On the Relation between Data-Dependent Beamforming and Multichannel Linear Prediction for Dereverberation

Thomas Dietzen[1,3], Ann Spriet[1], Wouter Tirry[1], Simon Doclo[2], Marc Moonen[3], and Toon van Waterschoot[3]

[1]*NXP Software, Leuven, Belgium*

[2]*University of Oldenburg, Dept. of Medical Physics and Acoustics, Oldenburg, Germany*

[3]*KU Leuven, Dept. of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium*

Correspondence should be addressed to Thomas Dietzen (thomas.dietzen@nxp.com)

ABSTRACT

The generalized sidelobe canceler, a data-dependent beamformer that is commonly used in noise suppression, is able to perform dereverberation if the source signal is a random white signal. A similar statement can be made for multichannel linear prediction, which may be used to blindly invert any time-invariant multichannel transmission system that is excited by a random white signal, provided that the transmission channels do not share common zeros. If the source signal is colored on the other hand, as it is the case for speech signals, both the generalized sidelobe canceler and multichannel linear prediction tend to additionally invert the source coloration, and different approaches have been proposed to tackle this problem. In this paper we give an overview on multichannel linear prediction methods and formally analyze the generalized sidelobe canceler for dereverberation, which reveals close relations between the two approaches.

1. INTRODUCTION

Reverberation, an acoustic phenomenon determined by the transmission channel between a source and a microphone in a room, can degrade the quality and intelligibility of speech due to the multitude of reflections from the enclosure interfering with the direct sound component. Therefore dereverberation is needed in many applications such as hands-free mobile communication, hearing aids, teleconferencing and automatic speech recognition.

Approaches based on microphone arrays take advantage of spatial diversity and, according to the **m**ultiple input/output **in**verse **t**heorem (MINT) [1], theoretically allow perfect inversion of the room transfer functions, if the different channels do not share common zeros. In practical applications however, the room transfer functions are unknown and hard to estimate, while MINT inversion has been shown to be rather sensitive to transfer function estimation errors [2]. Therefore explicit inversion is not favorable and other multichannel approaches such as multichannel linear prediction (MLP) [3–14] and, to a somewhat lesser extent, beamforming [15–19] have been proposed for dereverberation.

MLP algorithms are able to blindly invert any time-invariant multichannel transmission system that is excited by a white random signal, provided the channels do not share common zeros and the prediction filter is of sufficient length, see e.g. [5–7] and references therein. If the source signal is not white however, as it is the case for speech, then the source coloration is also inverted, a problem that is known as excessive whitening and has been tackled in different ways [3–14]. A more detailed review on MLP methods is given in section 2.

Beamforming has been used in noise reduction before it was applied to dereverberation and traditionally does not target channel inversion, but aims at steering a beam into the direction of the target source while suppressing interfering noise or reflections from other directions. One can distinguish between data-independent, i.e. fixed (e.g. superdirective) beamforming and data-dependent beam-

forming. In noise reduction the latter often performs better due to the adaptation to a time-varying noise field. In [15], the former is used in a dereverberation stage and the latter in a noise reduction stage. In [16], the so-called MINTFormer is introduced, which provides a trade-off between the performance of MINT and the robustness of beamforming in a unified framework.

The generalized sidelobe canceler (GSC) [20], a data-dependent beamformer widely employed in noise reduction, consists of three components: a fixed beamformer steering a beam into the target direction, a blocking matrix that provides so-called noise references by blocking the target signal, and an unconstrained adaptive filter shaping the noise references such that remaining noise in the fixed beamformer output is suppressed.

If however the GSC is applied to suppress reverberation instead of noise, i.e. convolutive interference instead of additive interference, then the coloration of the source signal will generally bias the filter estimate leading to distortions as shown in [17]. To circumvent this problem, the authors proposed to perform source signal pre-whitening, i.e. to remove the coloration from the source inherent in the microphone signals.

Other approaches combine blocking-matrix-based beamforming with speech enhancement to provide an estimate of the reverberant signal energy. In [18], the GSC is used in a spectral-subtraction-based method to estimate the reverberant signal energy after the delay-and-sum beamformer from the output of the blocking matrix. In [19], 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.

In fact, although both approaches evolved from very different ideas, there are close relations between MLP and data-dependent beamforming using the GSC if applied for dereverberation. The aim of this paper is to provide new insights into these relations and to derive formal equivalence conditions.

