Article

On the Single-Sideband Transform for MVDR Beamformers

Vitor Probst Curtarelli [1,*], **Israel Cohen** [1]

[1] Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 3200003, Israel

* Correspondence: vitor.c@campus.technion.ac.il

Abstract: In order to explore different beamforming applicaitons, this paper investigates the application of the Single-Sideband Transform (SSBT) for constructing a Minimum-Variance Distortionless-Response (MVDR) beamformer in the context of the convolutive transfer function (CTF) model for short-window time-frequency transforms by making use of filter-banks and their properties. Our study aims to optimize the appropriate utilization of SSBT in this endeavor, by examining its characteristics and traits. We address a reverberant scenario with multiple noise sources, aiming to minimize both undesired interference and reverberation in the output. Through simulations reflecting real-life scenarios, we show that employing the SSBT correctly leads to a beamformer that outperforms the one obtained when via the Short-Time Fourier Transform (STFT), while exploiting the SSBT's property of it being real-valued. Two approaches were developed with the SSBT, one naive and one refined, with the later being able to ensure the desired distortionless behavior, which is not achieved by the former.

Keywords: Single-sideband transform; MVDR beamformer; Filter-banks; Array signal processing; Signal enhancement.

1. Introduction

Beamformers are an important tool for signal enhancement, being employed in a plethora of applications from hearing aids [1] to source localization [2] to imaging [3,4]. Among the possible ways to use such devices is to implement them the time-frequency domain [5], which allows the exploitation of frequency-related information while also dynamically adapting to signal changes over time. The most widely used instruments for time-frequency analysis are transforms, from which the Short-Time Fourier Transform (STFT) [6,7] stands out in terms of spread and commonness. However, alternative transforms can also be employed implemented [8–10], each offering unique perspective and information regarding the signal, possibly leading to different outputs.

Among these alternatives, the Single-Sideband Transform (SSBT) [11,12] is of great interest, given its real-valued frequency spectrum. It has been shown that the SSBT works particularly well with short analysis windows [11]. Therefore, if we use the convolutive transfer function (CTF) model [13] to study the desired signal model, the SSBT can lend itself to be useful, if we think about the beamforming process through the lenses of filter-banks [14,15]. Thus, by applying this transform within this context it is possible to pull off

Citation: Curtarelli, V. P.; Cohen, I. On the Single-Sideband Transform for MVDR Beamformers. *Algorithms* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

superior performances than only with the STFT. However, it is important to be aware of the limitations of the transform, in order to properly utilize it to try and achieve better outputs.

Two of the most important goals in beamforming are the minimization of noise in the output signal, and the distortionless-ness of the desired signal, both being achieved by the Minimum-Variance Distortionless-Response (MVDR) beamformer [16,17]. As the MVDR beamformer can be used on the time-frequency domain without restrictions on the transform chosen, it is possible to explore and compare the performance of this filter, when designing it through different time-frequency transforms.

Motivated by this, our paper explores the SSB transform and its application on the subject of beamforming within the context of the CTF model. We propose an approach for the CTF that allows the separation of desired and undesired speech components for reverberant environments, and employ this approach for designing the MVDR beamformer. We also explore the traits and limitations of the SSBT, and how to properly adapt the MVDR beamformer to this new transform's constraints. We show that a beamformer designed using the SSBT can surpass the STFT one, while also conforming to the distortionless constraint.

We organized the paper as follows: in Section 2 we introduce the proposed time-frequency transforms, how they're related and what are their relevant properties; Section 3 the considered signal model in the time domain is presented, and how it is transferred into the time-frequency domain; and in Section 4 we develop a true-MVDR beamformer with the SSBT, taking into account its features.

2. STFT and the Single-Sideband Transform

When studying signals and systems, often frequency and time-frequency transforms are used in order to change the signal domain [18], allowing the exploitation of different patterns and informations inherent to the signal.

Given a time-domain signal $x[n]$, its Short-time Fourier Transform (STFT) [6,7] is

$$X_{\mathcal{F}}[l,k] = \sum_{n=0}^{K-1} w[n]x[n+l\cdot O]e^{-\mathrm{j}2\pi k\frac{(n+l\cdot O)}{K}} \tag{1}$$

where $w[n]$ is an analysis window of length K; and O is the overlap between windows of the transform, usually $O = \lfloor K/2 \rfloor$. Even though the STFT is the most traditionally used time-frequency transform, it isn't the only one available. Thus, exploring different possibilities for such an operation can be useful and lead to interesting results.

