Multichannel Signal Enhancement using Non-Causal, Time-Domain Filters

Jensen, Jesper Rindom; Christensen, Mads Græsbøll; Benesty, Jacob

Published in:
2013 IEEE International Conference on Acoustics, Speech, and Signal Processing

DOI (link to publication from Publisher):
10.1109/ICASSP.2013.6639075

Publication date:
2013

Document Version
Accepted author manuscript, peer reviewed version

MULTICHANNEL SIGNAL ENHANCEMENT USING NON-CAUSAL, TIME-DOMAIN FILTERS

Jesper Rindom Jensen[†]*, Mads Græsbøll Christensen*[†]*, and Jacob Benesty*[‡]

[†]Audio Analysis Lab, AD:MT
Aalborg University, Denmark
{jrj,mgc}@create.aau.dk

[‡] INRS-EMT, University of Quebec
Montreal, QC H5A 1K6, Canada
benesty@emt.inrs.ca

ABSTRACT

In the vast amount of time-domain filtering methods for speech enhancement, the filters are designed to be causal. Recently, however, it was shown that the noise reduction and signal distortion capabilities of such single-channel filters can be improved by allowing the filters to be non-causal. While non-causal filters require knowledge of the future, they can be implemented in practice by introducing a short delay. In this paper, we generalize the idea of exploiting non-causality in optimal filter designs to the multichannel scenario. More specifically, a set of optimal, non-causal, multichannel filters for enhancement based on an orthogonal decomposition is proposed. The evaluation shows that there is a potential gain in noise reduction and signal distortion by introducing non-causality. Moreover, experiments on real-life speech show that we can improve the perceptual quality.

Index Terms— Signal enhancement, time-domain filtering, multichannel, non-causal.

1. INTRODUCTION

Speech enhancement techniques are utilized in numerous applications such as telecommunications, teleconferencing, hearing-aids, surveillance systems, and human-machine interfaces. Before utilization of the speech, it has to be captured using one or more microphones. Unfortunately, background noise such as interfering speakers, car noise, fan noise, etc., are present in most real-life recording settings, and the noise will most likely have a detrimental impact on the aforementioned applications. For example, the noise will reduce the speech quality which can cause undesirable listener fatigue for hearing-aid users. Therefore, reduction of noise, aka. enhancement, is essential in various signal processing applications. In the past decades, a multitude of methods for combating noise have been proposed. A thorough overview of speech enhancement methods can be found in, e.g., [1, 2] and the references therein. These methods can generally be divided into four categories: spectral subtractive methods [3], filtering methods [4–6], statistical model-based methods [7–10], and subspace methods [11–14]. In this paper, we focus on filtering-based enhancement.

In many filtering methods for enhancement, a linear filter is applied to the observed signal. The filter should be designed to fulfill at least two criterias: the noise should be attenuated significantly, and the distortion of the desired signal after filtering should be negligible. Many equivalent filters can be obtained by deriving them in either the time domain or in different transform domains such as the Fourier [3,12,15] and Karhunen-Loève [16,17] domains. Herein, we restrict ourselves to the study of time-domain filters for multichannel signals. Many of such time-domain, enhancement filters are derived to be causal, with some recent examples being the single-channel, orthogonal decomposition (OD) based filters in [18, 19], the generalized singular value decomposition based filters in [20], and the multichannel spatio-temporal prediction method in [21]. If some of the involved signals are nonstationary, however, it is beneficial to introduce non-causality in the filter design. As reported in [22], this can be exploited to increase the amount of noise reduction of, e.g., the single-channel OD filters without introducing additional distortion of the desired signal. Inspired by the ideas presented in [22], we therefore derive a set of novel, optimal, non-causal filters for multichannel enhancement of speech in this paper. The proposed filters are based on the orthogonal decomposition, and can be seen as extensions of the multichannel filters in [18, 23]. The filters can be implemented in practice by introducing a short delay; in many cases, a significant noise reduction improvement can be obtained with only a few samples of delay. Moreover, we present closed-form performance measures for the proposed filters when the desired signal is periodic, i.e., these expressions hold for voiced speech [2, 24], and they facilitate the evaluation of the filters' performance without having to estimate any signal or noise statistics. That is, the evaluations conducted in this way are not disturbed by estimation errors in the statistics.

