance of 3 m.

An identification procedure for the RETFs was proposed and used to block the early speech components at the output of the BM. The late reverberation was modeled as a diffuse sound field, while the reverberation level was estimated by the average of the marginal reverberation levels at the microphones. We also derived an expression for the NC that requires less calculations per frame when the SNR is slowly time-varying. The presented experimental study consists of both simulated and recorded signals. The algorithm was tested in a room with a reverberation time of 0.5 s for various source-array distances (1–4 m), and for several signal-to-noise levels and compared with various competing algorithms. In terms of objective quality measures and ASR performance the proposed algorithm significantly outperforms: 1) a baseline single-channel algorithm; 2) the same NO-GSC without the PF; 4) and a simpler combination of SDBF and a PF. By using these algorithms we would like to evaluate the role of the PF and its combination with the BF, and the advantage of using

[4] Sound examples are available at http://www.eng.biu.ac.il/gannot/speech-enhancement.

a more complex spatial correlation matrix for the interference signal and of considering early speech reflections rather than only the direct-path in the MVDR design.

IX. Acknowledgment

This work was supported by Samsung Electronics. Special thanks to Dr. Namgook Cho for his valuable comments and ideas.

References

[1] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. Thesis, Technische Universiteit Eindhoven, Jun. 2007.

[2] M. Triki and D. T. Slock, "Iterated delay and predict equalization for blind speech dereverberation," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2006.

[3] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 430–440, 2007.

[4] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 392–396, 1996.

[5] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2003, pp. 92–95.

[6] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, pp. 1074–1090, 2003.

[7] M. Miyoshi and Y. Kenda, "Inverse filtering of room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[8] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping - a study," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[9] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 249–259, Feb. 2010.

[10] W. Zhang, E. A. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2010.

[11] I. Kodrasi and S. Doclo, "Robust partial multichannel equalization techniques for speech dereverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 537–540.

[12] J. D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, France, 1988.

[13] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, pp. 359–366, 2001.

[14] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, 2004, pp. 250–254.

[15] ——, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Mar. 2005, pp. 173–176.

[16] E. A. P. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation in a noisy environment," in *IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 651–655.

[17] H. W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 1, 2009.

[18] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

[19] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, 2009.

[20] M. Togami and Y. Kawaguchi, "Noise robust speech dereverberation with Kalman smoother," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7447–7451.

[21] L. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

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

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

[24] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Sep. 2009.

[25] ——, "Multichannel speech enhancement using convolutive transfer function approximation in reverberant environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3885–3888.

[26] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, The Netherlands, Jun. 2007.

[27] E. A. P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. IEEE, 2007, pp. IV–901.

[28] E. A. P. Habets, "Towards multi-microphone speech dereverberation using spectral enhancement and statistical reverberation models," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, 2008, pp. 806–810.

[29] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *Reverb Challenge*. Florence, Italy: IEEE Audio and Acoustic Signal Processing TC, May 2014.

[30] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 945–958, 2013.

[31] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," *Microphone Arrays: Signal Processing Techniques and Applications*, vol. 3, pp. 39–60, 2001.

[32] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *IEEE Sensor Array and Multichannel Signal Processing Workshop*. IEEE, 2002, pp. 209–213.

[33] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Proc. Mag.*, pp. 4–24, Apr. 1988.

[34] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 35, no. 10, pp. 1365–1376, 1987.

[35] A. T. Parsons, "Maximum directivity proof for three-dimensional arrays," *J. Acoust. Soc. Am.*, vol. 82, p. 179, 1987.

[36] J. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[37] H. Kuttruff, *Room acoustics*. Taylor and Francis, 2000.

[38] K. Lebart, J.-M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[39] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," *Technical University of Denmark*, pp. 7–15, 2008.

[40] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 158–170, 2010.

[41] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

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

[43] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[44] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, 2009.

[45] ——, "Speech dereverberation using backward estimation of the late reverberant spectral variance," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, 2008, pp. 384–388.

[46] 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.

[47] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, pp. 3464–3470, Dec. 2007.

[48] E. A. P. Habets, "A distortionless subband beamformer for noise reduction in reverberant environments," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel-Aviv, Israel, Aug. 2010.

[49] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011.

[50] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012.

[51] R. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[52] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[53] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[54] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[55] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[56] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Am.*, vol. 76, no. 5, pp. 1527–1529, Nov. 1986.

Ofer Schwartz Ofer Schwartz received his BSc (Summa Cum Laude) and MSc degrees in Electrical Engineering from Bar-Ilan University, Israel in 2010 and 2013, respectively. He is currently a PhD student at the Speech and Signal Processing laboratory of the Faculty of Engineering at Bar-Ilan. His research interests include statistical signal processing and in particular noise reduction and dereverberation using microphone arrays and speaker localization and tracking.

Sharon Gannot (S'92-M'01-SM'06) received his B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in electrical engineering. In 2001 he held a post-doctoral position at the department of Electrical Engineering (ESAT-SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is an Associate Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory. Prof. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for 2010 and 2014.

Prof. Gannot has served as an Associate Editor of the EURASIP Journal of Advances in Signal Processing in 2003-2012, and as an Editor of two special issues on Multi-microphone Speech Processing of the same journal. He has also served as a guest editor of ELSEVIER Speech Communication and Signal Processing journals. Prof. Gannot has served as an Associate Editor of IEEE Transactions on Speech, Audio and Language Processing in 2009-2013. Currently, he is a Senior Area Chair of the same journal. He also serves as a reviewer of many IEEE journals and conferences. Prof. Gannot is a member of the Audio and Acoustic Signal Processing (AASP) technical committee of the IEEE since Jan., 2010. He is also a member of the Technical and Steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the general co-chair of IWAENC held at Tel-Aviv, Israel in August 2010. Prof. Gannot has served as the general co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. Prof. Gannot was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013 and EUSIPCO 2013. Prof. Gannot research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation; dereverberation; single microphone speech enhancement and speaker localization and tracking.

Emanuël A.P. Habets (S'02-M'07-SM'11) is an Associate Professor at the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg and Fraunhofer IIS), and Head of the Spatial Audio Research Group at Fraunhofer IIS, Germany. He received the B.Sc. degree in electrical engineering from the Hogeschool Limburg, The Netherlands, in 1999, and the M.Sc. and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, The Netherlands, in 2002 and 2007, respectively.

From 2007 until 2009, he was a Postdoctoral Fellow at the Technion - Israel Institute of Technology and at the Bar-Ilan University, Israel. From 2009 until 2010, he was a Research Fellow in the Communication and Signal Processing Group at Imperial College London, U.K.

His research activities center around audio and acoustic signal processing, and include spatial audio signal processing, spatial sound recording and reproduction, speech enhancement (dereverberation, noise reduction, echo reduction), and sound localization and tracking.

Dr. Habets was a member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC) in Eindhoven, The Netherlands, a general co-chair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in New Paltz, New York, and general co-chair of the 2014 International Conference on Spatial Audio (ICSA) in Erlangen, Germany. He is a Senior Member of the IEEE, a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2011-2016) and a member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013-2015). Currently, he is an Associate Editor of the IEEE Signal Processing Letters, and a Guest Editor for the IEEE Journal of Selected Topics in Signal Processing.