Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Harmonic beamformers for speech enhancement and dereverberation in the time domain

Jensen, Jesper Rindom; Karimian-Azari, Sam; Christensen, Mads Græsbøll; Benesty, Jacob

Published in:
Speech Communication

DOI (link to publication from Publisher):
10.1016/j.specom.2019.11.003

Publication date:
2020

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Jensen, J. R., Karimian-Azari, S., Christensen, M. G., & Benesty, J. (2020). Harmonic beamformers for speech enhancement and dereverberation in the time domain. *Speech Communication*, *116*, 1-11. Advance online publication. https://doi.org/10.1016/j.specom.2019.11.003

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Harmonic Beamformers for Speech Enhancement and Dereverberation in the Time Domain

J. R. Jensen[a,*], S. Karimian-Azari[a], M. G. Christensen[a,**], J. Benesty[a,b]

[a] *Audio Analysis Lab, CREATE, Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark.*
[b] *University of Quebec, 800 de la Gauchetiere Ouest, QC H5A 1K6 Montreal, Canada.*

Abstract

This paper presents a framework for parametric broadband beamforming that exploits the frequency-domain sparsity of voiced speech to achieve more noise reduction than traditional nonparametric broadband beamforming without introducing additional distortion. In this framework, the harmonic model is used to parametrize the signal of interest by a single parameter, the fundamental frequency, whereby both speech enhancement and derevereration is performed. This framework thus exploits both the spatial and temporal properties of speech signals simultaneously and includes both fixed and adaptive beamformers, such as (1) delay-and-sum, (2) null forming, (3) Wiener, (4) minimum variance distortionless response (MVDR), and (5) linearly constrained minimum variance beamformers. Moreover, the framework contains standard broadband beamforming as a special case, whereby the proposed beamformers can also handle unvoiced speech. The reported experimental results demonstrate the capabilities of the proposed framework to perform both speech enhancement and dereverberation simultaneously. The proposed beamformers are evaluated in terms of speech distortion and objective measures for speech quality and speech intelligibility, and are compared to nonparametric broadband beamformers. The

*Corresponding author.
**EURASIP member.
Email addresses: jrj@create.aau.dk (J. R. Jensen), s.karimian.azari@gmail.com (S. Karimian-Azari), mgc@create.aau.dk (M. G. Christensen), benesty@emt.inrs.ca (J. Benesty)

results show that the proposed beamformers perform well compared to traditional methods, including a state-of-the-art dereverberation method, particularly in adverse conditions with high amounts of noise and reverberation.
Keywords: microphone arrays, beamforming, noise reduction, enhancement, dereverberation, time domain

---

1. Introduction

Speech signals recorded in voice communication systems are often accompanied by unwanted noise, reflections from the acoustic environment, and interfering sources in real life. These nuisance signals, which degrade the quality and intelligibility of the speech signals, have a profound, negative impact on voice communication systems, so an effective speech enhancement method is required to mitigate or eliminate the effects of added noise, reverberation, and interference. Nowadays, many voice communication systems are equipped with microphone arrays that provide spatial sampling in addition to the temporal sampling. Microphone arrays increase the performance of voice communication systems as the number of microphones increases, as the ability to attenuate noise, remove reflections, and suppress interferences is hereby increased [1].

Beamforming is an approach for noise reduction using microphone arrays. Beamforming applies filters across the elements of the array, and an optimal filter is desired that minimizes the noise and competing interference with a reasonable distortion on the desired speech signal. A beamformer is typically applied to discriminate signals from different direction of arrivals (DOAs), other than that of the desired signal [2]. Narrowband beamformers, which are generally used on communication and radar signals at certain frequency bands, attenuate signals from other directions. They are designed to pass the signal of interest, attenuate noise, and reject interferers [3]. Broadband beamformers for acoustic signals, such as speech, are generally designed using narrowband beamformers, one for each bands of the broadband signal decomposed using a filterbank, for example using the short-time Fourier transform. To accomplish effective noise reduction,

numerous adaptive beamformers have been developed throughout the years (see [4, 5, 1, 6] and the references therein). The linearly constrained minimum variance (LCMV) beamformer [3] minimizes the residual noise, and enforces a set of linear constraints on the desired signal and interferers. Also, the Wiener post-filtering of the output of the minimum variance distortionless response (MVDR) beamformer [7] provides a minimum mean-squared error (MMSE) solution [8] which is equivalent to the multichannel wiener filter. Another adaptive approach is based on the Karhunen-Loève expansion [9] by means of which multichannel linear filtering can be performed based on joint diagonalization of either the correlation matrices of the noisy speech and the noise signals [10] or the correlation matrices of the speech and the noise signals [11]. Using such an expansion, filters can be designed to minimize the speech distortion subject to a flexible noise reduction level [12, 10, 13].

In general, broadband beamformers are designed across all frequency bands covering the spectrum of the speech signal. However, large parts of speech signals, and many audio signals too, exhibit sparsity in their spectrum. An example of this is the spectrum of voiced speech, which comprised a finite number of harmonics due to its quasi-periodic nature. In other words, only a few frequency bands constitute the signal of interest, and nonparametric broadband beamformers, e.g., the delay-and-sum and the MVDR beamformers, ignore this and may retain noise in frequency bands where the signal of interest is actually not present. Various filters based on the harmonic model have been proposed for single-channel signal enhancement [14] and dereverberation [15]. Also, adaptive filters based on the Capon spectral estimator [16] have been proposed [17]. This harmonic model-based filter passes periodic signals undistorted while minimizing the power of noise and interferers. For multichannel signal enhancement, the use of the harmonic model have not been widely considered to the best of the authors' knowledge. A few examples are: in [18], where a harmonic transform is used as a preprocessor before localizing periodic sounds; in [19, 20] where the harmonic model is used to reduce the effect of periodic ego noise and; in [21] where APES-like [22] harmonic filters for multichannel enhancement were

proposed.

In this paper, we propose several new solutions to the multichannel signal enhancement problem based on the harmonic model, which describes the signal of interest as a sum of sinusoids whose frequencies are integral multiples of the fundamental frequency. More specifically, we generalize the principles of the single-channel filterbank [17] and the spatio-temporal filtering technique [23], and propose harmonic model-based beamforming that resembles a filterbank designed for the given spatial and spectral characteristics of the signal of interest. As an example, these model-based beamformers enable us to achieve more noise reduction than traditional non-parametric beamformers without introducing further signal distortion, since they can remove noise at non-harmonic frequencies even in the steering direction. Moreover, utilizing the harmonic model, the beamformers can reduce spectral smearing and thereby reduce the effects of reverberation. The DOA and the fundamental frequency of the signal of interest are treated as known parameters. The problem of estimating these parameters from noisy observed signals is considered beyond the scope of this paper, and we instead refer the interested readers to the many existing methods for finding them, e.g. [24, 25, 26, 27, 28, 23]. We design fixed delay-and-sum and null forming beamformers herein with distortionless constraints based on the aforementioned spatial and spectral parameters of the multichannel signals. To reduce incoherent noise as well as the interferers, we derive adaptive harmonic model-based beamformers based on the MVDR and LCMV beamformers as well as the multichannel Wiener filter.

The remainder of this paper is organized as follows. Section II describes the multichannel signal model and problem formulation that form the basis of the paper. Section III outlines the conventional beamforming approach. In Section IV, objective performance metrics are introduced, namely the noise reduction factor, speech distortion index, and mean-squared error criterion. Then, Sections V and VI develop fixed and adaptive harmonic model-based beamformers, respectively, followed by Section VII, wherein it is shown how traditional non-parametric beamformers can be obtained as a case of the harmonic model-based

beamformers. Finally, experimental results are presented in Section VIII and Section IX concludes on this work.

2. Signal Model and Problem Formulation

We consider a signal model in which a microphone array with M sensors receives the unknown speech signal $s(t)$, at the discrete-time index t, in some noise field. The received signals across the array are expressed as [1]

$$\begin{aligned} y_m(t) &= g_m(t) * s(t) + v'_m(t) \\ y_m(t) &= g^{\mathrm{d}}_m(t) * s(t) + g^{\mathrm{r}}_m(t) * s(t) + v'_m(t) \qquad (1) \\ &= x_m(t) + v_m(t),\ m = 1, 2, \ldots, M, \qquad (2) \end{aligned}$$

where m denotes the microphone index, $g_m(t)$ is the acoustic impulse response from the speech signal source to the microphone, which can be be decomposed into the impulse response for the direct speech component, $g^{\mathrm{d}}_m(t)$, and the impulse response for the reverberation $g^{\mathrm{r}}_m(t)$, i.e., $g_m(t) = g^{\mathrm{d}}_m(t) + g^{\mathrm{r}}_m(t)$. The variables $x_m(t) = g^{\mathrm{d}}_m(t) * s(t)$, $v'_m(t)$, and $v_m(t) = g^{\mathrm{r}}_m(t) * s(t) + v'_m(t)$ are the speech, additive noise, and reverberation-plus-additive noise signals, respectively. The speech and noise components are assumed to be uncorrelated and zero-mean. The terms $x_m(t),\ m = 1, 2, \ldots, M$, can be seen to be coherent across the array. The noise signals, $v_m(t),\ m = 1, 2, \ldots, M$, are typically only partially coherent across the array. We here choose microphone 1 as the reference sensor, whereby $x_1(t)$ is the desired signal that we seek to recover from the sensors' observations. Moreover, we assume that the unknown speech source signal is quasi-stationary over a short interval, e.g., 20–30 ms. Hence, over the most recent time samples, $\{\ s(t),\ \ s(t-1),\ \cdots,\ s(t-L+1)\ \}$, the spectral and statistical properties of the signal are constant for small L.