The remainder of the paper is organized as follows. First, the signal model and the problem of designing non-causal, time-domain filters for multichannel enhancement are defined in Section 2. In Section 3, we propose three optimal, non-causal filter designs. To facilitate evaluation of the filters, we present performance measures, we show that these have closed-forms for periodic desired signals, and we evaluate the theoretical gain of exploiting non-causality in Section 4. Then, the filters are evaluated on real-life speech in Section 5, and, in Section 6, a discussion relating the results presented herein to the state of the art is found.

2. PROBLEM FORMULATION

In the scenario considered in this paper, an array of N_s microphones capture a speech source signal $s(n_\mathrm{t})$ in some noise field. With this conventional setup, we have the following model for the signal captured by the n_s'th microphone at the discrete time instance n_t [25]:

$$\begin{aligned} y_{n_\mathrm{s}}(n_\mathrm{t}) &= g_{n_\mathrm{s}}(n_\mathrm{t}) * s(n_\mathrm{t}) + v_{n_\mathrm{s}}(n_\mathrm{t}) \\ &= x_{n_\mathrm{s}}(n_\mathrm{t}) + v_{n_\mathrm{s}}(n_\mathrm{t}) \,, \end{aligned} \tag{1}$$

where $g_{n_\mathrm{s}}(n_\mathrm{t})$ is the impulse response from the source location to the n_s'th microphone, $*$ is the linear convolution operator, and $v_{n_\mathrm{s}}(n_\mathrm{t})$ is the additive noise. It is assumed that the convolved source signal

This work was supported in part by the Villum Foundation.

$x_{n_s}(n_t)$ and the additive noise $v_{n_s}(n_t)$ are uncorrelated and zero-mean. Furthermore, $x_{n_s}(n_t)$ is coherent between the different microphones, whereas $v_{n_s}(n_t)$ is only partially coherent.

To facilitate the derivation of optimal, non-causal, noise reduction filters, we define a vector model counterpart of (1) as

$$\mathbf{y}_{n_s}(n_{t,k}) = \mathbf{x}_{n_s}(n_{t,k}) + \mathbf{v}_{n_s}(n_{t,k}), \tag{2}$$

with

$$\mathbf{y}_{n_s}(n_{t,k}) = \begin{bmatrix} y_{n_s}(n_{t,k}) & \cdots & y_{n_s}(n_{t,k} - M_t + 1) \end{bmatrix}^T \tag{3}$$

being a length M_t vector, $n_{t,k} = n_t + k$, k is the number of future samples utilized, and the vectors $\mathbf{x}_{n_s}(n_{t,k})$ and $\mathbf{v}_{n_s}(n_{t,k})$ are defined similarly to $\mathbf{y}_{n_s}(n_{t,k})$. To even further ease the filters' derivation, we stack the vectors related to the individual microphones, i.e.,

$$\bar{\mathbf{y}}(n_{t,k}) = \begin{bmatrix} \mathbf{y}_1^T(n_{t,k}) & \cdots & \mathbf{y}_{N_s}^T(n_{t,k}) \end{bmatrix}^T \tag{4}$$

and with similar definitions of $\bar{\mathbf{x}}(n_{t,k})$ and $\bar{\mathbf{v}}(n_{t,k})$. In consequence of $x_{n_s}(n_{t,k})$ and $v_{n_s}(n_{t,k})$ being uncorrelated by assumption, the correlation matrix of the stacked microphone observations is given by

$$\mathbf{R}_{\bar{\mathbf{y}}} = \mathrm{E}\left[\bar{\mathbf{y}}(n_{t,k})\bar{\mathbf{y}}^T(n_{t,k})\right] = \mathbf{R}_{\bar{\mathbf{x}}} + \mathbf{R}_{\bar{\mathbf{v}}}, \tag{5}$$

where $\mathrm{E}[\cdot]$ is the mathematical expectation operator, and $\mathbf{R}_{\bar{\mathbf{x}}}$ and $\mathbf{R}_{\bar{\mathbf{v}}}$ are the correlation matrices of $\bar{\mathbf{x}}(n_{t,k})$ and $\bar{\mathbf{v}}(n_{t,k})$, respectively.