The Single-Sideband Transform (SSBT) [11] is one such alternative, being cleverly constructed such that its frequency spectrum is real-valued, without loss of information. The SSB transform of $x[n]$ is defined as

$$X_{\mathcal{S}}[l,k] = \sqrt{2}\Re\left\{\sum_{n=0}^{K-1} w[n]x[n+l\cdot O]e^{-\mathrm{j}2\pi k\frac{(n+l\cdot O)}{K}+\mathrm{j}\frac{3\pi}{4}}\right\} \tag{2}$$

Assuming that $x[n]$ is real-valued, one advantage of using the STFT is that we only need to work with $\lfloor (K+1)/2 \rfloor + 1$ frequency bins, given its complex-conjugate behavior. Meanwhile, the SSBT requires all K bins to correctly capture all information of $x[n]$, however it is real-valued.

Assuming that all K bins of the STFT are available, from Eqs. (1) and (2) we have

$$X_{\mathcal{S}}[l,k] = \sqrt{2}\Re\left\{X_{\mathcal{F}}[l,k]e^{\mathrm{j}\frac{3\pi}{4}}\right\} \\ = -\Re\{X_{\mathcal{F}}[l,k]\} - \Im\{X_{\mathcal{F}}[l,k]\} \tag{3}$$

It is easy to see that[1]

$$X_{\mathcal{S}}[l,k] = \frac{1}{\sqrt{2}}\left(e^{\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{F}}[l,k] + e^{-\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{F}}[l,K-k]\right) \tag{4}$$

from which it we deduce

$$X_{\mathcal{F}}[l,k] = \frac{1}{\sqrt{2}}\left(e^{-\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{S}}[l,k] + e^{\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{S}}[l,K-k]\right) \tag{5}$$

One disadvantage of the SSBT is that the convolution theorem does not hold when employing it (see Appendix A), not even as an approximation. Nonetheless, by converting any result in the SSBT domain to the STFT domain (using Eq. (3)) before utilization, it remains feasible to employ the transform to study of the problem at hand.

3. Signal Model and Beamforming

Let there be a device that consists of M sensors and a loudspeaker (LS) in a reverberant environment. In this setting there also are a desired source, an interfering source, and uncorrelated noise impinging at each sensor, all traveling with a speed c. For simplicity we assume that all sources are spatially stationary, although this condition can be easily removed.

We denote $y_m[n]$ as the signal at the m-th sensor, being defined as

$$y_m[n] = h_m[n] * x[n] + g_m[n] * w[n] + e_m[n] * s[n] + r_m[n] \tag{6}$$

in which $h_m[n]$ is the impulse response between the desired source and the m-th sensor ($1 \leq m \leq M$), with $x[n]$ being the desired source's signal; similarly for the interfering noise $w[n]$ and its IR $g_m[n]$, and th speaker's signal $s[n]$ and its IR $e_m[n]$; and $r_m[n]$ is the uncorrelated noise.

We let m' be the reference sensor's index, for simplicity assume $m' = 1$, and also $x_1[n] = h_1[n] * x[n]$ (and similarly for $v_1[n]$ and $s_1[n]$). We define $a_m[n]$ as the *relative* impulse response between the desired signal (at the reference sensor) and the m-th sensor, such that

$$a_m[n] * x_1[n] = h_m[n] * x[n] \tag{7}$$

We similarly define $b_m[n]$ such that $b_m[n] * w_1[n] = g_m[n] * w[n]$, and $c_m[n]$ from $e_m[n]$ and $s[n]$. Therefore, Eq. (6) becomes

$$y_m[n] = a_m[n] * x_1[n] + b_m[n] * w_1[n] + c_m[n] * s_1[n] + r_m[n] \tag{8}$$

Here, the impulse responses ($a_m[n]$, $b_m[n]$, $c_m[n]$) can be non-causal, depending on the direction of arrival and features of the reverberant environment.