The paper is organized as follows. Section 2 gives a review on MLP, section 3 introduces the problem statement for the GSC, and section 4 points out the relation between both approaches. Section 5 concludes the paper.

2. OVERVIEW ON MLP BASED APPROACHES

MLP approaches are motivated by the observation that a multichannel transmission system may be blindly inverted by linear prediction of sufficient order if the following conditions are fulfilled [5–7]:

1. The transmission system is time-invariant or at most slowly time-varying. This is a commonly made assumption in room acoustics.
2. The transmission channels are relatively prime to each other, i.e. their transfer functions do not share common zeros. This requirement is given by the MINT theorem and is crucial for invertibility.
3. The multichannel transmission system is excited by an independently and identically distributed (iid) random sequence, i.e. by a white noise signal.

The most critical condition is the last one, as the source signals of interest like speech signals do not fulfill the iid-assumption, but may rather be modeled as the output of a time-varying speech production system which is driven by an iid-like innovations process. Hence, if the room transmission system is excited by a speech signal, then pure MLP will also invert at least the average characteristics of the speech production system – a problem that is referred to as excessive whitening [5, 6, 9, 11] or over-whitening [8]. Dereverberation approaches based on MLP may be distinguished by their strategy on how to circumvent the excessive whitening problem.

In [3], an average pre-whitening stage aiming at transforming the speech signal into an iid-like signal is used, where the average whitening filter is estimated on the same window of reverberant speech as the MLP coefficients. The correct prediction coefficients are then found by performing MLP on the whitened microphone signals. The authors further propose to align the microphone signals in order to increase prediction performance by avoiding early reflections arriving in one microphone before the direct component in another [4].

In the **li**near-predictive **m**ulti-input **e**qualization algorithm (LIME) [5–8] in contrast, MLP is performed on the microphone signals in the first step directly. In the second step, the source signal is recovered by estimating the speech production system and applying it to the MLP output. Numerical problems in LIME caused by room transfer functions with zeros in the same region are discussed in [6]. In [7], LIME is extended to additionally perform noise reduction. In [8], the LIME algorithm is reformulated in order to adapt to time-varying acoustic environments.

Unlike pre-whitening approaches and LIME, the algorithm proposed in [9] jointly estimates the inverse filters of both the room acoustics and the speech production system in an iterative, alternating manner.

Delayed linear prediction in combination with spectral subtraction has been proposed in [10], assuming that the desired signal and late reverberations are uncorrelated. This approach might be seen as conceptionally somewhat related to the spectral-subtraction-based blocking matrix approaches in [18, 19].

Probabilistic approaches [11–14] commonly model the speech signal as a time-varying Gaussian signal, and perform maximum likelihood estimation based on MLP to transform the observed reverberant speech signal into one that is probabilistically more like non-reverberant speech, without targeting exact inversion. In [11], a codebook based on short-time speech spectra was used. In [13], maximum likelihood estimation is combined with delayed linear prediction. In [14], the desired speech signal is modeled using a general sparse prior, which is interpreted as a generalization of the time-varying Gaussian model.

3. PROBLEM STATEMENT FOR THE GSC

Consider the GSC in a reverberant, but noise-free environment as shown in fig. 1, where the cascade of room and GSC is excited by a single source signal $s(n)$. Instead of separately modeling the time-invariant room impulse response (RIR), the fixed beamformer, and the blocking matrix, we make use of joint representations of the cascade of room and fixed beamformer as well as the cascade of room and blocking matrix.

3.1. Signal Model

Let $\mathbf{h}_q \in \mathbb{R}^L$ denote the zero-padded, delay-free part (i.e. excluding the dead time) of the joint impulse response of length L_{h_q} of the room and the fixed beamformer of the GSC in vector form,

$$\mathbf{h}_q = \begin{pmatrix} h_{q|0} & \cdots & h_{q|L_{h_q}-1} & 0 & \cdots & 0 \end{pmatrix}^T, \tag{1}$$

and let $\mathbf{H}_u \in \mathbb{R}^{N_u L_w \times L}$ denote a matrix stacking the N_u delay-free joint impulse responses of the room and the blocking matrix in toeplitz matrix form $\mathbf{H}_{u|i} \in \mathbb{R}^{L_w \times L}$,

$$\mathbf{H}_u = \begin{pmatrix} \mathbf{H}_{u|1}^T & \cdots & \mathbf{H}_{u|N_u}^T \end{pmatrix}^T, \tag{2a}$$

$$\mathbf{H}_{u|i} = \begin{pmatrix} h_{u|i,0} & \cdots & h_{u|i,L_{h_u}-1} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & h_{u|i,0} & \cdots & h_{u|i,L_{h_u}-1} \end{pmatrix}. \tag{2b}$$