Based on the presented model, the objective considered in this paper is to obtain a "good" estimate of $x_{n_s}(n_t)$ from the observed signal vector $\bar{\mathbf{y}}(n_{t,k})$. Conventionally, "good" means the noise should be reduced significantly, while the distortion of the desired signal should be negligible. Note that the desired signal in this work is the convolved source signal. While not considered here, the source signal can be obtained from the convolved source signal if needed by applying dereverberation (see, e.g., [26] and the references therein). Recently, an orthogonal decomposition approach was considered in the derivation of optimal, causal, noise reduction filters for multichannel signals [18, 23]. Here, we extend this approach to enable derivation of similar non-causal filters. The non-causal orthogonal decomposition of $\bar{\mathbf{x}}(n_{t,k})$ is

$$\bar{\mathbf{x}}(n_{t,k}) = x_{n_s}(n_{t,k})\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k} + \bar{\mathbf{x}}_i(n_{t,k}), \tag{6}$$

where

$$\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k} = \begin{bmatrix} \boldsymbol{\rho}^T_{\mathbf{x}_1 x_{n_s},k} & \cdots & \boldsymbol{\rho}^T_{\mathbf{x}_{N_s} x_{n_s},k} \end{bmatrix}^T \tag{7}$$

is the normalized cross-corelation vector between $\bar{\mathbf{x}}(n_{t,k})$ and $x_{n_s}(n_{t,k})$, and $\bar{\mathbf{x}}_i(n_{t,k})$ is the so-called interference vector being orthogonal to $x_{n_s}(n_{t,k})\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k}$. The subvectors $\boldsymbol{\rho}_{\mathbf{x}_{n_s}x_{n_s},k}$ of $\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k}$ are the cross-correlation vectors between $\mathbf{x}_{n_s}(n_{t,k})$ and $x_{n_s}(n_{t,k})$, i.e.,

$$\boldsymbol{\rho}_{\mathbf{x}_{n_s}x_{n_s},k} = \frac{\mathrm{E}\left[\mathbf{x}_{n_s}(n_{t,k})x_{n_s}(n_{t,k})\right]}{\mathrm{E}\left[x_{n_s}^2(n_{t,k})\right]}. \tag{8}$$

Combining (4) and (6) yields the orthogonal decomposition based signal model:

$$\bar{\mathbf{y}}(n_{t,k}) = x_{n_s}(n_{t,k})\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k} + \bar{\mathbf{x}}_i(n_{t,k}) + \bar{\mathbf{v}}(n_{t,k}). \tag{9}$$

Perhaps against intuition, the observed signal contains two noise components when utilizing the orthogonal decomposition approach: the interference signal vector $\bar{\mathbf{x}}_i(n_{t,k})$ and the additive noise vector $\bar{\mathbf{v}}(n_{t,k})$.

3. NON-CAUSAL FILTERS

Equipped with the signal model, the task is then to reduce the noise by applying a non-causal, finite impulse response (FIR) filter to the observed signal. This yields the following estimate of the desired signal:

$$\hat{x}_{n_s,k}(n_t) = \sum_{n_s=1}^{N_s} \mathbf{h}_{n_s}^T \mathbf{y}_{n_s}(n_{t,k}) = \bar{\mathbf{h}}^T \bar{\mathbf{y}}(n_{t,k}). \tag{10}$$

where $\mathbf{h}_{n_s}$ are filters of length M_t and

$$\bar{\mathbf{h}} = \begin{bmatrix} \mathbf{h}_1^T & \cdots & \mathbf{h}_{N_s}^T \end{bmatrix}^T. \tag{11}$$

Several optimal filter designs for multichannel noise reduction can be derived from the orthogonal decomposition based model in (9). In this section, we present the maximum signal-to-noise ratio (SNR) filter to motivate the introduction of non-causality in the filter design. Moreover, we propose non-causal, multichannel Wiener and minimum variance distortionless response (MVDR) filters.