In this paper, we consider a uniform linear array (ULA) consisting of M omnidirectional microphones, where the distance between two successive sensors is equal to δ and the direction of the source signal to this ULA is parameterized by the azimuthal angle θ lying inside the range 0 to π. The speech signal is

modeled as a sum of sinusoids, which is a particularly good model for voiced speech. Therefore, we model the desired, direct speech component, $x_m(t)$, at the mth microphone as a harmonic signal source. By enhancing the signal according to this model we expect, not only reduce the additive noise, but also combat reverberation, since this will lead to spectral and temporal smearing of the signal source, which is not included in the harmonic model. That is, by reconstructing the harmonic components and suppressing the residual noise and nonharmonic components we can enhance the signal of interest without assuming a priori knowledge about the indirect-path components of the acoustic impulse response. Thus, our model for the desired signal is formulated as [29]:

$$x_m(t) = \sum_{n=-N}^{N} a_n e^{\jmath n\omega_0[t - f_s\tau_m(\theta)]}, \tag{3}$$

where N is the model order, the complex amplitude a_n is associated with the nth harmonic, $\jmath = \sqrt{-1}$ is the imaginary unit, ω_0 is the pitch or fundamental frequency, f_s is the sampling frequency,

$$\tau_m(\theta) = (m-1)\frac{\delta \cos\theta}{c} \tag{4}$$

is the relative delay of an impinging plane wave on the ULA, and c is the propagation speed of sound in the air. We note that $a_0 = 0$ since the signals, $x_m(t)$, $m = 1, \ldots, M$, are assumed zero-mean. Basically, the broadband signal, $x_m(t)$, whose fundamental frequency is ω_0, is the sum of $2N$ narrowband signals. Using (3), we can express (1) as

$$\begin{aligned} y_m(t) &= \sum_{n=-N}^{N} a_n e^{\jmath n\omega_0[t - f_s\tau_m(\theta)]} + v_m(t) \\ &= \sum_{n=-N}^{N} a_n e^{\jmath n\omega_0 t} e^{-\jmath n\omega_0 f_s\tau_m(\theta)} + v_m(t). \end{aligned} \tag{5}$$

Putting together the samples of the mth microphone observations in a vector

of length L, we get

$$\begin{aligned}\mathbf{y}_m(t) &= [\ y_m(t)\ \ y_m(t-1)\ \cdots\ y_m(t-L+1)\ ]^T \\ &= \mathbf{x}_m(t) + \mathbf{v}_m(t) \\ &= \mathbf{D}_{m,N}(\theta,\omega_0)\mathbf{a}(t,\omega_0) + \mathbf{v}_m(t), \end{aligned} \tag{6}$$

where the superscript ^{T} is the transpose operator, $\mathbf{x}_m(t) = \mathbf{D}_{m,N}(\theta,\omega_0)\mathbf{a}(t,\omega_0)$, and the nth column of the $L \times 2N$ matrix, $\mathbf{D}_{m,N}(\theta,\omega_0)$, is given by

$$\mathbf{d}_{m,n}(\theta,\omega_0) = e^{-\jmath n\omega_0 f_s \tau_m(\theta)} \times \left[\ 1\ \ e^{-\jmath n\omega_0}\ \cdots\ e^{-\jmath n\omega_0(L-1)}\ \right]^T$$

being a vector of length L. Furthermore, we have that

$$\mathbf{a}(t,\omega_0) = [\ a_{-N}e^{-\jmath N\omega_0 t}\ \ a_{-N+1}e^{-\jmath(N-1)\omega_0 t}\ \cdots\ a_N e^{\jmath N\omega_0 t}\ ]^T \tag{7}$$

is a vector of length $2N$, and

$$\mathbf{v}_m(t) = [\ v_m(t)\ \ v_m(t-1)\ \cdots\ v_m(t-L+1)\ ]^T. \tag{8}$$

The complex amplitudes, $[\ a_{-N}\ \ a_{-N+1}\ \cdots\ a_N\ ]$, are assumed to be zero-mean circular complex random variables that have independent phases uniformly distributed on the interval $(-\pi\,,\,\pi]$. Therefore $E\left[a_i a_j^*\right] = 0$ for $i \neq j$, and the correlation matrix of $\mathbf{a}$ (of size $2N \times 2N$) is

$$\mathbf{R}_{\mathbf{a}} = \text{diag}\left(E\left[|a_{-N}|^2\right], E\left[|a_{-N+1}|^2\right], \ldots, E\left[|a_N|^2\right]\right), \tag{9}$$

where $E[\cdot]$ is the expectation operator, and the superscript * is the complex-conjugate operator. Define the vector of length $2N$:

$$\mathbf{1}_{2N} = [\ 1\ \ 1\ \cdots\ 1\ ]^T. \tag{10}$$

It is obvious that $\mathbf{1}_{2N}^T\mathbf{a}(t,\omega_0) = x_1(t)$, which is the desired signal. Now, concatenating all microphone signal vectors, we obtain the vector of length ML:

$$\begin{aligned}\underline{\mathbf{y}}(t) &= \left[\ \mathbf{y}_1^T(t)\quad \mathbf{y}_2^T(t)\quad \cdots\quad \mathbf{y}_M^T(t)\ \right]^T \\ &= \underline{\mathbf{x}}(t) + \underline{\mathbf{v}}(t) \\ &= \underline{\mathbf{D}}_N(\theta,\omega_0)\mathbf{a}(t,\omega_0) + \underline{\mathbf{v}}(t), \end{aligned} \tag{11}$$

where $\underline{\mathbf{x}}(t) = \underline{\mathbf{D}}_N(\theta, \omega_0)\mathbf{a}(t, \omega_0)$,

$$\underline{\mathbf{D}}_N(\theta, \omega_0) = \begin{bmatrix} \mathbf{D}_{1,N}(\theta, \omega_0) \\ \mathbf{D}_{2,N}(\theta, \omega_0) \\ \vdots \\ \mathbf{D}_{M,N}(\theta, \omega_0) \end{bmatrix} \tag{12}$$

is a matrix of size $ML \times 2N$, and

$$\underline{\mathbf{v}}(t) = \begin{bmatrix} \mathbf{v}_1^T(t) & \mathbf{v}_2^T(t) & \cdots & \mathbf{v}_M^T(t) \end{bmatrix}^T. \tag{13}$$

We deduce that the correlation matrix of $\underline{\mathbf{y}}(k)$ (of size $ML \times ML$) is

$$\begin{aligned} \mathbf{R}_{\underline{\mathbf{y}}} &= E\left[\underline{\mathbf{y}}(t)\underline{\mathbf{y}}^H(t)\right] \\ &= \mathbf{R}_{\underline{\mathbf{x}}} + \mathbf{R}_{\underline{\mathbf{v}}} \\ &= \underline{\mathbf{D}}_N(\theta, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta, \omega_0) + \mathbf{R}_{\underline{\mathbf{v}}}, \end{aligned} \tag{14}$$

where the superscript ^{H} is the conjugate-transpose operator,

$$\mathbf{R}_{\underline{\mathbf{x}}} = \underline{\mathbf{D}}_N(\theta, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta, \omega_0) \tag{15}$$

is the correlation matrix of $\underline{\mathbf{x}}(t)$, and $\mathbf{R}_{\underline{\mathbf{v}}} = E\left[\underline{\mathbf{v}}(t)\underline{\mathbf{v}}^H(t)\right]$ is the correlation matrix of $\underline{\mathbf{v}}(t)$. It is important to observe that the matrix $\mathbf{R}_{\underline{\mathbf{x}}}$ is rank deficient only if $ML > 2N$, which is easy to satisfy by just increasing M or (especially) L; this will always be assumed. We will see how to exploit the nullspace of $\mathbf{R}_{\underline{\mathbf{x}}}$ to derive all kind of broadband beamformers. In the rest, it is assumed that the desired signal propagates from the fixed direction θ_0; so in (11) and (14), θ is replaced by θ_0. Therefore, our signal model is now

$$\underline{\mathbf{y}}(t) = \underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{a}(t, \omega_0) + \underline{\mathbf{v}}(t). \tag{16}$$

