SPEECH DEREVERBERATION BY DATA-DEPENDENT BEAMFORMING WITH SIGNAL PRE-WHITENING

T. Dietzen[*‡], *N. Huleihel*[*‡], *A. Spriet*[*], *W. Tirry*[*], *S. Doclo*[†], *M. Moonen*[‡], *T. van Waterschoot*[‡]

[*] NXP Software, Leuven, Belgium
[†] University of Oldenburg, Dept. of Medical Physics and Acoustics, Oldenburg, Germany
[‡] KU Leuven, Dept. of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium

ABSTRACT

Among different microphone array processing techniques, data-dependent beamforming has been proven to be effective in suppressing ambient noise. When applied for dereverberation, however, the adaptation process results in a biased estimate of the beamformer coefficients leading to strong distortions at the beamformer output. In this paper, we investigate the origin of this bias for the generalized sidelobe canceller. It is shown that an unbiased estimate of the beamformer coefficients and thus dereverberation can be achieved if the source signal is a white random signal. Based on these findings, a pre-whitening approach for speech signals is proposed and combined with a generalized sidelobe canceller for speech dereverberation. The concept is demonstrated for the case of stationary speech-shaped noise as a source signal.

Index Terms— Dereverberation, beamforming, generalized sidelobe canceller, estimation bias, whitening

1. INTRODUCTION

Microphone signals recorded in a room generally do not capture the desired source signal only, but also numerous reflections from the enclosure – a phenomenon that is referred to as reverberation. While reverberation may be beneficial for musical performances, it may also reduce speech intelligibility and signal quality. Hence it is undesirable in many applications such as hands-free mobile phone communication, teleconferencing, hearing aids, and automatic speech recognition.

The need for dereverberation has led to a number of approaches, where two general classes are based on microphone arrays. Firstly, channel equalization approaches are targeted at the inversion of the room transfer function from the source to the microphones, either by prior system identification followed by MINT-based inversion [1], or by estimating the inverse directly [2]. Secondly, beamforming approaches are aimed at steering a beam into the direction of the desired source while suppressing reflections from other directions [3].

Commonly data-independent, i.e. fixed beamforming (e.g. superdirective beamforming) is used for the suppression of reverberation, while for noise reduction data-dependent beamforming often performs better due to the adaptation to a time-varying noise field.

A data-dependent beamformer that has found wide usage in interference cancellation is the generalized sidelobe canceller (GSC) [4], which consists of three components: a fixed beamformer, e.g. a delay-and-sum beamformer (DSB), a blocking matrix that provides so-called noise references by blocking the desired signal, and an unconstrained adaptive filter to shape the noise references such that remaining noise in the output of the fixed beamformer is canceled. The adaptation process of the GSC however relies on the assumption that the desired signal and the noise or interference to be cancelled are statistically independent – an assumption that does not hold in the case of reverberation, where the interference stems from reflections of the desired source signal. Accordingly, the estimated filter coefficients will be biased resulting in distortion at the beamformer output, and so the classical GSC scheme is not directly suitable for dereverberation. In the scope of noise reduction, desired signal distortion due to reverberation is avoided through an improved design of the blocking matrix, e.g. by using a transfer function model instead of a delay model [5].

In the context of dereverberation, data-dependent beamforming has mainly been used in combination with speech enhancement to provide an estimate of the reverberant signal energy [6,7]. In [6], the GSC is used in a spectral-subtraction-based method to estimate the reverberant signal energy after the DSB from the output of the blocking matrix. In [7], the blocking matrix is designed to additionally block early reflections, serving spectral enhancement of the microphone signals under the assumption that late reverberation can be modeled as diffuse noise.

This research work was carried out in the frame of KU Leuven Research Council CoE PFV/10/002 (OPTEC), KU Leuven Impulse Fund IMP/14/037, and the FP7-PEOPLE Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)", funded by the European Commission under Grant Agreement no. 316969. The scientific responsibility is assumed by its authors.

In this paper we analyze the bias induced by reverberation in the estimate of the filter coefficients of the GSC found by the Wiener solution. Assuming ideal conditions, i.e. perfect steering of the DSB and perfect blocking of the desired signal in the blocking matrix, it is shown that the bias strongly depends on the autocorrelation of the source signal, implying that the filter estimate is unbiased if the source signal is a white random signal. A combination of signal pre-whitening and GSC is hence proposed and its performance evaluated for the case of stationary speech-shaped noise as a source signal.