3.1. Maximum SNR

The maximum SNR filter $\bar{\mathbf{h}}_{\max,k}$ is a filter maximizing the output SNR (oSNR). In the non-causal orthogonal decomposition approach to noise reduction, the oSNR is given by [18, 23]

$$\mathrm{oSNR}(\bar{\mathbf{h}}_k) = \frac{\sigma^2_{x_{n_s}} \bar{\mathbf{h}}_k^T \boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k}\boldsymbol{\rho}^T_{\bar{\mathbf{x}}x_{n_s},k}\bar{\mathbf{h}}_k}{\bar{\mathbf{h}}_k^T \mathbf{R}_{\mathrm{in},k}\bar{\mathbf{h}}_k}, \tag{12}$$

where $\mathbf{R}_{\mathrm{in},k} = \mathbf{R}_{\bar{\mathbf{x}}_{i,k}} + \mathbf{R}_{\bar{\mathbf{v}}}$, and $\mathbf{R}_{\bar{\mathbf{x}}_{i,k}}$ is the correlation matrix of the interference vector $\bar{\mathbf{x}}_i(n_{t,k})$. The oSNR can be recognized as a generalized Rayleigh quotient that is maximized when $\bar{\mathbf{h}}_k$ equals the maximum eigenvector of the matrix $\sigma^2_{x_{n_s}}\mathbf{R}^{-1}_{\mathrm{in},k}\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k}\boldsymbol{\rho}^T_{\bar{\mathbf{x}}x_{n_s},k}$. Clearly, this matrix is rank one, so the maximum oSNR is given by the maximum eigenvalue:

$$\lambda_{\max,k} = \mathrm{oSNR}(\bar{\mathbf{h}}_{\max,k}) = \sigma^2_{x_{n_s}}\boldsymbol{\rho}^T_{\bar{\mathbf{x}}x_{n_s},k}\mathbf{R}^{-1}_{\mathrm{in},k}\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k}. \tag{13}$$

It is important to note that, in general, $\lambda_{\max,p} \neq \lambda_{\max,q}$ for $p \neq k$. In other words, the oSNR will be different for different ks, so we may be able to improve the oSNR by introducing non-causality in the filter design. Obviously, the maximum SNR filter is given by

$$\bar{\mathbf{h}}_{\max,k} = \eta\mathbf{R}^{-1}_{\mathrm{in},k}\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k}, \tag{14}$$

where η is an arbitrary scaling constant. While η has no influence on the oSNR, it may affect the distortion of the desired signal.

3.2. Wiener

To obtain a Wiener filter design, we introduce an error function:

$$e_k(n_t) = \hat{x}_{n_s,k}(n_t) - x_{n_s}(n_t). \tag{15}$$

Minimizing the variance of the error $\mathrm{E}[e_k^2(n_t)]$ with respect to the filter response yields the Wiener design:

$$\bar{\mathbf{h}}_{\mathrm{W},k} = \sigma^2_{x_{n_s}}\mathbf{R}^{-1}_{\bar{\mathbf{y}}}\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_s},k}. \tag{16}$$

It can be shown that choosing

$$\eta = \frac{\sigma^2_{x_{n_s}}}{1 + \lambda_{\max,k}} \tag{17}$$

in (14) gives $\bar{\mathbf{h}}_{\mathrm{W},k}$, so the Wiener filter also maximizes the oSNR.

3.3. MVDR

The maximum SNR filter and the Wiener filters will most likely distort the desired signal. To tackle this issue, the MVDR principle can be used for designing the filter in (10). First, we introduce the speech reduction factor that is defined as the ratio between the power of the desired signal before and after noise reduction, i.e.,

$$\xi_{\text{sr}}(\bar{\mathbf{h}}_k) = \left(\bar{\mathbf{h}}_k^T \boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k}\right)^{-2}. \tag{18}$$

According to this measure, a filter $\bar{\mathbf{h}}_k$ is distortionless for $\xi_{\text{sr}}(\bar{\mathbf{h}}_k) = 1$. That is, a distortionless, non-causal, noise reduction filter can be derived by solving

$$\min_{\bar{\mathbf{h}}_k} \bar{\mathbf{h}}_k^T \mathbf{R}_{\text{in},k} \bar{\mathbf{h}}_k \quad \text{s.t.} \quad \bar{\mathbf{h}}_k^T \boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k} = 1. \tag{19}$$

The well-known solution to this type of optimization problems is given by

$$\begin{aligned}\bar{\mathbf{h}}_{\text{M},k} &= \mathbf{R}_{\text{in},k}^{-1} \boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k} \left(\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k}^T \mathbf{R}_{\text{in},k}^{-1} \boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k}\right)^{-1} \\ &= \mathbf{R}_{\bar{\mathbf{y}}}^{-1} \boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k} \left(\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k}^T \mathbf{R}_{\bar{\mathbf{y}}}^{-1} \boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k}\right)^{-1}. \end{aligned} \tag{20}$$

It can be shown that the MVDR filter is a scaled version of the Wiener filter, so it maximizes the oSNR while being distortionless in terms of the speech reduction factor [18, 23].