3. Broadband Beamforming

The conventional way to perform beamforming is by applying a complex-valued temporal linear filter of length L at the output of each microphone and

summing the filtered signals. The beamformer output is then

$$\begin{aligned} z(t) &= \sum_{m=1}^{M} \mathbf{h}_m^H \mathbf{y}_m(t) \\ &= \underline{\mathbf{h}}^H \underline{\mathbf{y}}(t) \\ &= x_{\mathrm{fd}}(t) + v_{\mathrm{rn}}(t), \end{aligned} \tag{17}$$

where

$$\underline{\mathbf{h}} = \begin{bmatrix} \mathbf{h}_1^T & \mathbf{h}_2^T & \cdots & \mathbf{h}_M^T \end{bmatrix}^T \tag{18}$$

is the spatiotemporal linear filter of length ML, with $\mathbf{h}_m$, $m = 1, 2, \ldots, M$ being the temporal filters of length L,

$$\begin{aligned} x_{\mathrm{fd}}(t) &= \sum_{m=1}^{M} \mathbf{h}_m^H \mathbf{D}_{m,N}(\theta_0, \omega_0) \mathbf{a}(t, \omega_0) \\ &= \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{a}(t, \omega_0) \end{aligned} \tag{19}$$

is the filtered desired signal, and

$$\begin{aligned} v_{\mathrm{rn}}(t) &= \sum_{m=1}^{M} \mathbf{h}_m^H \mathbf{v}_m(t) \\ &= \underline{\mathbf{h}}^H \underline{\mathbf{v}}(t) \end{aligned} \tag{20}$$

is the residual noise. We deduce that the variance of $z(t)$ is

$$\begin{aligned} \sigma_z^2 &= \underline{\mathbf{h}}^H \mathbf{R}_{\underline{\mathbf{y}}} \underline{\mathbf{h}} \\ &= \sigma_{x_{\mathrm{fd}}}^2 + \sigma_{v_{\mathrm{rn}}}^2, \end{aligned} \tag{21}$$

where

$$\sigma_{x_{\mathrm{fd}}}^2 = \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{h}} \tag{22}$$

is the variance of $x_{\mathrm{fd}}(t)$ and

$$\sigma_{v_{\mathrm{rn}}}^2 = \underline{\mathbf{h}}^H \mathbf{R}_{\underline{\mathbf{v}}} \underline{\mathbf{h}} \tag{23}$$

is the variance of $v_{\mathrm{rn}}(t)$.

4. Performance Measures

In this section, we derive some very useful performance measures that are needed not only for the derivation of different kind of beamformers but also for their evaluation. The performance measures are special cases of the well-known general expressions in [1, 30] by using the harmonic model. We parameterize the signal correlation matrix, and discuss the noise reduction performance, as well as the speech distortion performance, and the mean-squared error (MSE) criterion. We show how the MSE is naturally related to all second-order performance measures.

4.1. Noise Reduction

Since microphone 1 is the reference, the input signal-to-noise ratio (SNR) is computed from the first L components of $\underline{\mathbf{y}}(t)$ as defined in (16), i.e., $\mathbf{y}_1(t) = \mathbf{D}_{1,N}(\theta_0, \omega_0)\mathbf{a}(t, \omega_0) + \mathbf{v}_1(t)$. We easily find that

$$\begin{aligned}\text{iSNR} &= \frac{\text{tr}\left[\mathbf{D}_{1,N}(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\mathbf{D}_{1,N}^H(\theta_0, \omega_0)\right]}{\text{tr}\left(\mathbf{R}_{\mathbf{v}_1}\right)} \\ &= \frac{\mathbf{1}_{2N}^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_{2N}}{\sigma_{v_1}^2},\end{aligned} \tag{24}$$

where $\text{tr}(\cdot)$ denotes the trace of a square matrix, $\mathbf{R}_{\mathbf{v}_1}$ is the correlation matrix of $\mathbf{v}_1(t)$, and $\sigma_{v_1}^2$ is the variance of $v_1(t)$.

The output SNR is obtained from (21). It is given by

$$\begin{aligned}\text{oSNR}\left(\underline{\mathbf{h}}\right) &= \frac{\sigma_{x_{\text{fd}}}^2}{\sigma_{v_{\text{rn}}}^2} \\ &= \frac{\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta_0, \omega_0)\underline{\mathbf{h}}}{\sigma_{v_1}^2\underline{\mathbf{h}}^H\mathbf{\Gamma}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}},\end{aligned} \tag{25}$$

where $\mathbf{\Gamma}_{\underline{\mathbf{v}}} = \mathbf{R}_{\underline{\mathbf{v}}}/\sigma_{v_1}^2$ is the pseudo-correlation matrix of $\underline{\mathbf{v}}(t)$. We see from (25) that the gain in SNR is

$$\begin{aligned}\mathcal{G}\left(\underline{\mathbf{h}}\right) &= \frac{\text{oSNR}\left(\underline{\mathbf{h}}\right)}{\text{iSNR}} \\ &= \frac{1}{\mathbf{1}_{2N}^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_{2N}} \times \frac{\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta_0, \omega_0)\underline{\mathbf{h}}}{\underline{\mathbf{h}}^H\mathbf{\Gamma}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}}.\end{aligned} \tag{26}$$

The white noise gain (WNG), $\mathcal{W}(\underline{\mathbf{h}})$, is obtained by taking $\mathbf{\Gamma}_{\underline{\mathbf{v}}} = \mathbf{I}_{ML}$, where $\mathbf{I}_{ML}$ is the $ML \times ML$ identity matrix.

The noise reduction factor quantifies the amount of noise being attenuated by the beamformer. This quantity is defined as the ratio of the power of the original noise over the power of the noise remaining after filtering, i.e.,

$$\xi_{\mathrm{nr}}(\underline{\mathbf{h}}) = \frac{\mathrm{tr}(\mathbf{R}_{\mathbf{v}_1})}{L\sigma^2_{v_{\mathrm{rn}}}} = \frac{1}{\underline{\mathbf{h}}^H \mathbf{\Gamma}_{\underline{\mathbf{v}}} \underline{\mathbf{h}}}. \tag{27}$$

For optimal filters, it is desired that $\xi_{\mathrm{nr}}(\underline{\mathbf{h}}) \geq 1$.

4.2. Speech Distortion

The desired speech signal can be distorted by the beamformer. Therefore, the speech reduction factor is defined as

$$\begin{aligned}\xi_{\mathrm{sr}}(\underline{\mathbf{h}}) &= \frac{\mathrm{tr}(\mathbf{R}_{\mathbf{x}_1})}{L\sigma^2_{x_{\mathrm{fd}}}} \\ &= \frac{\mathbf{1}_{2N}^T \mathbf{R}_{\mathbf{a}} \mathbf{1}_{2N}}{\underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0,\omega_0) \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0,\omega_0) \underline{\mathbf{h}}}.\end{aligned} \tag{28}$$

For optimal filters, it is preferred that $\xi_{\mathrm{sr}}(\underline{\mathbf{h}}) \geq 1$. In the distortionless case, we have $\xi_{\mathrm{sr}}(\underline{\mathbf{h}}) = 1$. Hence, a beamformer that does not affect the desired signal requires the constraint:

$$\underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0,\omega_0) = \mathbf{1}_{2N}^T. \tag{29}$$

It is clear that we always have

$$\mathcal{G}(\underline{\mathbf{h}}) = \frac{\xi_{\mathrm{nr}}(\underline{\mathbf{h}})}{\xi_{\mathrm{sr}}(\underline{\mathbf{h}})}. \tag{30}$$

The distortion can also be measured with the speech distortion index:

$$\begin{aligned}\upsilon_{\mathrm{sd}}(\underline{\mathbf{h}}) &= L\frac{E\left[|x_{\mathrm{fd}}(t) - x_1(t)|^2\right]}{\mathrm{tr}(\mathbf{R}_{\mathbf{x}_1})} \\ &= \frac{E\left[\left|\underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0,\omega_0)\mathbf{a}(t,\omega_0) - \mathbf{1}_{2N}^T \mathbf{a}(t,\omega_0)\right|^2\right]}{\mathbf{1}_{2N}^T \mathbf{R}_{\mathbf{a}} \mathbf{1}_{2N}}. \\ &= \frac{\left[\underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0,\omega_0) - \mathbf{1}_{2N}^T\right] \mathbf{R}_{\mathbf{a}} \left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\underline{\mathbf{h}} - \mathbf{1}_{2N}\right]}{\mathbf{1}_{2N}^T \mathbf{R}_{\mathbf{a}} \mathbf{1}_{2N}}.\end{aligned} \tag{31}$$

It has been proven in [31] that $0 \leq \upsilon_{\mathrm{sd}}(\underline{\mathbf{h}}) \leq 1$, and a value of $\upsilon_{\mathrm{sd}}(\underline{\mathbf{h}})$ close to 0 is preferred for optimal filters.