The outline of the paper is as follows. In section 2, the behavior of the GSC when applied for dereverberation with a focus on the estimation bias is studied, leading to the pre-whitening approach. In section 3, the results from section 2 are verified for stationary speech-shaped noise as a source signal. Section 4 summarizes the paper.

2. GSC APPLIED FOR DEREVERBERATION

2.1. Signal Model

Consider the GSC in a reverberant but noise-free environment, as shown in Fig. 1. Assuming M microphones, let $h_m(n)$ be the room impulse response (RIR) of length L at sample n between the source and the mth microphone,

$$h_m(n) = \vec{h}_m(n) + \check{h}_m(n). \tag{1}$$

In (1) and in the following, $\vec{(\cdot)}$ and $\check{(\cdot)}$ denote the direct and the reverberant component, respectively, where the direct component of the RIR is defined as the first non-zero sample. Let $s(n)$ denote the source signal at time instant n. For simplicity, we assume far-field propagation and the source to be positioned in the broadside direction of the microphone array. Then the following relation holds,

$$\sum_{m=0}^{M-1} \vec{h}_m(n) \sim \delta(n - n_0), \tag{2}$$

where n_0 and $(\sim)$ denote the arrival time of the direct component and proportionality, respectively. The DSB hence simplifies to a summation of the M microphone signals $y_m(n)$ with the output given as,

$$\begin{aligned} q(n) &= \sum_{m=0}^{M-1} y_m(n) && \text{(3a)} \\ &= \sum_{m=0}^{M-1} \Big(\vec{h}_m(n) + \check{h}_m(n)\Big) * s(n) && \text{(3b)} \\ &= \vec{q}(n) + \check{q}(n), && \text{(3c)} \end{aligned}$$

where $(*)$ denotes the convolution operation. From (2) and (3) we find

$$\vec{q}(n) \sim s(n - n_0). \tag{4}$$

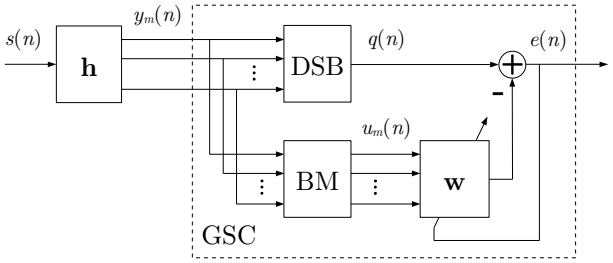


Fig. 1. The GSC in a reverberant environment.

In the remainder of this paper, we will call $\vec{q}(n)$ the desired signal. We estimate it by applying a filter of length L_w to the blocking matrix outputs $u_m(n)$, which serve as reverberation references. We construct the references $u_m(n)$ with $m = 1, 2, \ldots, M-1$, from the microphone signals based on the Griffiths-Jim blocking matrix,

$$\begin{aligned} u_m(n) &= y_0(n) - y_m(n) && \text{(5a)} \\ &= s(n) * h_{\text{ref},m}(n) && \text{(5b)} \\ &\stackrel{\ominus}{=} s(n) * \check{h}_{\text{ref},m}(n), && \text{(5c)} \end{aligned}$$

where the reference RIR $h_{\text{ref},m}$ is defined as

$$h_{\text{ref},m}(n) = h_0(n) - h_m(n). \tag{6}$$

The symbol $\ominus$ in (5c) denotes that equal gain is assumed for all microphones in the broadside direction, in which case we indeed achieve perfect blocking, i.e. the references do not contain any direct path contribution. For ease of presentation, let $\mathbf{u}(n)$ contain stacked references $u_m(n)$ for the last L_w samples and $M-1$ channels,

$$\mathbf{u}_m(n) = \big[\, u_m(n) \;\; \ldots \;\; u_m(n - L_w + 1) \,\big]^T \in \mathbb{R}^{L_w}, \tag{7}$$

$$\mathbf{u}(n) = \big[\, \mathbf{u}_1^T(n) \;\; \ldots \;\; \mathbf{u}_{M-1}^T(n) \,\big]^T \in \mathbb{R}^{(M-1)L_w}. \tag{8}$$