4. THEORETICAL PERFORMANCE

The performance measures such as the oSNRs and the signal reduction factors of the proposed filters are functions of the statistics of the desired signal and the noise. These statistics need to be estimated in practice, so it is difficult to evaluate the performance gain of introducing non-causality without also measuring the impact of errors in the estimated signal and noise statistics. In this section, we therefore assume a specific and realistic model of the observed signal that enables the derivation of closed-form performance measure expressions. Using these expressions, the potential gain of introducing non-causality can be clearly identified.

A widely used and accepted model for voiced speech, is the harmonic model:

$$s(n_\text{t}) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 n_\text{t}} + \alpha_l^* e^{-jl\omega_0 n_\text{t}}, \tag{21}$$

where $\alpha_l = \frac{A_l}{2} e^{j\phi_l}$ is the complex amplitude of the lth harmonic, A_l and ϕ_l are the real amplitude and phase of the lth harmonic, respectively, ω_0 is the fundamental frequency, and $(\cdot)^*$ is the complex conjugate. By using the covariance matrix model [27] and the fact that the acoustical room impulse response and the source signal are stationary per assumption in the considered time window, the covariance matrix of the convolved source signal can be written as

$$\mathbf{R}_{\bar{\mathbf{x}}} = \bar{\mathbf{Z}}_g \mathbf{P} \bar{\mathbf{Z}}_g^H, \tag{22}$$

with

$$\bar{\mathbf{Z}}_g = \begin{bmatrix}\mathbf{G}_1^T & \cdots & \mathbf{G}_{N_\text{s}}^T\end{bmatrix}^T \mathbf{Z}, \tag{23}$$

$$\mathbf{Z} = \begin{bmatrix}\mathbf{z}(\omega_0) & \mathbf{z}(-\omega_0) & \cdots & \mathbf{z}(L\omega_0) & \mathbf{z}(-L\omega_0)\end{bmatrix}, \tag{24}$$

$$\mathbf{z}(l\omega_0) = \begin{bmatrix}1 & e^{-jl\omega_0} & \cdots & e^{-jl\omega_0(M_g-1+M_\text{t}-1)}\end{bmatrix}^T, \tag{25}$$

$$\mathbf{P} = \text{diag}\left\{\begin{bmatrix}|\alpha_1|^2 & |\alpha_1^*|^2 & \cdots & |\alpha_L|^2 & |\alpha_L^*|^2\end{bmatrix}^T\right\}, \tag{26}$$

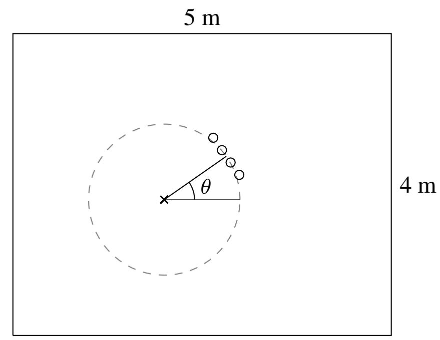

Fig. 1. Top-down view of the simulated room setup where $\times$ and $\circ$ denotes the source and sensor locations, respectively.

$$\mathbf{G}_{n_\text{s}} = \begin{bmatrix}\mathbf{g}_{n_\text{s}} & \mathbf{S}_1 \mathbf{g}_{n_\text{s}} & \cdots & \mathbf{S}_{M_\text{t}-1} \mathbf{g}_{n_\text{s}}\end{bmatrix}^T, \tag{27}$$

$$\mathbf{g}_{n_\text{s}} = \begin{bmatrix}g_{n_\text{s}}(0) & \cdots & g_{n_\text{s}}(-M_g+1) & \mathbf{0}_{1\times(M_\text{t}-1)}\end{bmatrix}^T, \tag{28}$$

$$\mathbf{S}_{m_\text{t}} = \begin{bmatrix}\mathbf{0}_{m_\text{t}\times(M_g-m_\text{t})} & \mathbf{0}_{m_\text{t}\times m_\text{t}} \\ \mathbf{I}_{(M_g-m_\text{t})\times(M_g-m_\text{t})} & \mathbf{0}_{(M_g-m_\text{t})\times m_\text{t}}\end{bmatrix}, \tag{29}$$

where M_g is the length of the acoustical impulse response, diag$\{\cdot\}$ denotes transformation of a vector into a diagonal matrix, $(\cdot)_{p\times q}$ denotes a matrix of size $p \times q$, $\mathbf{0}$ is a matrix of zeros, and $\mathbf{I}$ is the identity matrix.