[1] For the abuse of notation, we let $X_{\mathcal{S}}[l,K] \equiv X_{\mathcal{S}}[l,0]$, and equally for $X_{\mathcal{F}}[l,K]$.

We use a time-frequency transform (such as the STFT or SSBT, as in Section 2) with the convolutive transfer-function (CTF) model [13] to obtain our time-frequency signal model,

$$Y_m[l,k] = A_m[l,k] * X_1[l,k] + B_m[l,k] * W_1[l,k] + C_m[l,k] * S_1[l,k] + R_m[l,k] \tag{9}$$

where $Y_m[l,k]$ is the transform of $y_m[n]$ (resp. all other signals); l and k are the window (or decimated-time) and bin indexes, with $0 \leq k \leq K-1$; and the convolution is in the window-index axis.

Using that $A_m[l,k]$ is a finite (possibly truncated) response with L_A windows, then

$$A_m[l,k] * X_1[l,k] = \mathbf{a}_m^\mathsf{T}[k]\mathbf{x}_1[l,k] \tag{10}$$

in which

$$\mathbf{a}_m[k] = \Big[\, A_m[-\Delta,k]\,,\ \cdots,\ A_m[0,k]\,,\ \cdots,\ A_m[L_B-\Delta-1,k]\,\Big]^\mathsf{T} \tag{11a}$$

$$\mathbf{x}_1[l,k] = \Big[\, X_1[l+\Delta,k]\,,\ \cdots,\ X_1[l,k]\,,\ \cdots,\ X_1[l-L_B+\Delta+1,k]\,\Big]^\mathsf{T} \tag{11b}$$

and in the same way we define $\mathbf{b}_m[k]$, $\mathbf{w}_1[l,k]$, $\mathbf{d}_m[k]$ and $\mathbf{s}_1[l,k]$. Note that $\mathbf{a}_m[k]$ and $\mathbf{b}_m[k]$ don't depend on the index l, given the spatial stationarity assumption. Also, Δ is the number of non-causal windows in the reference sensor necessary to capture the whole signal. With this, Eq. (9) becomes

$$Y_m[l,k] = \mathbf{a}_m^\mathsf{T}[k]\mathbf{x}_1[l,k] + \mathbf{b}_m^\mathsf{T}[k]\mathbf{w}_1[l,k] + \mathbf{c}_m^\mathsf{T}[k]\mathbf{s}_1[l,k] + R_m[l,k] \tag{12}$$

Vectorizing the signals sensor-wise, we finally get

$$\mathbf{y}[l,k] = \mathbf{A}[k]\mathbf{x}_1[l,k] + \mathbf{B}[k]\mathbf{w}_1[l,k] + \mathbf{C}[k]\mathbf{s}_1[l,k] + \mathbf{r}[l,k] \tag{13}$$

where

$$\mathbf{y}[l,k] = \Big[\, y_1[l,k]\,,\ \cdots,\ y_M[l,k]\,\Big]^\mathsf{T} \tag{14}$$

and similarly for the other variables. In this situation, $\mathbf{A}[k]$, $\mathbf{B}[k]$ and $\mathbf{C}[k]$ are $M \times L_A$, $M \times L_B$ and $M \times L_C$ matrices respectively; $\mathbf{x}_1[l,k]$, $\mathbf{w}_1[l,k]$ and $\mathbf{s}_1[l,k]$ are $L_A \times 1$, $L_B \times 1$ and $L_C \times 1$ vectors respectively; and $\mathbf{y}[l,k]$ and $\mathbf{r}[l,k]$ are $M \times 1$ vectors.

3.1. *Reverb-aware formulation*

Let $l = 0$ be the desired window which we would like to retrieve from the signal $\mathbf{A}[k]\mathbf{x}_1[l,k]$. We can write $\mathbf{A}[k]\mathbf{x}_1[l,k]$ as

$$\mathbf{A}[k]\mathbf{x}_1[l,k] = \mathbf{d}_x[k]X_1[l,k] + \mathbf{q}[l,k] \tag{15}$$

where $X_1[l,k]$ is the desired speech signal, and $\mathbf{q}[l,k]$ is an undesired component, uncorrelated to $X_1[l,k]$. Through a similar process as exposed in [19] (sec. 7.1.1) (see Appendix B for details), we can deduce that $\mathbf{d}_x[k]$ can be defined as