The error signal to be minimized is the difference between the DSB output and the output of the data-dependent filter with coefficients $\mathbf{w}$,

$$\mathbf{w}_m = \big[\, w_m(0) \;\; \ldots \;\; w_m(L_w - 1) \,\big]^T \in \mathbb{R}^{L_w}, \tag{9}$$

$$\mathbf{w} = \big[\, \mathbf{w}_1^T \;\; \ldots \;\; \mathbf{w}_{M-1}^T \,\big]^T \in \mathbb{R}^{(M-1)L_w}. \tag{10}$$

With (8) and (10) we may represent the error signal as

$$e(n) = q(n) - \mathbf{w}^T \mathbf{u}(n). \tag{11}$$

Ideally, $\mathbf{w}$ is chosen such that the error signal corresponds to the desired signal $\vec{q}(n)$.

2.2. Wiener Solution

Assuming wide-sense stationarity, let us consider the Wiener solution that minimizes the cost function

$$\mathcal{J}(\mathbf{w}) = \mathrm{E}\{\big(q(n) - \mathbf{w}^T \mathbf{u}(n)\big)^2\} \tag{12a}$$

By setting the derivative of $\mathcal{J}(\mathbf{w})$ to zero and solving for $\mathbf{w}$, we obtain the Wiener solution for the filter coefficients,

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{r}. \tag{13}$$

In (13), the terms $\mathbf{r}$ and $\mathbf{R}$ refer to the covariance vector between $q(n)$ and $\mathbf{u}(n)$, and the autocovariance matrix of $\mathbf{u}(n)$, respectively,

$$\mathbf{r} = \mathrm{E}\{q(n)\mathbf{u}(n)\}, \tag{14a}$$
$$\mathbf{R} = \mathrm{E}\{\mathbf{u}(n)\mathbf{u}^T(n)\}. \tag{14b}$$

2.3. Bias

The covariance $\mathbf{r}$ may be decomposed into

$$\mathbf{r} = \mathrm{E}\{\vec{q}(n)\mathbf{u}(n)\} + \mathrm{E}\{\check{q}(n)\mathbf{u}(n)\} \tag{15a}$$
$$= \mathbf{r}_{\vec{q}u} + \mathbf{r}_{\check{q}u}. \tag{15b}$$

The vector $\mathbf{r}_{\vec{q}u}$ in (15b) expresses the covariance between the desired signal and the references. Note that in noise reduction, this term is considered to be zero if the desired signal is perfectly blocked by the blocking matrix, since the desired signal and noise are assumed to be statistically independent. This assumption is however violated in case of reverberation, which is due to its physical nature – the reflections from the room enclosure and the desired signal are both linearly related to the source signal. The term $\mathbf{r}_{\vec{q}u}$ is then generally non-zero and introduces a dependency of the filter coefficients $\mathbf{w}$ on the desired signal, which results in a bias in the filter estimate and causes distortion in the GSC output. For the following analysis of $\mathbf{r}_{\vec{q}u}$, we draw attention to its vector elements,

$$\mathbf{r}_{\vec{q}u,m} = \begin{bmatrix} r_{\vec{q}u,m}(0) & \dots & r_{\vec{q}u,m}(-L_w+1) \end{bmatrix}^T \in \mathbb{R}^{L_w}, \tag{16}$$
$$\mathbf{r}_{\vec{q}u} = \begin{bmatrix} \mathbf{r}_{\vec{q}u,0}^T & \dots & \mathbf{r}_{\vec{q}u,M-1}^T \end{bmatrix}^T \in \mathbb{R}^{(M-1)L_w}, \tag{17}$$

with the scalar elements $r_{\vec{q}u,m}(\eta)$ given by

$$r_{\vec{q}u,m}(\eta) = \mathrm{E}\{\vec{q}(n)u_m(n+\eta)\}. \tag{18}$$

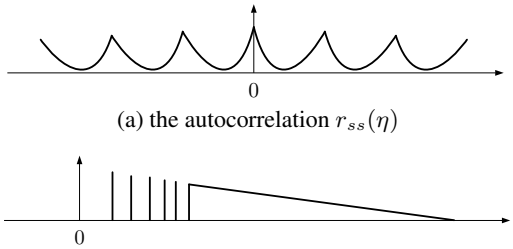
Under the assumption of ergodicity, $r_{\vec{q}u_m}(\eta)$ is proportional to the sample cross-correlation of $\vec{q}(n)$ and $u_m(n)$ at lag η,

$$r_{\vec{q}u,m}(\eta) \sim \sum_{n=-\infty}^{\infty} \vec{q}(n)u_m(n+\eta). \tag{19}$$