With the expression for $\mathbf{R}_{\bar{\mathbf{x}}}$ given the parameters of the periodic signals, we can find closed-form expression for the performance measures of the proposed filters. First, we write the normalized cross-correlation vector $\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k}$ as

$$\boldsymbol{\rho}_{\bar{\mathbf{x}}x_{n_\text{s}},k} = \sigma_{x_{n_\text{s}}}^{-2} \mathbf{R}_{\bar{\mathbf{x}}} \mathbf{i}_{(n_\text{s}-1)M_\text{t}+k} = \sigma_{x_{n_\text{s}}}^{-2} \bar{\mathbf{Z}}_g \mathbf{P} \bar{\mathbf{Z}}_g^H \mathbf{i}_{q(n_\text{s},k)}, \tag{30}$$

where $\mathbf{i}_{q(n_\text{s},k)}$ is a unit vector having a one at the $q(n_\text{s},k)$th entry and zeros at all other entries, and $q(n_\text{s},k) = (n_\text{s}-1)M_\text{t} + k$. By combining (30) with (13), respectively, we get the following closed-form oSNR expression:

$$\begin{aligned}\text{oSNR}(\mathbf{h}_{\text{W},k}) &= \text{oSNR}(\mathbf{h}_{\text{M},k}) \\ &= \sigma_{x_{n_\text{s}}}^{-2} \bar{\mathbf{z}}_{\text{r},(n_\text{s},k)} \mathbf{P} \bar{\mathbf{Z}}_g^H \mathbf{R}_{\text{in},k}^{-1} \bar{\mathbf{Z}}_g \mathbf{P} \bar{\mathbf{z}}_{\text{r},(n_\text{s},k)}^H. \end{aligned} \tag{31}$$

A closed-form expression for the signal reduction factor of the Wiener filter can be obtained by writing it as a function of the oSNR, i.e.,

$$\xi_{\text{sr}}(\bar{\mathbf{h}}_{\text{W},k}) = \left[\text{oSNR}^{-1}(\bar{\mathbf{h}}_{\text{W},k})) + 1\right]^2. \tag{32}$$

We then proceed to evaluate the potential gain of exploiting non-causality by using the closed-form expressions in (31) and (32). For this evaluation, we assumed that the desired signal is modeled by (21) with $L = 8$, $\alpha_l = 1$, and $\omega_0 = 0.1578$. The desired signal was assumed to be generated by a source placed at $(2\text{ m}, 1.8\text{ m}, 1.5\text{ m})$ in a room with the dimensions $(5\text{ m}, 4\text{ m}, 3\text{ m})$. A top-down view of the room is shown in Fig. 1. Furthermore, the speed of sound in the room was 340 m/s, and the reverberation time was $T_{60} \approx 0.4$ s. The signal source was then assumed to be recorded by a uniform linear array (ULA), with four omnidirectional microphones and a microphone spacing of $d = 0.02$ m, and the noise on each microphone was assume to be white Gaussian with an SNR of 10 dB. The ULA was placed at the same height as the source (1.5 m), and was otherwise located as depicted in Fig. 1. Using this setup, we measured the performance of the filters of order $M_\text{t} = 30$ for 500 equidistant array angles in the interval $\theta \in [0°; 360°[$. For each angle, the acoustical

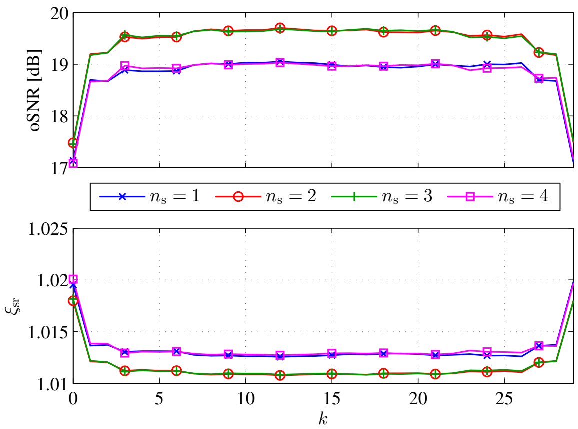

Fig. 2. Performance results obtained using the expressions in (top) (31) and (bottom) (32).