Both $\mathbf{H}_u$ and $\mathbf{h}_q$ are considered to be unknown. L_{h_u}, N_u, and L_w respectively denote the length of the joint impulse response of the room and the blocking matrix, the number of interference references and the length of the data-dependent filter. The superscript T denotes the transpose of a matrix. The length L is defined as

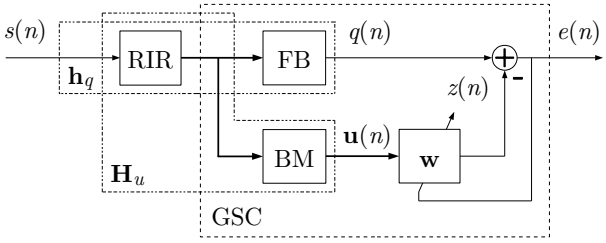


Fig. 1: The GSC in a reverberant, but noise-free environment with a single source.

$$L = L_w + L_{h_u} - 1, \tag{3}$$

and we assume here and in the following,

$$L_w \geq \frac{L_{h_u} - 1}{N_u - 1}, \tag{4}$$

such that $N_u L_w \geq L$ and the matrix $\mathbf{H}_u$ has at least as many rows as columns. Considerations on its rank follow later in this section. Further, we made the implicit assumption that $L_{h_q} \leq L$, implying a limitation of the order of the fixed beamformer depending on L_w and L_{h_u}. Apart from this restriction, we do not make any assumptions on the fixed beamformer design at this point.

With $\bar{n} = n - n_0$ and n_0 the dead time delay, let $\mathbf{s}(\bar{n}) \in \mathbb{R}^L$ be a vector stacking the latest L samples of $s(\bar{n})$,

$$\mathbf{s}(\bar{n}) = \begin{pmatrix} s(\bar{n}) & \cdots & s(\bar{n}-L+1) \end{pmatrix}^T, \tag{5}$$

The speech reference $q(n)$ can then be expressed by

$$q(n) = \mathbf{s}^T(\bar{n})\mathbf{h}_q. \tag{6}$$

Let us consider the use of the Griffiths-Jim blocking matrix to create the reverberation reference signals, where the references are constructed by subtracting the remaining microphone signals from the first microphone signal. For simplicity, let the source be in broadside direction of the microphone array and further assume equal microphone gains in the source direction as well as far-field propagation. Further, let $h_{i,j}$ denote the j^{th} sample of the RIR of the length L_h after dead time at the i^{th} microphone, with $i = 0, \ldots, M-1$ and M the number of microphones. The vector composed of the direct components $h_{i,0}$ of the individual RIRs then lies in the null

space of the blocking matrix, such that the direct component is canceled. The single entries of the joint impulse response of the room and the blocking matrix in $\mathbf{H}_u$ in (2) then take the values

$$h_{u|i,j} = h_{0,j+1} - h_{i,j+1}, \tag{7}$$

with $i = 1,\dots,N_u$ and $j = 0,\dots,L_{h_u} - 1$, where $N_u = M - 1$ and $L_{h_u} = L_h - 1$. Using this definition for $\mathbf{H}_u$, the reverberation reference signals stacked in the vector $\mathbf{u}(n) \in \mathbb{R}^{N_u L_w}$ over the latest L_w samples and N_u channels can be written as

$$\mathbf{u}(n) = \mathbf{H}_u \mathbf{s}(\bar{n} - 1). \tag{8}$$

Let us take a few considerations on the rank and the nullity of $\mathbf{H}_u$ in the Griffiths-Jim case. The joint transfer functions modeled by $\mathbf{H}_u$ are the differences between the individual room transfer functions, hence they can share a common zero only if all individual transfer functions take the same value in some point of the z-plane. This is very likely if $M = 2$ microphones are used only, in which case $N_u = 1$ reverberation reference is available only such that (4) cannot be satisfied. For values $M > 2$ however the likelihood of common zeros in the joint transfer functions drops quickly given that the individual transfer functions do not share common zeros, i.e. $\mathbf{H}_u$ will be likely to have full column rank. If $\mathbf{H}_u$ has full column rank, then the nullity of $\mathbf{H}_u$, i.e. the rank of its null space, must be zero according to the rank-nullity theorem. Hence, while the nullity of the blocking matrix itself is greater than zero per definition, this does not necessarily apply for the joint matrix $\mathbf{H}_u$.