Let $(\star)$ denote the cross-correlation operation. Making use of the relation $(x \star y)(\eta) = x(-\eta) * y(\eta)$ and inserting (3) and (5) in (19) we easily derive

$$r_{\vec{q}u,m}(\eta) \sim \vec{q}(\eta) \star u_m(\eta) \tag{20a}$$
$$= r_{ss}(\eta) * r_{\vec{h}h_{\mathrm{ref},m}}(\eta). \tag{20b}$$

The terms $r_{ss}(\eta)$ and $r_{\vec{h}h_{\mathrm{ref},m}}(\eta)$ express the autocorrelation of the source signal and the cross-correlation between the summed direct components of the RIR and the reference RIR, respectively,



(a) the autocorrelation $r_{ss}(\eta)$

(b) the cross-correlation $r_{\vec{h}h_{\mathrm{ref},m}}(\eta)$

Fig. 2. Schematic illustration of the two correlations defining the estimation bias.

$$r_{ss}(\eta) = s(\eta) \star s(\eta), \tag{21a}$$
$$r_{\vec{h}h_{\mathrm{ref},m}}(\eta) = \sum_{m'=0}^{M-1} \vec{h}_{m'}(\eta) \star h_{\mathrm{ref},m}(\eta) \tag{21b}$$
$$\sim h_{\mathrm{ref},m}(\eta + n_0) \tag{21c}$$
$$\stackrel{\ominus}{=} \check{h}_{\mathrm{ref},m}(\eta + n_0). \tag{21d}$$

Let us consider the implications of the convolution in (20b) on the bias in the filter estimate $\mathbf{w}$, caused by the component $\mathbf{r}_{\vec{q}u}$ given in (16) and (17). In order to obtain an unbiased estimate, we require the convolution result to be zero for $-L_w + 1 \leq \eta \leq 0$. Schematic depictions of the two convolved terms are shown in Fig. 2.

For perfect blocking, the term $r_{\vec{h}h_{\mathrm{ref},m}}(\eta)$ is proportional to the reverberant component of the reference RIR shifted by n_0 samples to the left, see (21d). Since n_0 indicates the arrival time of the direct component and, consequently, the reverberant component is zero at any lag $n \leq n_0$, it is guaranteed that $r_{\vec{h}h_{\mathrm{ref},m}}(\eta)$ is zero for any $\eta \leq 0$. In contrast, the autocorrelation $r_{ss}(\eta)$ is symmetrical around $\eta = 0$, and thus the convolution of both will generally be non-zero for $\eta \leq 0$. An exception however is given if $r_{ss}(\eta)$ fulfills the condition,

$$r_{ss}(\eta) \sim \delta(\eta), \tag{22}$$

or, in other words, if the source signal $s(n)$ is a white random signal. If this condition is met, the Wiener solution will yield an unbiased estimate of the filter coefficients required for dereverberation.

2.4. Pre-Whitening

We therefore propose to pre-whiten the microphone signals $y_m(n)$ with respect to the source signal $s(n)$, i.e. to compensate for the coloration of $s(n)$ inherent in $y_m(n)$. The resulting estimate of the filter coefficients $\mathbf{w}$ can then be copied to a second GSC structure that is applied to the unaltered microphone signals, yielding an estimate $\hat{s}(n)$ of the source signal,

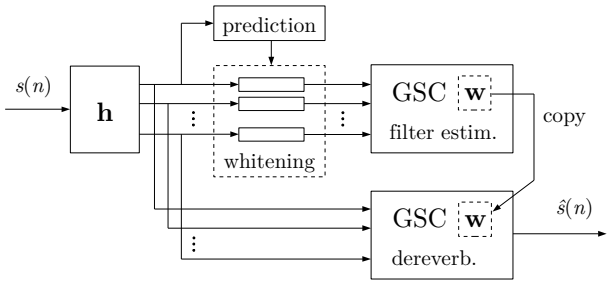


Fig. 3. The proposed pre-whitening approach.