room impulse responses (RIRs) of length $M_g = 4,096$ were generated using an online toolbox [28] based on the image method [29] at a sampling frequency of $f_s = 8$ kHz. The performance measures were averaged over all different θs, and the results are shown in Fig. 2 as a function of the number of future samples k used by the filter and as a function of the reference sensor number n_s for the desired signal. We observe that the performance measures varies for the different ks and n_s, and that we can improve both measures by changing these values. As an example, the oSNR can be improved by ≈ 2.5 dB by choosing $k = 12$ and $n_s = 2$ instead of the traditional choice of $k = 0$ and $n_s = 1$. Note that this also implies a small improvement wrt. the signal reduction factor.

5. EXPERIMENTAL RESULTS

The proposed filtering methods were also evaluated on real-life speech. In these experiments, the room was again simulated using the image method, and the speech source was assumed to be placed at $(2\text{ m}, 3.5\text{ m}, 2\text{ m})$ in a room with dimensions $(5\text{ m}, 4\text{ m}, 4\text{ m})$. The source was recorded by an array of three microphones with coordinates $x = \{0.98, 1.00, 1.02\}$ m, $y = 2.5$ m, and $z = 2$ m at $f_s = 8$ kHz. As in Sec. 4, we then generated RIRs for the microphones using an online MATLAB toolbox for a reverberation time of $T_{60} = 0.4$ s, and the RIRs were used to generate the multichannel speech signal. Using this setup, we evaluated the causal and non-causal Wiener and MVDR filters. Two different implementations of the non-causal filters were considered: one with $k = M_t$ and one where k is chosen at each time instance to maximize the estimated oSNR. The results obtained using these different implementations are denoted by $(\cdot)^{\text{NC}}$ and $(\cdot)^{\text{NC}}_{\text{max}}$, respectively. The statistics of the noise needed in the filter designs were estimated from the past 400 samples at each time instance, and a small amount of regularization was added to the so-obtained observed signal statistics as suggested in [30] with $\lambda = 0.05$. The evaluation was conducted for a male and a female speech excerpt each of length ≈ 2 s from the Keele database [31]. Each excerpt was then enhanced in different noise scenarios (car, exhibition, street, babble, white), at different filter lengths (30, 40, 50), and at different iSNRs (0 dB, 5 dB, 10 dB). The noise was generated to be diffuse, and the iSNR was the same at each microphone. For each filter length and iSNR, the differences between the PESQ scores[1] [32] of the causal filters and the corresponding non-causal filters were averaged over the different speech and noise scenarios. The resulting means are depicted in Fig. 3 with 95 % confidence intervals. From these results, we observe that the perceptual quality in terms of PESQ scores can indeed be improved by exploiting non-causality, especially for low filter lengths and low iSNRs. In many cases, this can be concluded with 95 % confidence as 0 is not included in the confidence interval. While the actual perceptual improvement may be difficult to assess from the PESQ scores, our informal listening tests confirmed that the improvement is significant and audible in most cases. Moreover, the results suggest that there is only a slight difference between using a fixed $k = \lfloor M_t/2 \rfloor$ and using the optimal k in many cases. This is an important observation, as the non-causal filters are easily implemented in practice when k is fixed.

[1]The PESQ scores are predicted mean opinion scores (MOS).