In the following, we will generalize our analysis to arbitrary full column rank matrices $\mathbf{H}_u$ including a general reference delay d. We express the generalized reverberation reference signals as

$$\mathbf{u}(n) = \mathbf{H}_u \mathbf{s}(\bar{n} - d). \tag{9}$$

The reference delay d has to be chosen according to the definition of the desired signal, which is introduced in section 3.2.

The filter output $z(n)$ is given by

$$z(n) = \mathbf{u}^T(n)\mathbf{w} \tag{10a}$$
$$= \mathbf{s}^T(\bar{n} - d)\mathbf{H}_u^T \mathbf{w} \tag{10b}$$

with the filter coefficients $\mathbf{w} \in \mathbb{R}^{N_u L_w}$, composed of L_w coefficients per reference channel stacked over N_u channels. The filter coefficients are chosen according to the Wiener solution, which is introduced in section 3.3.

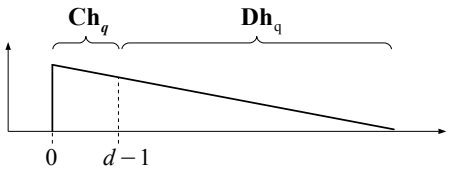


Fig. 2: Schematic depiction of the joint impulse response of room and fixed beamformer $\mathbf{h}_q$, separated in desired and undesired component $\mathbf{Ch}_q$ and $\mathbf{Dh}_q$, respectively.

3.2. Desired Signal

We can split the speech reference $q(n)$ into a desired and undesired component. Assuming that the first d samples of the impulse response $\mathbf{h}_q$ are desired, i.e. we allow early reflections up to delay d relative to the direct component, but intend to suppress later reflections, we derive

$$q(n) = \mathbf{s}^T(\bar{n})\mathbf{C}\mathbf{h}_q + \mathbf{s}^T(\bar{n} - d)\mathbf{D}\mathbf{h}_q, \tag{11}$$

with $\mathbf{s}^T(\bar{n})\mathbf{C}\mathbf{h}_q$ describing the desired component. The cutoff matrix $\mathbf{C} \in \mathbb{R}^{L \times L}$ and the delay matrix $\mathbf{D} \in \mathbb{R}^{L \times L}$ are defined as

$$\mathbf{C} = \begin{pmatrix} \mathbf{I}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_r \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{12}$$

with the identity matrices $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ and $\mathbf{I}_r \in \mathbb{R}^{(L-d)\times(L-d)}$. The cutoff matrix $\mathbf{C}$ selects the desired, first d coefficients of $\mathbf{h}_q$, while the delay matrix $\mathbf{D}$ shifts the coefficients in $\mathbf{h}_q$ upwards by d rows. The undesired and the desired component of $\mathbf{h}_q$ are illustrated schematically in fig. 2. The delay d also determines the delay to be applied in the blocking matrix, as given in (9).

3.3. Wiener Solution

The well-known Wiener solution filter coefficients that minimizes the variance of the GSC output for stationary signals is given by

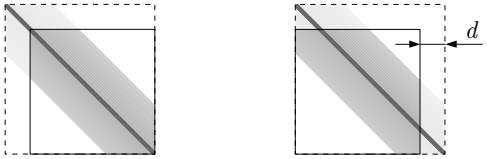
$$\mathbf{w} = \mathbf{R}_{uu}^{+} \mathbf{r}_{uq}, \tag{13}$$

where the superscript $+$ denotes the Moore-Penrose pseudoinverse. Using (6) and (9), the covariance vector $\mathbf{r}_{uq} = E\{\mathbf{u}(n)q(n)\}$ can be written as

$$\mathbf{r}_{uq} = \mathbf{H}_u E\{\mathbf{s}(\bar{n} - d)\mathbf{s}^T(\bar{n})\}\mathbf{h}_q \tag{14a}$$
$$= \mathbf{H}_u \mathbf{R}_{ss|d} \mathbf{h}_q, \tag{14b}$$

where $E\{\cdot\}$ denotes the expected value operator. Using (9) the autocovariance matrix $\mathbf{R}_{uu} = E\{\mathbf{u}(n)\mathbf{u}^T(n)\}$ can be written as

$$\mathbf{R}_{uu} = \mathbf{H}_u E\{\mathbf{s}(\bar{n} - d)\mathbf{s}^T(\bar{n} - d)\}\mathbf{H}_u^T \tag{15a}$$
$$= \mathbf{H}_u \mathbf{R}_{ss} \mathbf{H}_u^T, \tag{15b}$$