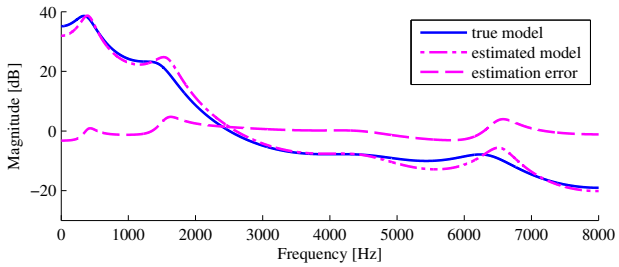


Fig. 4. The all-pole model applied to synthesize the source signal and its estimation from the microphone signal $y_0(n)$.

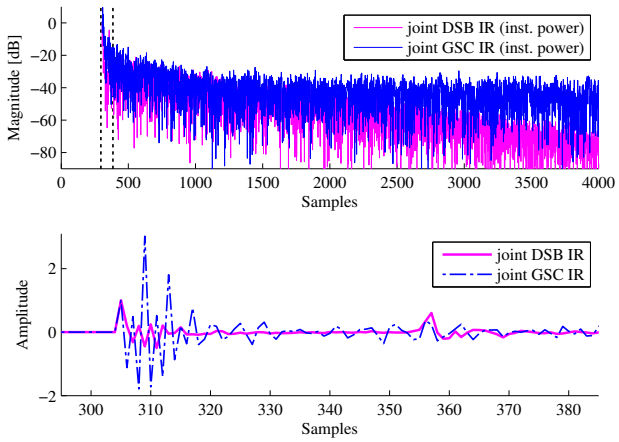


Fig. 5. The joint impulse responses after DSB and GSC if no pre-whitening is applied.

as shown in Fig. 3. In case of speech, whitening may generally be done by combining short-term and long-term linear predictive filtering so to invert the transfer function of the vocal tract and the glottal excitation, respectively.

In the following simulations however, we consider stationary speech-shaped noise as a source signal, i.e. we focus on the inversion of the vocal tract only. In order to whiten the microphone signals $y_m(n)$, we apply the predictive filter $\mathbf{a}$ with the coefficients given by

$$\mathbf{a} = \begin{bmatrix} 1 & -a_1 & \dots & -a_p \end{bmatrix}^T \in \mathbb{R}^{p+1}, \tag{23}$$

where p denotes the filter order. The coefficients are estimated from the microphone signal $y_0(n)$ by minimizing the prediction error $x(n)$ of the autoregressive model,

$$y_0(n) = \sum_{i=1}^{p} a_i y_0(n-i) + x(n). \tag{24}$$

3. SIMULATIONS

The following experiments aim at evaluating the effect of signal pre-whitening on the performance of the GSC for dereverberation. We focus on the Wiener solution for a stationary source signal as given in (13).

3.1. Experimental Setup

The RIRs are chosen from the multichannel audio database [8], downsampled to 16 kHz. Their reverberation time is 360 ms, thus resembling moderate reverberation. The source is positioned at 2 m distance in the broadside direction of the microphone array with a microphone spacing of 8 cm, from which three microphones are selected for the simulation. The maxima of the RIRs are located at the same lag and considered to be the direct components. To induce ideal conditions as assumed in section 2, we equalize the impulse responses with respect to their direct components, so to obtain perfect blocking as given in (5c). Further, to simulate a perfectly clean dead time, we consider the samples preceding the direct contribution as measurement noise and set them to zero. The impulse responses are truncated after 0.5 s (8000 samples), where they reach the measurement noise floor at about -85 dB below the level of the direct component. The filter of the GSC is chosen to have the same length as the RIRs.

We synthesize a source signal from stationary Gaussian white noise of duration 30 s, shaped by a 10th order all-pole filter resembling the vocal tract. The pre-whitening filter is chosen to have the same order and is estimated as described in (23) and (24). The magnitude of the true all-pole filter and the autoregressive approximation are shown in Fig. 4. It can be seen that the reverberation causes an estimation error of up to 4 dB around the formants.

3.2. Results

We evaluate the dereverberation performance by looking at the joint impulse response of the room and the GSC using filter coefficients found by the Wiener solution. Ideally, the joint impulse response should be proportional to $\delta(n - n_0)$.