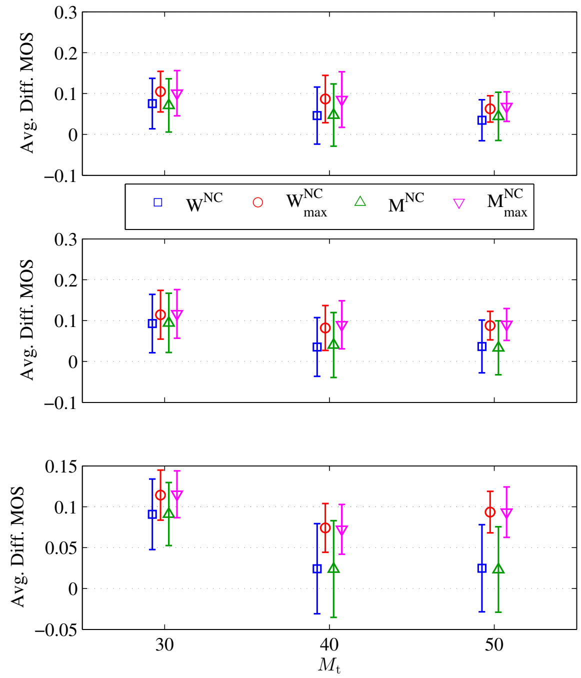

Fig. 3. Average difference in PESQ scores between the causal and non-causal Wiener filters (W^{NC}, $\text{W}^{\text{NC}}_{\text{max}}$) and between the causal and non-causal MVDR filters (M^{NC}, $\text{M}^{\text{NC}}_{\text{max}}$) in scenarios with iSNRs of (top) 0 dB, (middle) 5 dB, and (bottom) 10 dB, respectively.

6. DISCUSSION

The work presented in this paper is focused on the derivation of non-causal, multichannel filters for speech enhancement in the time domain. More specifically, the proposed filters are based on an orthogonal decomposition of the desired signal. The orthogonal decomposition approach was first considered in [18, 19] for single-channel speech enhancement in the time domain, and it was also considered in the frequency domain [33]. Recently, it was showed that the performance (oSNR and distortion) of these time-domain filters can be improved by allowing the filters to be non-causal. This motivated the work presented in this paper, which can be seen as non-causal counterparts to the causal, multichannel filters proposed in [18, 23]. As in the single-channel case, the reported results reveal a significant performance improvement by introducing non-causality compared to the corresponding causal filters [18]. The non-causal filters presented herein were also briefly mentioned in [34], but no evaluation of the filters were presented.

7. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Signals and Communication Technology. Springer, 2005.

[2] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[4] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, Jun. 1978.

[5] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[6] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[7] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[9] K. V. Sørensen and S. V. Andersen, "Rayleigh mixture model-based hidden Markov modeling and estimation of noise in noisy speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 901–917, Mar. 2007.

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[11] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Commun.*, vol. 10, no. 1, pp. 45–57, Feb. 1991.

[12] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[13] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[14] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. on Advances in Signal Process.*, vol. 2007, no. 1, pp. 24, Jun. 2007.

[15] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[16] J. Chen, J. Benesty, and Y. Huang, "Study of the noise-reduction problem in the Karhunen-Loève expansion domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 787–802, May 2009.

[17] J. Benesty, J. Chen, and Y. Huang, "Speech enhancement in the Karhunen-Loève expansion domain," *Synthesis Lectures on Speech and Audio Processing*, vol. 7, no. 1, pp. 1–112, 2011.

[18] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters – A Theoretical Study*, Number VII. Springer, 1 edition, 2011.

[19] J. Benesty, J. Chen, Y. Huang, and T. Gaensler, "Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction," *J. Acoust. Soc. Am.*, vol. 132, no. 1, pp. 452–464, Jul. 2012.

[20] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[21] J. Benesty, Y. Huang, and J. Chen, *Microphone Array Signal Processing*, vol. 1, Springer-Verlag, 2008.

[22] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Non-causal time-domain filters for single-channel noise reduction," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 5, pp. 1526–1541, Jul. 2012.

[23] J. Benesty, M. Souden, and J. Chen, "A perspective on multichannel noise reduction in the time domain," *Applied Acoust.*, vol. 74, no. 3, pp. 343–355, Mar. 2013.

[24] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[25] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*, Springer-Verlag, 2001.

[26] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Signals and Communication Technology. Springer, 2010.

[27] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Education, Inc., 2005.

[28] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2010, Ver. 2.0.20100920.

[29] 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.

[30] F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation - An Engineering Approach using MATLAB®*, John Wiley & Sons Ltd, 2004.

[31] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. Eurospeech*, Sep. 1995, pp. 837–840.

[32] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," , no. P.862, pp. 1–30, Feb. 2001.

[33] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 273–276.

[34] M. R. Bai, J.-G. Ih, and J. Benesty, *Acoustic Array Systems – Theory, Implementation and Application*, Wiley-IEEE Press, 2013, to appear.