4.3. Mean-Squared Error Criterion

We define the error signal between the estimated and desired signals as

$$e(t) = z(t) - x_1(t) = e_{\mathrm{ds}}(t) + e_{\mathrm{rs}}(t), \tag{32}$$

where

$$e_{\mathrm{ds}}(t) = x_{\mathrm{fd}}(t) - x_1(t) = \left[\underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) - \mathbf{1}_{2N}^T\right] \mathbf{a}(t, \omega_0) \tag{33}$$

represents the signal distortion and $e_{\mathrm{rs}}(t) = v_{\mathrm{rn}}(t)$ represents the residual noise. We deduce that the mean-squared error (MSE) criterion is

$$\begin{aligned} J\left(\underline{\mathbf{h}}\right) = & \ E\left[\left|e(t)\right|^2\right] & (34)\\ = & \ \mathbf{1}_{2N}^T \mathbf{R}_{\mathbf{a}} \mathbf{1}_{2N} & (35)\\ & + \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{h}} \\ & - \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \mathbf{1}_{2N} \\ & - \mathbf{1}_{2N}^T \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{h}} + \underline{\mathbf{h}}^H \mathbf{R}_{\underline{\mathbf{v}}} \underline{\mathbf{h}}. \end{aligned}$$

Since $E\left[e_{\mathrm{ds}}(t) e_{\mathrm{rs}}^*(t)\right] = 0$, $J\left(\underline{\mathbf{h}}\right)$ can also be expressed as

$$\begin{aligned} J\left(\underline{\mathbf{h}}\right) &= E\left[\left|e_{\mathrm{ds}}(t)\right|^2\right] + E\left[\left|e_{\mathrm{rs}}(t)\right|^2\right] \\ &= J_{\mathrm{ds}}\left(\underline{\mathbf{h}}\right) + J_{\mathrm{rs}}\left(\underline{\mathbf{h}}\right), \end{aligned} \tag{36}$$

where $J_{\mathrm{ds}}\left(\underline{\mathbf{h}}\right) = \mathrm{tr}\left(\mathbf{R}_{\mathbf{x}_1}\right) \upsilon_{\mathrm{sd}}\left(\underline{\mathbf{h}}\right) / L$, and $J_{\mathrm{rs}}\left(\underline{\mathbf{h}}\right) = \mathrm{tr}\left(\mathbf{R}_{\mathbf{v}_1}\right) / L \xi_{\mathrm{nr}}\left(\underline{\mathbf{h}}\right)$. Finally, we have

$$\begin{aligned} \frac{J_{\mathrm{ds}}\left(\underline{\mathbf{h}}\right)}{J_{\mathrm{rs}}\left(\underline{\mathbf{h}}\right)} &= \mathrm{iSNR} \times \xi_{\mathrm{nr}}\left(\underline{\mathbf{h}}\right) \times \upsilon_{\mathrm{sd}}\left(\underline{\mathbf{h}}\right) \\ &= \mathrm{oSNR}\left(\underline{\mathbf{h}}\right) \times \xi_{\mathrm{sr}}\left(\underline{\mathbf{h}}\right) \times \upsilon_{\mathrm{sd}}\left(\underline{\mathbf{h}}\right). \end{aligned} \tag{37}$$

This shows how the MSEs are related to the most fundamental performance measures.

5. Harmonic Model-based Beamforming

Based on the harmonic signal model and the performance measures as reported in the previous sections, this section presents the derivations of the proposed harmonic beamformers for noise reduction and dereverberation.

5.1. Fixed Beamformers

The first harmonic beamformers considered are fixed beamformers. While these cannot adapt to the spatial characteristics of the noise, they are computationally efficient and very practical in the sense that they do not require estimates of second-order statistics.

5.1.1. Delay-and-Sum

The delay-and-sum (DS) beamformer is obtained by maximizing the WNG subject to the distortionless constraint, i.e.,

$$\min_{\underline{\mathbf{h}}} \underline{\mathbf{h}}^H \underline{\mathbf{h}} \quad \text{subject to} \quad \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) = \mathbf{1}_{2N}^T. \tag{38}$$

We deduce that the optimal solution is

$$\underline{\mathbf{h}}_{\mathrm{DS-HB}} = \underline{\mathbf{D}}_N(\theta_0, \omega_0) \left[\underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{D}}_N(\theta_0, \omega_0) \right]^{-1} \mathbf{1}_{2N}. \tag{39}$$

As a result, the WNG is

$$\mathcal{W}\left(\underline{\mathbf{h}}_{\mathrm{DS-HB}}\right) = \frac{1}{\mathbf{1}_{2N}^T \left[\underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{D}}_N(\theta_0, \omega_0) \right]^{-1} \mathbf{1}_{2N}}. \tag{40}$$

In the presence of spatially white noise, the DS beamformer is optimal in the sense that it gives the maximum gain in SNR without distorting the desired signal. However, in the presence of other noises, we should not expect very high gains. Moreover, we can obtain $\lim_{ML \to \infty} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \underline{\mathbf{D}}_N(\theta_0, \omega_0) = ML \times \mathbf{I}_{2N}$. Therefore, the WNG of the DS-HB depends directly to both M and L, i.e., $\mathcal{W}\left(\underline{\mathbf{h}}_{\mathrm{DS-HB}}\right) \to ML/2N$.

5.1.2. Null Forming

Let us assume that there is a broadband interference with fundamental frequency ω_1 and model order N_1 in the direction θ_1. The matrix $\underline{\mathbf{D}}_{N_1}(\theta_1, \omega_1)$ of size $ML \times 2N_1$ is associated with this interference.

Now, we would like to perfectly recover the desired signal and completely cancel the interference. The constraint is then

$$\underline{\mathbf{h}}^H \underline{\mathbf{C}} = \left[\begin{array}{cc} \mathbf{1}_{2N}^T & \mathbf{0}_{2N_1}^T \end{array} \right], \tag{41}$$

where

$$\underline{\mathbf{C}} = \left[\begin{array}{cc} \underline{\mathbf{D}}_N(\theta_0, \omega_0) & \underline{\mathbf{D}}_{N_1}(\theta_1, \omega_1) \end{array} \right] \tag{42}$$

is the constraint matrix of size $ML \times 2(N + N_1)$ and $\mathbf{0}_{2N_1}$ is the zero vector of length $2N_1$. Then, our criterion is

$$\min_{\underline{\mathbf{h}}} \underline{\mathbf{h}}^H \underline{\mathbf{h}} \quad \text{subject to} \quad \underline{\mathbf{h}}^H \underline{\mathbf{C}} = \left[\begin{array}{cc} \mathbf{1}_{2N}^T & \mathbf{0}_{2N_1}^T \end{array} \right], \tag{43}$$

from which we find the optimal solution:

$$\underline{\mathbf{h}}_{\mathrm{NF-HB}} = \underline{\mathbf{C}} \left(\underline{\mathbf{C}}^H \underline{\mathbf{C}} \right)^{-1} \left[\begin{array}{c} \mathbf{1}_{2N} \\ \mathbf{0}_{2N_1} \end{array} \right]. \tag{44}$$

Obviously, we must have $ML > 2(N+N_1)$. The generalization of this approach to any number of interferences is straightforward.

5.2. Adaptive Beamformers

Having considered different fixed harmonic beamforming techniques, this subsection deals with a class of adaptive beamformers, where some signal statistics need to be estimated. In theory, adaptive beamformers give better noise reduction results than fixed beamformers since they can adjust to the spatial characteristics of the noise.

5.2.1. Wiener

The harmonic model-based Wiener beamformer is easily derived by taking the gradient of the MSE, $J(\underline{\mathbf{h}})$ [eq. (34)], with respect to $\underline{\mathbf{h}}$ and equating the result to zero:

$$\underline{\mathbf{h}}_{\mathrm{W-HB}} = \left[\underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) + \mathbf{R}_{\underline{\mathbf{v}}} \right]^{-1} \underline{\mathbf{D}}_N(\theta_0, \omega_0) \mathbf{R}_{\mathbf{a}} \mathbf{1}_{2N}. \tag{45}$$

Determining the matrix inverse with the Woodbury identity leads to another interesting formulation of the harmonic model-based Wiener beamformer:

$$\begin{aligned} \underline{\mathbf{h}}_{\mathrm{W-HB}} &= \mathbf{R}_{\underline{\mathbf{v}}}^{-1} \underline{\mathbf{D}}_N(\theta_0, \omega_0) \left[\mathbf{R}_{\mathbf{a}}^{-1} + \underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \mathbf{R}_{\underline{\mathbf{v}}}^{-1} \underline{\mathbf{D}}_N(\theta_0, \omega_0) \right]^{-1} \mathbf{1}_{2N} \\ &= \mathbf{R}_{\underline{\mathbf{v}}}^{-1} \underline{\mathbf{D}}_N(\theta_0, \omega_0) \left[\underline{\mathbf{D}}_N^H(\theta_0, \omega_0) \mathbf{R}_{\underline{\mathbf{v}}}^{-1} \underline{\mathbf{D}}_N(\theta_0, \omega_0) \right]^{-1} \mathbf{P}(\theta_0, \omega_0) \mathbf{1}_{2N}, \end{aligned} \tag{46}$$

where

$$\mathbf{P}\left(\theta_0,\omega_0\right)=\left(\mathbf{R}_{\mathbf{a}}^{-1}\left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\right]^{-1}+\mathbf{I}_{2N}\right)^{-1}.$$

In spatially white noise, we can approximate $\mathbf{P}\left(\theta_0,\omega_0\right)$ as

$$\mathbf{P}=\left(\frac{\sigma_{v_1}^2}{ML}\mathbf{R}_{\mathbf{a}}^{-1}+\mathbf{I}_{2N}\right)^{-1} \tag{47}$$

for a large filter, i.e., $ML\rightarrow\infty$.