(a) the covariance $\mathbf{R}_{ss}$ (b) the covariance $\mathbf{R}_{ss|d}$

Fig. 3: Schematic illustration of the covariance matrices $\mathbf{R}_{ss}$ and $\mathbf{R}_{ss|d}$ as different sections (continuous line frame) of a larger covariance matrix (dashed line frame).

and we can express the Wiener solution filter coefficients by

$$\mathbf{w} = (\mathbf{H}_u \mathbf{R}_{ss} \mathbf{H}_u^T)^+ \mathbf{H}_u \mathbf{R}_{ss|d} \mathbf{h}_q \tag{16a}$$

$$= \mathbf{H}_u^{+T} \mathbf{R}_{ss}^{-1} \mathbf{R}_{ss|d} \mathbf{h}_q, \tag{16b}$$

where the last transition is valid under the assumption of $\mathbf{H}_u$ having full column rank and $\mathbf{R}_{ss}$ being positive-definite, which is true by construction.

The autocovariance matrices $\mathbf{R}_{ss}$ and $\mathbf{R}_{ss|d}$ have a specific relation. As shown in fig. 3, we can interpret both as different sections of a larger covariance matrix, while the section defining $\mathbf{R}_{ss|d}$ is off-diagonal and shifted by d columns leftwards as compared to the section defining $\mathbf{R}_{ss}$. Assuming that the autocovariance of $s(\bar{n})$ is zero for lags greater than $L-d$, we can express $\mathbf{R}_{ss|d}$ in terms of $\mathbf{R}_{ss}$, and $\mathbf{C}$ and $\mathbf{D}$ as given in (12) by

$$\mathbf{R}_{ss|d} = \mathbf{R}_{ss}\mathbf{D} + \mathbf{D}\mathbf{R}_{ss}\mathbf{C}. \tag{17}$$

The product $\mathbf{R}_{ss}\mathbf{D}$ shifts the coefficients in $\mathbf{R}_{ss}$ to the right by d columns, inserting zero columns on the left. The product $\mathbf{D}\mathbf{R}_{ss}\mathbf{C}$ covers for the zero columns by selecting the first d columns of $\mathbf{R}_{ss}$ and shifting them up by d rows. If we approximate the autocovariance $\mathbf{R}_{ss}$ by a time-averaging operation for non-stationary signals, then the relation in (17) only holds approximately if the autocorrelation is more or less invariant over d samples, i.e. if the delay d is reasonably small as compared to the time window on which the autocorrelation is estimated.

4. RELATION TO LINEAR PREDICTION

In the following we will derive the conditions that need to be satisfied in order to cancel the undesired component of $q(n)$. Further, we will study the behavior of the output signal of the GSC on specific conditions for d and $\mathbf{R}_{ss}$. This will lead us to formal relations to MLP.

4.1. Cancellation Requirement

We wish to remove the undesired component of $q(n)$ in (11) from the GSC output $e(n) = q(n) - z(n)$, and hence the filter output $z(n)$ in (10b) must satisfy the condition

$$\mathbf{s}^T(\bar{n}-d)\mathbf{H}_u^T\mathbf{w} = \mathbf{s}^T(\bar{n}-d)\mathbf{D}\mathbf{h}_q. \tag{18}$$

As we have chosen the blocking matrix delay d on the left hand side according to the delay of the undesired component on the right hand side, we attain a MINT formulation of the cancellation requirement in (18) for arbitrary source signals,

$$\mathbf{H}_u^T\mathbf{w} = \mathbf{D}\mathbf{h}_q, \tag{19}$$

i.e. we seek a Wiener solution for the filter coefficients that is equivalent to the coefficients obtained by solving the MINT relation in (19). In contrast to the MINT solution however, the Wiener solution does not require any knowledge on $\mathbf{H}_u^T$ or $\mathbf{h}_q$ on the one hand, but may be biased by the source signal on the other hand. Similar statements can generally be made on MLP methods.