$$\mathbf{d}_x[k] = \frac{\sum_i \mathbf{a}_m[k]_i \sum_{n=0}^{K-1-|i\mathcal{O}|} w[n]w[n+|i\mathcal{O}|]}{\sum_{n'=0}^{K-1} w[n']^2} \tag{16}$$

in which $w[n]$ and $\mathcal{O}$ are the window-function and the decimation factor used for the time-frequency transform; and $\mathbf{a}_m[k]_i$ is the $(\Delta+i)$-th element of $\mathbf{a}_m[k]$. From Appendix A, we also have that $-\lfloor (K-1)/\mathcal{O} \rfloor \leq i \leq \lfloor (K-1)/\mathcal{O} \rfloor$.

Note that $\mathbf{d}_x[k]X_1[l,k]$ also consists of some reverberation, since $x_1[n] = h_1[n] * x[n]$ is the desired signal at the reference sensor and not at source, therefore being affected by the environment. However, this formulation allows us to better estimate the influence of the neighboring time-frequency windows over the desired signal, given their overlap.

It's important to have in mind the sensor delay and window length. If the time for the signal to travel from the reference to the farthest sensor exceeds the window length (in seconds), different windows may represent the desired speech, for each sensor. This isn't a problem if $\frac{\bar{\delta}}{c} < \frac{K}{f_s}$, where $\bar{\delta}$ is the largest reference-to-sensor distance among all sensors.

Using Eq. (15) we define $\mathbf{v}[l,k]$ as the undesired signal (undesired speech components + speaker signal + interfering source + uncorrelated noise),

$$\mathbf{v}[l,k] = \mathbf{q}[l,k] + \mathbf{B}[k]\mathbf{w}_1[l,k] + \mathbf{C}[k]\mathbf{s}_1[l,k] + \mathbf{r}[l,k] \tag{17}$$

and therefore

$$\mathbf{y}[l,k] = \mathbf{d}_x[k]X_1[l,k] + \mathbf{v}[l,k] \tag{18}$$

3.2. Filtering and the MPDR beamformer

We estimate the desired signal at reference $X_1[l,k]$ as $Z[l,k]$ through a filter $\mathbf{f}[l,k]$, such that

$$\begin{aligned} Z[l,k] &= \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{y}[l,k] \\ &\approx X_1[l,k] \end{aligned} \tag{19}$$

with $(\cdot)^{\mathsf{H}}$ being the transposed-complex-conjugate operator. Since the source signals' properties can vary over time, so can the filter, adapting to the environment.

To minimize the variance of the output signal while obeying the distortionless constraint $\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[l,k] = 1$, a Minimum-Power Distortionless Response (MPDR) beamformer will be used, it being defined as

$$\mathbf{f}_{\mathrm{mpdr}}[l,k] = \min_{\mathbf{f}[l,k]} \mathbf{f}[l,k]^{\mathsf{H}}\mathbf{\Phi}_{\mathbf{y}}[l,k]\mathbf{f}[l,k] \text{ s.t. } \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[l,k] = 1 \tag{20}$$

where $\mathbf{\Phi}_{\mathbf{y}}[l,k]$ is the correlation matrix of the observed signal $\mathbf{y}[l,k]$. The solution to this minimization problem

$$\mathbf{f}_{\mathrm{mpdr}}[l,k] = \frac{\mathbf{\Phi}_{\mathbf{y}}^{-1}[l,k]\mathbf{d}_x[l,k]}{\mathbf{d}_x^{\mathsf{H}}[k]\mathbf{\Phi}_{\mathbf{y}}^{-1}[l,k]\mathbf{d}_x[k]} \tag{21}$$

4. True-MPDR with the Single-Sideband Transform

When carelessly using any of the established methods with the SSBT, the distortionless constraint ensures that the beamformer avoids causing distortion exclusively within the SSBT domain. However, as explained in Section 2 the SSBT beamformer must be carefully constructed to achieve the desired effects, such as the distortionless constraint.