5.2.2. Minimum Variance Distortionless Response

The celebrated minimum variance distortionless response (MVDR) beamformer proposed by Capon [7, 16] is easily derived by optimizing the following criterion:

$$\min_{\underline{\mathbf{h}}}\underline{\mathbf{h}}^H\mathbf{R}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}\quad\text{subject to}\quad\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0,\omega_0)=\mathbf{1}_{2N}^T. \tag{48}$$

We obtain

$$\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}=\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\right]^{-1}\mathbf{1}_{2N}. \tag{49}$$

The perfectly matched beamformer to the signal parameters results in

$$\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}^H\mathbf{R}_{\underline{\mathbf{y}}}\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}=\mathbf{1}_{2N}^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_{2N}+\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}^H\mathbf{R}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}_{\mathrm{MVDR-HB}}. \tag{50}$$

Therefore, minimizing the residual noise is equivalent to minimizing the noisy signal, i.e., $\underline{\mathbf{h}}^H\mathbf{R}_{\underline{\mathbf{y}}}\underline{\mathbf{h}}$, and we can express the MVDR beamformer alternatively as the minimum power distortionless response (MPDR) beamformer [32]. We obtain the MPDR beamformer interestingly by exploiting the correlation matrix of the noisy signals as

$$\underline{\mathbf{h}}_{\mathrm{MPDR-HB}}=\mathbf{R}_{\underline{\mathbf{y}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\left[\underline{\mathbf{D}}_N^H(\theta_0,\omega_0)\mathbf{R}_{\underline{\mathbf{y}}}^{-1}\underline{\mathbf{D}}_N(\theta_0,\omega_0)\right]^{-1}\mathbf{1}_{2N}. \tag{51}$$

We can identify the harmonic model-based Wiener beamformer in (46) as the weighted MVDR beamformer in (49). The diagonal weight matrix $\mathbf{P}\left(\theta_0,\omega_0\right)$ is related to the narrowband input SNRs of the harmonics. Therefore, we can conclude that the MVDR and Wiener beamformers are approximately equivalent

in high input SNRs. Moreover, it has also been shown in [9] that we always have a trade-off in noise reduction and speech distortion index between the MVDR and Wiener beamformers, i.e.,

$$\text{oSNR}\left(\underline{\mathbf{h}}_{\text{W}-\text{HB}}\right) \geq \text{oSNR}\left(\underline{\mathbf{h}}_{\text{MVDR}-\text{HB}}\right) \geq \text{iSNR}, \tag{52}$$

$$\upsilon_{\text{sd}}\left(\underline{\mathbf{h}}_{\text{W}-\text{HB}}\right) \geq \upsilon_{\text{sd}}\left(\underline{\mathbf{h}}_{\text{MVDR}-\text{HB}}\right) = 0, \tag{53}$$

$$\xi_{\text{sr}}\left(\underline{\mathbf{h}}_{\text{W}-\text{HB}}\right) \geq \xi_{\text{sr}}\left(\underline{\mathbf{h}}_{\text{MVDR}-\text{HB}}\right) = 1. \tag{54}$$

5.2.3. Linearly Constrained Minimum Variance

We can derive a linearly constrained minimum variance (LCMV) beamformer [3, 33], which can handle more than one linear constraint, by exploiting the nullspace of the desired signal correlation matrix. Again, we assume the presence of a unique interference as explained in Subsection 5.1.2. The criterion to be optimized is now

$$\min_{\underline{\mathbf{h}}} \underline{\mathbf{h}}^H \mathbf{R}_{\underline{\mathbf{v}}} \underline{\mathbf{h}} \quad \text{subject to} \quad \underline{\mathbf{h}}^H \underline{\mathbf{C}} = \begin{bmatrix} \mathbf{1}_{2N}^T & \mathbf{0}_{2N_1}^T \end{bmatrix}, \tag{55}$$

where $\underline{\mathbf{C}}$ is defined in Subsection 5.1.2. We obtain

$$\underline{\mathbf{h}}_{\text{LCMV}-\text{HB}} = \mathbf{R}_{\underline{\mathbf{v}}}^{-1} \underline{\mathbf{C}} \left(\underline{\mathbf{C}}^H \mathbf{R}_{\underline{\mathbf{v}}}^{-1} \underline{\mathbf{C}}\right)^{-1} \begin{bmatrix} \mathbf{1}_{2N} \\ \mathbf{0}_{2N_1} \end{bmatrix}. \tag{56}$$

While the LCMV beamformer completely cancels the interference, there is no guarantee that the output SNR is greater than the input SNR [34]. The generalization of this LCMV beamformer to any number of interferences is straightforward, as long as the filter length ML is larger than the number of constraints. Now, we can express the linearly constrained minimum power (LCMP) beamformer, which utilizes the correlation matrix of the noisy signals, by the following equation:

$$\underline{\mathbf{h}}_{\text{LCMP}-\text{HB}} = \mathbf{R}_{\underline{\mathbf{y}}}^{-1} \underline{\mathbf{C}} \left(\underline{\mathbf{C}}^H \mathbf{R}_{\underline{\mathbf{y}}}^{-1} \underline{\mathbf{C}}\right)^{-1} \begin{bmatrix} \mathbf{1}_{2N} \\ \mathbf{0}_{2N_1} \end{bmatrix}. \tag{57}$$

Although the MVDR/LCMV and the MPDR/LCMP beamformers are theoretically the same, an inaccurate estimate of the correlation matrix in practice

causes mismatch between the actual and the presumed signal in the MPDR/LCMP beamformers. Furthermore, the MVDR/LCMV beamformers are more robust to DOA estimation errors than the MPDR/LCMP beamformers [32, 35]. Therefore, for the sake of the maximum WNG, we can add the minimum filter norm constraint as $\underline{\mathbf{h}}^H\underline{\mathbf{h}} \leq k$ to the beamformers in addition to the distortionless constraints, where k is a positive constant. This modification corresponds to the so-called diagonal loading approach [16, 32, 36] which is given by $\mathbf{R}_{\underline{\mathbf{y}}} \leftarrow \mathbf{R}_{\underline{\mathbf{y}}} + \lambda\mathbf{I}_{ML}$, where λ is a positive constant. In general, the diagonal loading technique is applied to improve the performance of the beamformers with errors on the signal parameters (i.e., the DOA and frequency) and an inaccurate estimation of the correlation matrix.

5.3. Relation to Broadband Beamforming

The harmonic model used in deriving the parametric beamformers is well-suited for modeling periodic signals using a few harmonic frequencies, but it can also model general broadband signals as a special case. To achieve this, we can set the fundamental frequency to $\omega_0 = 2\pi/L$ and the number of harmonics to $N = \lfloor L/2 \rfloor$. With this choice, we can compute non-parametric version of the beamformer designs in the Sections 5.1 and 5.2 that does not rely on a temporal model of the desired signal. That is, the broadband beamformers are special cases of the proposed harmonic beamformers in contrast to common belief. While these broadband beamformers do not have potential problems with model mismatch, they can only yield spatial selectivity and not spectral selectivity as opposed to the proposed harmonic beamformers. The spectral selectivity that can be obtained with the proposed beamformers allows for reducing noise in between the harmonic frequencies, such as spectral smearing introduced by reverberation. The relationship between harmonic and broadband beamforming makes it easy to apply the proposed beamformers to signals containing both stochastic broadband parts and deterministic harmonic parts like speech, e.g., by using an voiced/unvoiced speech detector [37]. When voiced speech is detected, the estimated fundamental frequency is used to compute and apply the harmonic

Table 1: List of parameters used for RIR generation.

Parameter	**Value**	**Unit**
Sound speed	343	m/s
Reverberation time (T_{60})	0.5	s
RIR length	2048	samples
Room dimensions	[8, 6, 4]	m
Microphone spacing	5	cm
No. of microphones	3	
Array center	[3.5, 1, 1]	m
Microphone directivity	Omni	
Direction-of-arrival	150	$^\circ$
Range (in positive y-dir)	2	m

Table 2: List of parameters for computing the beamformers.

Parameter	**Value**	**Unit**
Segment length, L	20	ms
Time hop, T_{hop}	10	ms
Forgetting factor, α	0.05	
Smoothing parameter, β	0.98	
Regularization parameter, γ	10^{-6}	
Maximum no. of harmonics, N_{max}	15	

beamformers, while the broadband version is used for the unvoiced parts.

6. Experimental Results

In this section, we investigate the merits of the parametric beamformers presented in Sections 5.1 and 5.2 in relation to the more conventional broadband beamformers, e.g., those described in Section 5.3, but also in relation to a state-of-the-art dereverberation method, i.e., the weighted prediction error method using iteratively reweighted least squares (DR) in [38]. First, we provide some qualitative and illustrative examples of the differences between these approaches and their performance, which are then followed by thorough and quantitative experiments on both synthetic and real recorded data to uncover the general behaviour of the proposed beamformers.