By inserting (16b) in (19) and making use of (17) the cancellation requirement may be simplified to

$$\mathbf{R}_{ss}^{-1}\mathbf{R}_{ss|d} = \mathbf{D} \;\Leftrightarrow\; \mathbf{D}\mathbf{R}_{ss}\mathbf{C} = \mathbf{0}, \tag{20}$$

if we allow $\mathbf{h}_q$ and $s(n)$ to be arbitrary. Note that neither $\mathbf{H}_u$ nor $\mathbf{h}_q$ have any influence on the cancellation requirement, as long as $\mathbf{H}_u$ has full column rank and hence cancels out. I.e. whether or not we can cancel the undesired component of $q(n)$ depends on the properties of the source signal only, namely on the autocovariance of $s(n)$. Hence, the blocking matrix could also be replaced by a simple pass-through and a delay, and one might see the delay itself as a blocking matrix for convolutive interference, where the blocking capability depends on the characteristics of the source signal.

If classical blocking of a signal component coming from a specific direction, as it is applied in noise reduction, is not required however, then steering is not strictly required either, although recommended for MLP in [4]. In fact, by replacing the blocking matrix by a delay, the resulting structure is rather similar to a multichannel linear predictor with prediction delay d, as it has been used e.g. in [10, 13], and has similar properties, as shown in the following. A remaining difference is given by the fixed beamformer, which poses additional freedom of design as compared to conventional linear prediction, where the prediction is commonly performed from the first microphone signal or the sum of all microphone signals.

4.2. Behavior of the Output Signal

Let us study on what conditions the requirement $\mathbf{D}\mathbf{R}_{ss}\mathbf{C} = \mathbf{0}$ is satisfied, which gives an unbiased filter

estimate such that the undesired component is canceled, and how the output signal behaves if this is not the case.

4.2.1. Unbiased Estimate

The requirement $\mathbf{D}\mathbf{R}_{ss}\mathbf{C} = \mathbf{0}$ is fulfilled in two cases:

1. The trivial but not very reasonable case is given if we choose $d = 0$, which implies $\mathbf{C} = \mathbf{0}$. In this case none of the components of $q(n)$ is considered to be desired and the output is fully canceled. This situation may also appear if a classical blocking matrix is used and the direct component is not fully canceled, e.g. due to steering mismatch.
2. The more reasonable case that actually leads to dereverberation follows from the observation that the relation $\mathbf{D}\mathbf{C} = \mathbf{0}$ always holds, hence we find an unbiased filter estimate if $\mathbf{R}_{ss} \propto \mathbf{I}$, i.e. if the source signal is a random white sequence[1]. The output signal $e(n)$ then equals the desired signal $\mathbf{s}^T(\bar{n})\mathbf{C}\mathbf{h}_q$.

Based on the latter observation we proposed to prewhiten [17] the microphone signals with respect to the source signal in order to attain an unbiased filter estimate, i.e. to remove the coloration of the source signal inherent in the microphone signals. This approach has been verified for stationary speech-shaped random sequences as a source signal in [17].

4.2.2. Biased Estimate

If $d \neq 0$ and $\mathbf{R}_{ss} \not\propto \mathbf{I}$ the filter coefficients will be biased and, equivalently, differ from the MINT solution. The filter output will not equal the undesired component of $q(n)$ leading to data-dependent distortion at the GSC output. With (11), (10b), and (16b) we derive for the GSC output

$$e(n) = \mathbf{s}^T(\bar{n})\mathbf{C}\mathbf{h}_q + \mathbf{s}^T(\bar{n}-d)\mathbf{B}\mathbf{h}_q \tag{21}$$

with the matrix $\mathbf{B}$ determining the bias given by

$$\mathbf{B} = \mathbf{D} - \mathbf{R}_{ss}^{-1}\mathbf{R}_{ss|d} \tag{22a}$$

$$= -\mathbf{R}_{ss}^{-1}\mathbf{D}\mathbf{R}_{ss}\mathbf{C}. \tag{22b}$$

In the last transition (17) is used. Note the formal similarity between (21) and (11) and that $\mathbf{B}$ could also be derived from the cancellation requirement in (20) directly. The joint impulse response of the cascade of room and GSC is determined by $\mathbf{C}\mathbf{h}_q$ for the first d samples, which represents the desired component of the overall impulse response, and $\mathbf{B}\mathbf{h}_q$ for the subsequent samples, which represents the data-dependent undesired component. Interestingly, only the first d samples of $\mathbf{h}_q$ have impact on the result, as we find the product $\mathbf{C}\mathbf{h}_q$ in both the expressions for desired and undesired component.