Let us first consider the case that signal pre-whitening is not applied, i.e. the Wiener solution is computed from the unaltered microphone signals directly. In the top part of Fig. 5 the instantaneous power for the first 4000 samples of the joint

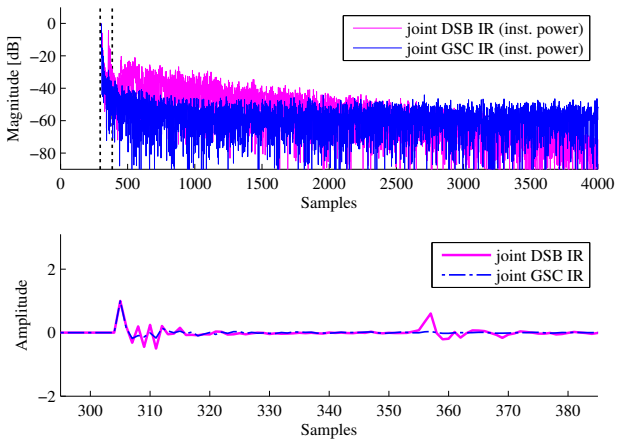


Fig. 6. The joint impulse responses after DSB and GSC if pre-whitening is applied.

impulse response after the DSB (which in our case comes down to the sum of the three RIRs), and the joint impulse response at the GSC output are shown in dB. The section of the impulse response indicated by the two vertical dotted lines covers the direct component as well as the first reflections and is shown in the bottom part of Fig. 5. The direct component given by the first non-zero sample is perfectly preserved, however it is obvious that the GSC does not remove reverberation but instead introduces strong distortions to subsequent samples. One might interpret these distortions as an attempt of the GSC to not only compensate for the RIR, but instead to compensate for the cascade of both the all-pole filter and the RIR. Seen from this perspective, it is easily understood that pre-whitening the microphone signals, i.e. filtering with the inverse of the all-pole filter, will cause the GSC to compensate for the RIR only, hence performing dereverberation.

Let us now study the outcome if signal pre-whitening is performed. Fig. 6 shows the joint impulse responses of the room and DSB as well as GSC for this case. Again the direct component is perfectly preserved, but instead of introducing additional distortion the GSC reduces reverberation. The joint GSC impulse response quickly drops to a level of about -50 dB, outperforming the joint DSB impulse response. However, note that unlike the latter one, the GSC impulse response does not further decay over time, but stays rather constant after reaching the -50 dB noise floor.

4. CONCLUSIONS AND FURTHER WORK

In the GSC, reverberation will generally lead to a biased estimate of the data-dependent filter coefficients and hence to a distorted estimate of the desired signal. Nonetheless an unbiased estimate can be obtained if the source signal is a white random signal, in which case dereverberation is achieved. Pre-whitening the microphone signals with respect to the source signal therefore bears potential to improve existing dereverberation methods based on the GSC scheme. It is worthwhile to investigate on what conditions perfect dereverberation is obtained, which is currently a topic of further research. For perfect dereverberation the GSC must behave as the exact inverse to the RIR (up to a factor and a delay), hence a relation to MINT can be assumed.

The concept of signal pre-whitening has been validated for stationary speech-shaped Gaussian noise as a source signal. However, further work is required to make the concept applicable in practice. Long-term prediction must be included in the whitening process, and the Wiener solution must be replaced by an adaptation algorithm so as to deal with time-varying scenarios. Additionally, robustness against noise and steering mismatch is required.

REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[2] H. Buchner and W. Kellermann, "Trinicon for dereverberation of speech and audio signals," in *Speech Dereverberation*, P. Naylor and N. Gaubitch, Eds., pp. 311–385. Springer, New York, 2010.

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

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

[5] D. Burshtein S. Gannot 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.

[6] E. A. P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," in *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2007)*, Honolulu, USA, Apr. 2007, vol. IV, pp. 901–904.

[7] A. Schwarz, K. Reindl, and W. Kellermann, "On blocking matrix-based dereverberation for automatic speech recognition," in *Proc. 2012 Int. Workshop Acoustic Echo Noise Control (IWAENC 2012)*, Aachen, Germany, Sept. 2012, pp. 1–4.

[8] P. Vary E. Hadad, F. Heese and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. 2014 Int. Workshop Acoustic Signal Enhancement (IWAENC 2014)*, Antibes – Juan les Pins, France, Sept. 2014, pp. 313–317.