6.1. Implementation Details

Before presenting the evaluations, this section provide an overview of the experimental setup. First, evaluation on synthetically generated data was con-

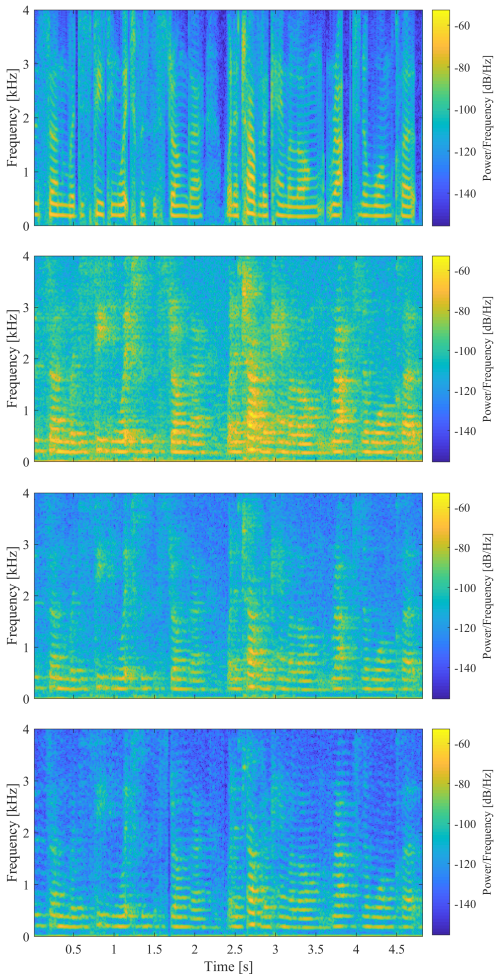

Figure 1: Spectrograms of (from top to bottom) the desired speech, the noisy speech with reverberation, diffuse noise, and thermal noise, the signal enhanced using the broadband DSB, and the signal enhanced using the harmonic DSB.

ducted, as this enabled us to accurately measure the performance of the different beamformers by having access to the individual speech and noise signals. The speech signals used for these evaluations were two male and two female speech signals comprising 20 seconds of speech in total. These signals, which are single-channel signals, were then synthesized spatially using room impulse responses (RIRs) obtained with a RIR generator based on the image source method [39]. The setup of the RIR generator is provided in Table 1. In addition to this,

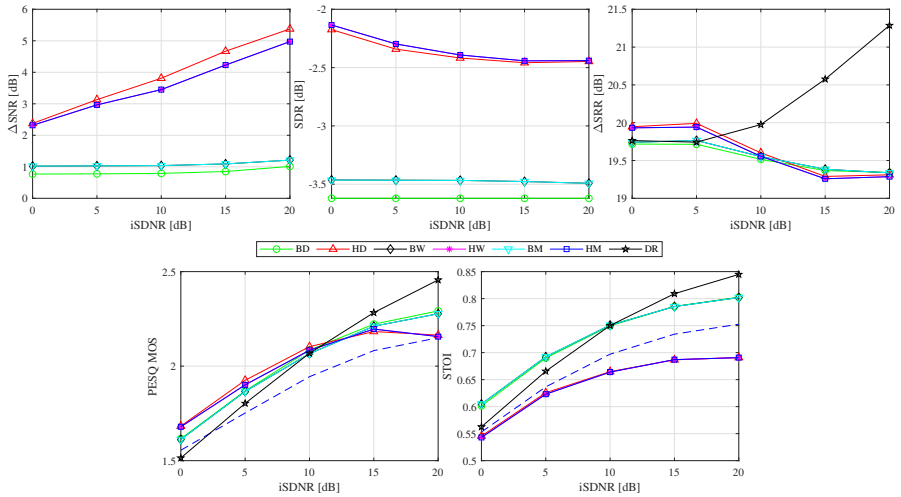

Figure 2: Evaluation of the broadband beamformers, the proposed parametric beamformers, and a dereverberation method [38] on synthetically generated data as a function of the input signal-to-diffuse-noise ratio.

two types of noise were added to the synthetic, multichannel speech signals: spherically isotropic (diffuse) babble noise and thermal sensor noise. The diffuse noise was generated using an online available noise generator implementing the algorithm in [40]. It was then added to the speech signals such that the signal-to-diffuse-noise ratio (SDNR) was 10 dB at the reference microphone, while the thermal noise was assumed to be white Gaussian, and it was added at a signal-to-thermal-noise ratio (STNR) of 30 dB at each microphone. In addition to this, evaluations on real speech data were carried out. For these experiments, female and male speech from the single- and multichannel audio recordings database (SMARD) [41] were used, more specifically the signals labeled FA03_09 and MD24_04, respectively. The two scenarios labelled 1011 and 1111 were considered, and, for each scenario, the two ULAs, A and B, were used for the evaluation. For each of these arrays, we used the three first microphones.

To implement the adaptive beamformers, an estimate of the multichannel noise covariance matrix is needed. In this paper, we focus on comparing the proposed parametric beamformers, with the more traditional broadband beamformers. Therefore, the speech direction-of-arrival, θ, was assumed known, and

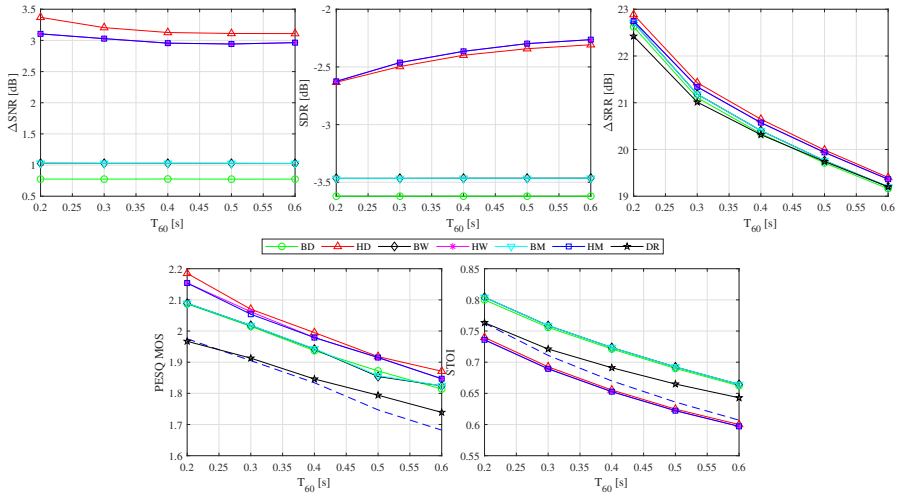

Figure 3: Evaluation of the broadband beamformers, the proposed parametric beamformers, and a dereverberation method [38] on synthetically generated data with respect to the reverberation time (T_{60}).

the noise statistics were simply estimated intrusively from the actual noise signals for both the broadband and parametric beamformers using a recursive update:

$$\widehat{\mathbf{R}}_{\mathbf{v}}(t) = \alpha \widehat{\mathbf{R}}_{\mathbf{v}}(t - T_{\text{hop}}) - (1 - \alpha)\underline{\mathbf{v}}(t)\underline{\mathbf{v}}^{T}(t). \tag{58}$$

As indicated by the update formula, the recursive update and the beamformers (except for the broadband DSB) were only computed every T_{hop} samples to ease the computational burden. Additionally, due to the large dimensions of the matrices involved in the filter designs, all matrices to be inverted ($\mathbf{X} \in \mathbb{R}^{K \times K}$) was regularized according to

$$\widetilde{\mathbf{X}} = \mathbf{X} + \gamma \frac{\text{tr}(\mathbf{X})}{K}\mathbf{I}, \tag{59}$$

where γ is the regularization parameter. Then, to perceptually reduce the effects of the beamformers not being computed every sample, they were smoothed every sample before application as

$$\widehat{\underline{\mathbf{h}}}(t) = \beta \widehat{\underline{\mathbf{h}}}(t - 1) + (1 - \beta)\underline{\mathbf{h}}, \tag{60}$$

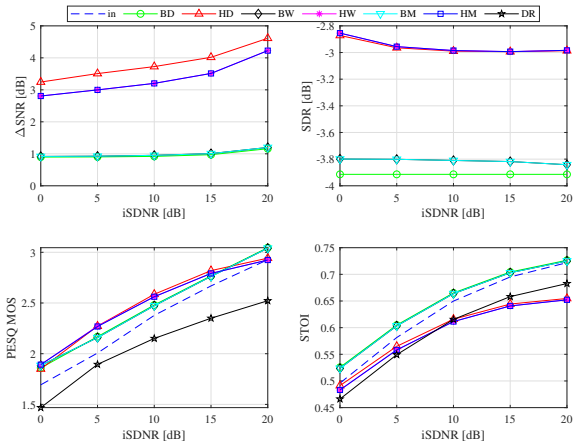

Figure 4: Evaluation of the broadband beamformers, the proposed parametric beamformers, and a dereverberation method [38] on SMARD data with respect to the input signal-to-diffuse-noise ratio.

where $\underline{\mathbf{h}}$ is the most recently computed beamformer and β is the smoothing parameter. An overview of the parameters used for the computation of the beamformers are provided in Table 2.