[1] from a merely algebraic point of view it was sufficient if $\mathbf{R}_{ss}$ was a diagonal matrix, which however violates the assumption made in (17).

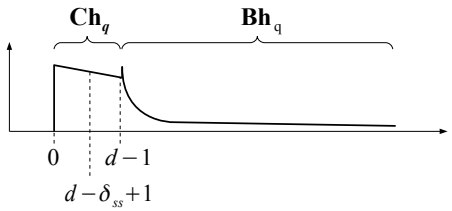


Fig. 4: Schematic depiction of the joint impulse response of room and GSC if the filter coefficients are biased. The matrix $\mathbf{B}$ depends on the last $\delta_{ss} - 1$ samples of the desired impulse response $\mathbf{C}\mathbf{h}_q$ and the autocovariance matrix $\mathbf{R}_{ss}$ of the source signal.

Interestingly, only the first d samples of $\mathbf{h}_q$ have impact on the result, as we find the product $\mathbf{C}\mathbf{h}_q$ in both the expressions for desired and undesired component.

A special case for the GSC output is given if $d = 1$ is chosen, i.e. if only the first sample $h_{q|0}$ of the impulse response $\mathbf{h}_q$ is desired. Then we derive for the output signal $e(n)$ in (21) the simplification

$$e(n) = h_{q|0}s(\bar{n}) - h_{q|0}\mathbf{s}^T(\bar{n}-1)\mathbf{R}_{ss}^{-1}\mathbf{r}_{ss|1}, \tag{23}$$

with $\mathbf{r}_{ss|1} = E\{\mathbf{s}(\bar{n}-1)s(\bar{n})\}$. The undesired component $h_{q|0}\mathbf{s}^T(\bar{n}-1)\mathbf{R}_{ss}^{-1}\mathbf{r}_{ss|1}$ in (23) equals the linear prediction of the desired component $h_{q|0}s(\bar{n})$, hence we can think of the GSC output signal $e(n)$ as the prediction residual of $s(\bar{n})$ weighted by $h_{q|0}$. In other words, both the transfer functions of speech production system and room are inverted, up to a factor and a delay, which is indeed the behavior that is expected for MLP of sufficient order.

Now suppose that the autocorrelation is zero for lags greater than δ_{ss}. We can state that if $d \geq \delta_{ss}$ is chosen, then the first $d - \delta_{ss} + 1$ columns of $\mathbf{B}$ are zero and therefore the first $d - \delta_{ss} + 1$ samples of $\mathbf{h}_q$ have no influence on the bias. The remaining bias will cause gradually delayed whitening of the reflections arriving with a delay in the range $[d - \delta_{ss} + 1, d - 1]$ relative to n_0, where reflections arriving with the delay $d - 1$ are fully whitened. Therefore, we can state that the larger the last $\delta_{ss} - 1$ samples of the desired impulse response $\mathbf{C}\mathbf{h}_q$, the higher will be the bias. The joint impulse response of the cascade of the room and the GSC and the aforementioned relations are schematically illustrated in fig. 4.

As compared to $\mathbf{h}_q$, the whitening effect may cause the actual joint impulse response to overshoot after the desired component, i.e. at samples with a delay greater than d, followed by a decay. An exemplary simulation is shown in fig. 5. The simulation setup is simi-

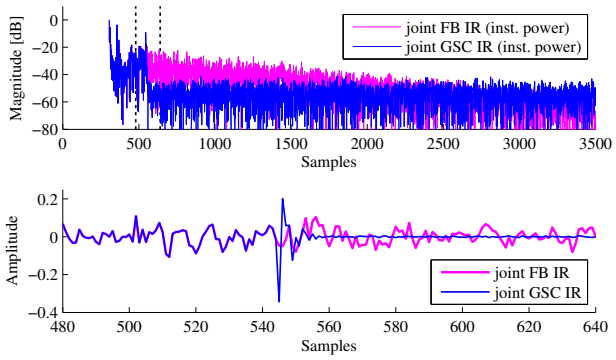


Fig. 5: Exemplary simulation result for the joint impulse response of the room and the fixed beamformer (in magenta) as well as the room and the GSC (in blue) for biased filter coefficients.