The tested beamformers are harmonic and broadband DSB (referred to in the figures as HD and BD, resp.), harmonic and broadband Wiener (referred to as HW and BW, resp.), and harmonic and broadband MVDR (referred to as HM and BM, resp.). For the proposed, harmonic beamformers, we also need estimates of the fundamental frequency, ω_0, and the number of harmonic components of the speech signal for every processed segment of speech. The fundamental frequency was estimated from the noisy recordings from the reference microphone using the fast fundamental frequency estimator described in [42, 43]. The model order was also estimated as part of the fundamental frequency estimation, but in the computation of the beamformers it was replaced by $\widehat{N} = \min(N_{\max}, \lfloor \pi/\widehat{\omega}_0 \rfloor)$ to reduce the distortion of the higher harmonics. Finally, the WPE-IRLS method (unregularized version) [38] included for comparison was implemented as follows: the ϵ value was set to $1 \cdot 10^{-8}$, the shape parameter was $p = 0.5$, and a fixed number of 10 iterations was run for each frequency bin.

6.2. Qualitative Experiments

The first results presented illustrate the difference between the broadband and harmonic beamformers. These results were obtained by applying the broadband and harmonic DSBs to one of the synthetically spatialized female speech signals mentioned before. The signal was then added with diffuse babble noise at a 20 dB iSDNR and reverberation with a reverberation time of 0.7 s. We then applied the two beamformers to this signal to reduce the effects of the noise. The spectrograms of all the signals are depicted in Figure 1. First of all, the spectrograms clearly indicate that the two beamformers reduce the effects of the noise. If we then compare the broadband and harmonic beamformers, we can see that the harmonic DSB seems to provide more noise reduction, especially in the high frequency region, while preserving the harmonic components of the speech. Moreover, the spectrograms indicate that the effects of reverberation are better mitigated with the harmonic DSB, with the harmonics being less smeared, e.g., in the time span from 2.5 s to 3 s.

6.3. Quantitivate Experiments

6.3.1. Synthetic Data

To support and strengthen the observations from the qualitative experiments, we conducted extensive evaluations of the proposed, parametric beamformers over various settings, set ups, and speech signals. The objective measures used to quantify the performance are the signal-to-noise ratio (SNR), the speech distortion index (SDI), the segmental speech-to-reverberation ratio (SRR) [44], the perceptual evaluation of speech quality (PESQ) scores [45], and the short-time objective intelligibility (STOI) measure [46]. In the computation of these measures, the direct speech component is considered the desired signal, while the addition of the diffuse and thermal noise components are considered as the noise signal. Moreover, the SRR measures are computed using the least squares level normalization proposed in [44] to reduce effects of signal distortion on this measure. The dereverberation method was only evaluated in terms of SRR, PESQ, and STOI, since it was not derived for noise reduction. For each

setting, the performance measures were averaged over all the speech signals of the evaluation. In this way, the performance measures were first computed versus the input SDNR on the reference microphone, yielding the results in Figure 2. First of all, we observe, through close inspection, that the Wiener and MVDR beamformers only provide slightly different performances, but they follow the same trend for all performance measures. Looking at the SNR measures, it is obvious that there is a higher SNR gain, measured with respect to the input SNR, when using the proposed, parametric beamformers over the broadband beamformers for all the considered iSDNRs. We observe that the harmonic Wiener and MVDR beamformers provide a slightly lower SNR gain compared to the harmonic DSB. While the adaptive beamformers in theory should perform better, they are relying on estimates of the noise statistics, resulting in a slightly lower practical performance. However, with directional noise components the benefit of using the adaptive beamformers is expected to be larger. The SDR measures shows that the harmonic beamformers gives more signal distortion, which is expected since there will be some practical model mismatch, e.g., due to fundamental frequency estimation errors. However, the results show that the proposed beamformers yield more suppression of reverberation compared the broadband ones measured in terms of the SRR gain. Moreover, compared to the dereverberation method, the parametric beamformers achieves higher SRR gain in noisy scenarios, i.e., for iSDNRs below 10 dB. The perceptual, but objective, measures indicate that, for low iSDNRs, the harmonic beamformers can give provide enhanced signals of a better quality, but the intelligibility is in general better when using the broadband beamformers. This suggest that it might be beneficial to combine the two types of beamforming to be able to control a trade off between quality and intelligibility. In comparison with the dereverberation, there is a tipping point around an iSDNR of 10 dB. Below this, better PESQ and STOI scores can be obtained with the beamformers, while the dereverberation method achieves better scores for higher iSDNRs.

To investigate further the abilities of the proposed beamformers to combat reverberation, we measured the performance again, but versus the reverberation

time (T_{60}). The results of this evaluation are depicted in Figure 3. First, they indicate that the SNR gain is not affected much by the reverberation, while the distortion increases slightly for the harmonic beamformers, when the reverberation time increases. Moreover, we see that for higher reverberation times, i.e., over 0.4 s, the harmonic beamformers provide more reverberation reduction compared to the broadband beamformers. For the considered iSDNR, the dereverberation method is generally outperformed in terms of SRR improvement by the beamformers, except for higher T_60's where it achieves comparable performance to the broadband ones. When it comes to the objective, perceptual measures, the quality is generally improved with all the beamformers, but slightly more with the proposed, harmonic beamformers for all reverberation times. The dereverberation provides the lowest PESQ scores for all reverberation times. The STOI scores indicate that only the broadband beamformers and the dereverberation method can generally improve the intelligibility, and particularly so for higher reverberation times, where the harmonic beamformers also provide the best STOI scores compared to the STOI scores of the noisy observations. This is in line with the SDR measurements, which showed that the harmonic beamformers yields more distortion of the desired speech, compared to the broadband beamformers. Moreover, it is supported by our informal listening tests, in which the distortion incurred by the mismatch between the harmonic model and the speech signal, becomes more noticeable for higher input SNRs. In other words, the harmonic beamformers are useful for improving the speech quality in low SNR conditions, through better noise reduction and dereverberation compared to the broadband beamformers. However, at this point it should be mentioned that the performance of the harmonic beamforming approach can be improved in a number of ways. For example, the model order, N, used in the harmonic beamforming is chosen to be the highest possible one for the estimated fundamental frequency within each STFT frame, but by choosing it adaptively depending on the number of actual harmonics it is possible to obtain further noise reduction. Secondly, the fundamental frequency estimates can be postprocessed (e.g., using smoothing tecniques) to reduce the

number of spurious estimates, which should result in less speech distortion. Finally, in our implementation of the harmonic beamformers, they are used on all parts of the speech, including unvoiced ones. Instead, it would be possible to switch between the harmonic and broadband beamformers for these parts using a voiced/unvoiced speech detector [37].

6.3.2. Real Data

In the final evaluations, the broadband and proposed beamformers were evaluated on real recorded speech from the SMARD. These recordings were then added with thermal and diffuse noise as in the previous experiments to enable us to compute the objective performance measures. In this regard, it is important to note that the performance measures are computed differently in this evaluation, since we do not have access to the clean, desired signal without reverberation. That is, we here consider the clean speech signal with reverberation at the reference microphone as the desired signal. We then conducted a series of experiments in which the SDNR was varied between 0 dB and 30 dB, and for each of considered SDNRs, the performance measures were calculated and averaged over the different speech signals and scenarios. The outcome of this evaluation is depicted in Figure 4. The results show a similar trend to those obtained on synthetic data. That is, the proposed harmonic beamformers has a higher SNR gain compared to their broadband alternatives. Distortion-wise, the broadband beamformers, however, yield better performance in terms of the SDR. Because we did not have access to the room impulse response in these experiments, the SRR was not calculated, but the methods were also compared in terms of PESQ and STOI scores. For low SDNRs (i.e., between 0 dB and 15 dB), the proposed harmonic beamformers yield enhanced signals with a better perceptual quality in terms of PESQ scores, while the broadband beamformers are preferred for higher SDNRs. This was also the case in the experiments on synthetic data, and can be explained by the fact that the modelling mismatch introduced by the harmonic model is greater than the noise reduction obtained by the beamformers at high SDNRs. In terms of STOI scores, the broadband

beamformers are generally performing better like we observed for the synthetic data experiments. The dereverberation method is generally achieving significantly lower PESQ scores than the beamforming methods, while it yields better STOI scores than the harmonic beamformers for iSDNRs greater than 10 dB.

7. Conclusions

In this paper, a new framework for beamforming has been presented, wherein the a priori knowledge about voiced speech signals and is properties have been exploited to develop model-based beamforming. This was done via a multichannel signal model that incorporates both the spatial and the spectral properties of periodic signals using the harmonic model. Based on this model, a number of fixed and adaptive beamformers have been proposed. Interestingly, these beamformers reduce to their broadband counterparts in special cases. Experiments on synthetic and real signals demonstrated the properties and good performance of the proposed harmonic model-based beamformers compared to traditional, broadband beamformers and a state-of-the-art dereverberation method. The most important observation from the experiments is that the harmonic model-based beamformers are capable of performing enhancement and dereverberation simultaneously, especially at high noise levels, where they outperform their broadband counterparts as well as the dereverberation method in terms of noise reduction, dereverberation, and PESQ scores.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This research was supported by the Villum Foundation and the Danish Council for Independent Research, grant ID: DFF - 1337-00084.

References

[1] J. Benesty, Y. Huang, J. Chen, Microphone Array Signal Processing, Vol. 1, Springer-Verlag, 2008.