lar as in [17]. The RIRs with 360 ms reverberation time are chosen from the multichannel audio database [21], downsampled to 16 kHz and truncated after 8000 samples. The dead time delay is 304 samples. The source is positioned in the broadside direction at 2 m distance and three microphones with 8 cm spacing are selected. The fixed beamformer simply sums up the microphone signals and the blocking matrix passes them through with a delay of 240 samples. The signal-dependent filter is chosen to have 4000 samples, satisfying (4). As a source signal stationary Gaussian noise shaped by a 10^{th} order all-pole filter resembling the speech production system of duration 30 s has been chosen [17]. The figure depicts the joint impulse response of the room and the fixed beamformer in magenta (corresponds to fig. 2) as well as the room and the GSC in blue (corresponds to fig. 4). The instantaneous power of the joint impulse responses in dB is shown in the top part of the figure. The section of the impulse response around the end of the desired component (sample index 544), indicated by the two vertical dotted lines, is shown again in the bottom part, displaying the aforementioned overshoot and decay in the undesired component of the joint impulse response of the room and the GSC.

It is advisable to bear in mind that the derivations and the conclusions taken in this paper are to some extent based on the assumption of a stationary source signal. The actual statistical properties of speech signals vary relatively quickly as compared to the length $L = L_w + L_{h_u} - 1$ on which the Wiener solution is estimated. In (23) we therefore rather expect to attain a prediction filter $\mathbf{R}_{ss}^{-1}\mathbf{r}_{ss|1}$ which inverts the average coloration of the source signal over a long time window. Further, since L is much greater than the usual number of coefficients in a speech production system, only a few of the prediction filter coefficients will be significantly different from zero.

5. CONCLUSION

Similar to MLP, the GSC is able to perform dereverberation but also causes excessive whitening for non-white source signals. In fact, although both approaches evolve from different ideas, they admit rather similar mathematical formulations. For the GSC, it has been shown that the last $\delta_{ss} - 1$ samples of the desired impulse response, which depends on the room acoustics and the fixed beamformer, play a crucial role for the data-dependent bias, where δ_{ss} denotes the length of the autocorrelation of the source signal. Therefore it may be worthwhile to investigate the influence of the design of the fixed beamformer on the overall performance in future research.

6. ACKNOWLEDGMENTS

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 commemoration of our Research Fellow Nejem Huleihel, we take this opportunity to express our gratitude for all his valuable contributions within the DREAMS Initial Training Network.

7. 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] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[3] M. Triki and D.T.M. Slock, "Blind dereverberation of a single source based on multichannel linear prediction," in *Proc. 2005 Int. Workshop Acoustic Echo Noise Control (IWAENC 2005)*, Eindhoven, Netherlands, Sep. 2005, pp. 173–176.

[4] M. Triki and D.T.M. Slock, "Delay and predict equalization for blind speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 06)*, Toulouse, France, May 2006, vol. 5, pp. 97–100.

[5] M. Delcroix, T. Hikichi, and M. Miyoshi, "Blind dereverberation algorithm for speech signals based on multi-channel linear prediction," *Acoust. Sci. Technol.*, vol. 26, no. 5, pp. 432–439, 2005.

[6] 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.

[7] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1791–1801, 2007.

[8] Jae-Mo Yang and Hong-Goo Kang, "Online speech dereverberation algorithm based on adaptive multichannel linear prediction," *Audio, Speech, Lang. Process., IEEE/ACM Trans. on*, vol. 22, pp. 608–619, 2014.

[9] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Dereverberation by using time-variant nature of speech production system," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[10] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, 2009.

[11] T. Nakatani, Biing-Hwang Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying gaussian source model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, 2008.

[12] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 09)*, Taipei, Taiwan, April 2009, pp. 3733–3736.

[13] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[14] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[15] 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.

[16] F. Lim, M.R.P. Thomas, and P.A. Naylor, "Mintformer: A spatially aware channel equalizer," in *Appl. Signal Process. Audio Acoust. (WASPAA), 2013 IEEE Workshop on*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[17] T. Dietzen, N. Huleihel, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Speech dereverberation by data-dependent beamforming with signal pre-whitening," in *Proc. 2015 Signal Process. Conf. (EUSIPCO 2015)*, Nice, France, Aug. 2015.

[18] 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.

[19] 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.

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

[21] E. Hadad, F. Heese, P. Vary, 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.