[2] J. Flanagan, J. Johnston, R. Zahn, G. Elko, Computer-steered microphone arrays for sound transduction in large rooms, J. Acoust. Soc. Am. 78 (5) (1985) 1508–1518.

[3] O. L. Frost, An algorithm for linearly constrained adaptive array processing, Proc. IEEE 60 (8) (1972) 926–935.

[4] B. D. Van Veen, K. M. Buckley, Beamforming: a versatile approach to spatial filtering, IEEE ASSP Mag. 5 (2) (1988) 4–24. doi:10.1109/53.665.

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

[6] S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech, IEEE Trans. Signal Process. 49 (8) (2001) 1614–1626. doi:10.1109/78.934132.

[7] J. Capon, High-resolution frequency-wavenumber spectrum analysis, Proc. IEEE 57 (8) (1969) 1408–1418. doi:10.1109/PROC.1969.7278.

[8] K. U. Simmer, J. Bitzer, C. Marro, Post-filtering techniques, in: Microphone Arrays, Springer, 2001, pp. 39–60.

[9] J. Benesty, J. Chen, Y. Huang, Speech enhancement in the karhunen-loève expansion domain, Synthesis Lectures on Speech and Audio Processing 7 (1) (2011) 1–112.

[10] Y. Lacouture-Parodi, E. A. Habets, J. Chen, J. Benesty, Multichannel noise reduction in the Karhunen-Loève expansion domain, IEEE Trans. Acoust., Speech, Signal Process. 22 (5) (2014) 923–936.

[11] S. Doclo, M. Moonen, GSVD-based optimal filtering for single and multimicrophone speech enhancement, IEEE Trans. Signal Process. 50 (9) (2002) 2230–2244. doi:10.1109/TSP.2002.801937.

[12] Y. Hu, P. C. Loizou, A subspace approach for enhancing speech corrupted by colored noise, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Vol. 1, IEEE, 2002, pp. I–573.

[13] J. R. Jensen, J. Benesty, M. G. Christensen, Noise reduction with optimal variable span linear filters, IEEE Trans. Acoust., Speech, Signal Process. 24 (4) (2015) 631–644.

[14] A. Nehorai, B. Porat, Adaptive comb filtering for harmonic signal enhancement, IEEE Trans. Acoust., Speech, Signal Process. 34 (5) (1986) 1124–1138.

[15] T. Nakatani, M. Miyoshi, K. Kinoshita, Single-microphone blind dereverberation, in: Speech Enhancement, Springer, 2005, pp. 247–270.

[16] R. T. Lacoss, Data Adaptive Spectral Analysis Methods, Geophysics 36 (Aug. 1971) 661–675. doi:10.1190/1.1440203.

[17] M. G. Christensen, A. Jakobsson, Optimal filter designs for separating and enhancing periodic signals, IEEE Trans. Signal Process. 58 (12) (2010) 5969–5983. doi:10.1109/TSP.2010.2070497.

[18] B. Harvey, S. O'Young, A harmonic spectral beamformer for the enhanced localization of propeller-driven aircraft, J. of Unmanned Vehicle Systems 7 (2) (2019) 156–174. arXiv:https://doi.org/10.1139/juvs-2018-0011, doi:10.1139/juvs-2018-0011.
URL `https://doi.org/10.1139/juvs-2018-0011`

[19] L. Wang, A. Cavallaro, Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles, IEEE Sensors Journal 17 (8) (2017) 2447–2455. doi:10.1109/JSEN.2017.2669262.

[20] A. Schmidt, H. W. Löllmann, W. Kellermann, A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 6583–6587. doi:10.1109/ICASSP.2018.8462211.

[21] J. R. Jensen, M. G. Christensen, A. Jakobsson, Harmonic minimum mean squared error filters for multichannel speech enhancement, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2017, pp. 501–505. doi:10.1109/ICASSP.2017.7952206.

[22] J. Li, P. Stoica, An adaptive filtering approach to spectral estimation and SAR imaging, IEEE Trans. Signal Process. 44 (6) (1996) 1469–1484. doi:10.1109/78.506612.

[23] J. R. Jensen, M. G. Christensen, J. Benesty, S. H. Jensen, Joint spatio-temporal filtering methods for DOA and fundamental frequency estimation, IEEE Trans. Audio, Speech, and Language Process. 23 (1) (2015) 174–185.

[24] M. G. Christensen, A. Jakobsson, Multi-pitch estimation, Synthesis Lectures on Speech and Audio Process. 5 (1) (2009) 1–160.

[25] M. S. Brandstein, H. Silverman, A practical methodology for speech source localization with microphone arrays, Comput. Speech Language.

[26] S. Karimian-Azari, J. R. Jensen, M. G. Christensen, Computationally efficient and noise robust DOA and pitch estimation, IEEE Trans. Audio, Speech, and Language Process. 24 (9) (2016) 1609–1621. doi:10.1109/TASLP.2016.2577501.

[27] S. Karimian-Azari, J. R. Jensen, M. G. Christensen, Fast joint DOA and pitch estimation using a broadband MVDR beamformer, in: Proc. European Signal Processing Conf., 2013, pp. 1–5.

[28] J. R. Jensen, M. G. Christensen, S. H. Jensen, Nonlinear least squares methods for joint DOA and pitch estimation, IEEE Trans.

Audio, Speech, and Language Process. 21 (5) (2013) 923 –933. doi:10.1109/TASL.2013.2239290.

[29] M. G. Christensen, Accurate estimation of low fundamental frequencies from real-valued measurements, IEEE Trans. Acoust., Speech, Signal Process. 21 (10) (2013) 2042–2056.

[30] J. Chen, J. Benesty, Y. Huang, S. Doclo, New insights into the noise reduction Wiener filter, IEEE Trans. Audio, Speech, and Language Process. 14 (4) (2006) 1218–1234. doi:10.1109/TSA.2005.860851.

[31] J. Benesty, J. Chen, Y. A. Huang, S. Doclo, Study of the wiener filter for noise reduction, in: Speech Enhancement, Springer, 2005, pp. 9–41.

[32] H. Cox, Resolving power and sensitivity to mismatch of optimum array processors, J. Acoust. Soc. Am. 54 (3) (1973) 771–785. doi:10.1121/1.1913659.

[33] M.-H. Er, A. Cantoni, Derivative constraints for broad-band element space antenna array processors, Acoustics, Speech and Signal Processing, IEEE Transactions on 31 (6) (1983) 1378–1393. doi:10.1109/TASSP.1983.1164219.

[34] M. Souden, J. Benesty, S. Affes, A study of the LCMV and MVDR noise reduction filters, IEEE Trans. Signal Process. 58 (9) (2010) 4925–4935. doi:10.1109/TSP.2010.2051803.

[35] L. Ehrenberg, S. Gannot, A. Leshem, E. Zehavi, Sensitivity analysis of MVDR and MPDR beamformers, in: Proc. IEEE Convention Electrical and Electronics Engineers in Israel, IEEE, 2010, pp. 416–420.

[36] B. D. Carlson, Covariance matrix estimation errors and diagonal loading in adaptive arrays, IEEE Trans. Aerosp. Electron. Syst. 24 (4) (1988) 397–401.

[37] E. Fisher, J. Tabrikian, S. Dubnov, Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model, IEEE

Trans. Audio, Speech, and Language Process. 14 (2) (2006) 502–510. doi:10.1109/TSA.2005.857806.

[38] A. Jukić, T. van Waterschoot, T. Gerkmann, S. Doclo, Multi-channel linear prediction-based speech dereverberation with sparse priors, IEEE/ACM Trans. Audio, Speech, and Language Process. 23 (9) (2015) 1509–1520. doi:10.1109/TASLP.2015.2438549.

[39] E. A. P. Habets, Room impulse response generator, Tech. rep., Technische Universiteit Eindhoven, Eindhoven, Netherlands, ver. 2.0.20100920 (2010).

[40] E. A. P. Habets, I. Cohen, S. Gannot, Generating nonstationary multisensor signals under a spatial coherence constraint, The Journal of the Acoustical Society of America 124 (5) (2008) 2911–2917.

[41] J. K. Nielsen, J. R. Jensen, S. H. Jensen, M. G. Christensen, The single-and multichannel audio recordings database (SMARD), in: Proc. Int. Workshop on Acoustic Signal Enhancement, 2014, pp. 40–44.

[42] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, S. H. Jensen, Bayesian model comparison with the g-prior, IEEE Trans. Signal Process. 62 (1) (2014) 225–238. doi:10.1109/TSP.2013.2286776.

[43] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, S. H. Jensen, Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient, Signal Processing 135 (2017) 188–197. doi:10.1016/j.sigpro.2017.01.011.

[44] P. A. Naylor, E. A. P. Habets, J. Y.-C. Wen, N. D. Gaubitch, Models, measurement and evaluation, in: Speech Dereverberation, Springer, 2010, Ch. 2, pp. 21–56.

[45] A. Rix, J. Beerends, M. Hollier, A. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Vol. 2, 2001, pp. 749–752 vol.2. doi:10.1109/ICASSP.2001.941023.

[46] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., IEEE, 2010, pp. 4214–4217.