We thus propose a framework for the SSBT in which we consider the bins k and $K-k$ simultaneously, since from Eq. (5) they both contribute to the k-th bin in the STFT domain. We define $\mathbf{y}'[l,k]$ as

$$\mathbf{y}'[l,k] = \begin{bmatrix} \mathbf{y}[l,k] \\ \mathbf{y}[l,K-k] \end{bmatrix}_{2M\times 1} \tag{22}$$

from which we define $\mathbf{\Phi}_{\mathbf{y}'}[l,k]$ as its correlation matrix. Under this idea, our filter $\mathbf{f}[l,k]$ is a $2M \times 1$ vector, with the first M values being for the k-th bin, and the last M values for the $[K-k]$-th bin. We let the STFT-equivalent filter for the SSBT beamformer $\mathbf{f}[l,k]$ be $\mathbf{f}_{\mathcal{F}}[l,k]$, given by

$$\mathbf{f}_{\mathcal{F}}[l,k] = \mathbf{\Lambda}\mathbf{f}[l,k] \tag{23}$$

in which

$$\mathbf{\Lambda} = \frac{1}{\sqrt{2}} \begin{bmatrix} e^{-\mathrm{j}\frac{3\pi}{4}} & 0 & \cdots & 0 & e^{\mathrm{j}\frac{3\pi}{4}} & 0 & \cdots & 0 \\ 0 & e^{-\mathrm{j}\frac{3\pi}{4}} & \cdots & 0 & 0 & e^{\mathrm{j}\frac{3\pi}{4}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \cdots & e^{-\mathrm{j}\frac{3\pi}{4}} & 0 & 0 & \cdots & e^{\mathrm{j}\frac{3\pi}{4}} \end{bmatrix}_{M\times 2M} \tag{24}$$

From Eq. (23) the constraint matrix within the SSBT domain becomes

$$\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_{x;\mathcal{S}}[k] = 1 \tag{25a}$$

$$\mathbf{d}_{x;\mathcal{S}}[k] = \mathbf{\Lambda}^{\mathsf{H}}\mathbf{d}_{x;\mathcal{F}}[k] \tag{25b}$$

where $\mathbf{d}_{x;\mathcal{F}}[l,k]$ is the constraint matrix within the STFT domain, and $\mathbf{d}_{x;\mathcal{S}}[k]$ is the new constraint matrix within the SSBT domain.

In this scheme, our minimization problem becomes

$$\mathbf{f}_{\mathrm{mpdr}}[l,k] = \min_{\mathbf{f}[l,k]} \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{\Phi}_{\mathbf{y}'}[l,k]\mathbf{f}[l,k] \text{ s.t. } \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_{x;\mathcal{S}}[k] = 1 \tag{26}$$

Although $\mathbf{\Phi}_{\mathbf{y}'}[l,k]$ is a matrix with real entries, $\mathbf{d}_{x;\mathcal{S}}[k]$ is complex-valued, and thus is the solution to Eq. (26), contradicting the purpose of utilizing the SSBT.

4.1. *Real-valued true-MPDR beamformer with SSBT*

To ensure the desired behavior of $\mathbf{f}[l,k]$ being real-valued, an additional constraint is necessary. By forcing $\mathbf{f}[l,k]$ to have real entries, from Eq. (25a) we trivially have that

$$\mathbf{f}^{\mathsf{T}}[l,k]\Re\{\mathbf{d}_{x;\mathcal{S}}[k]\} = 1 \tag{27a}$$

$$\mathbf{f}^{\mathsf{T}}[l,k]\Im\{\mathbf{d}_{x;\mathcal{S}}[k]\} = 0 \tag{27b}$$

which can be put in matricial form as $\mathbf{f}^{\mathsf{T}}[l,k]\mathbf{D}_x[k] = \mathbf{i}_2^{\mathsf{T}}$, with

$$\begin{aligned} \mathbf{D}_x[k] &= \begin{bmatrix} \Re\{\mathbf{d}_{x;\mathcal{S}}[k]\}\,, & \Im\{\mathbf{d}_{x;\mathcal{S}}[k]\} \end{bmatrix}_{2M\times 2} \\ &= \begin{bmatrix} \mathbf{d}_{x;\Re}[k]\,, & \mathbf{d}_{x;\Im}[k] \end{bmatrix} \end{aligned} \tag{28a}$$

$$\mathbf{i}_2 = \begin{bmatrix} 1\,, & 0 \end{bmatrix}^{\mathsf{T}} \tag{28b}$$

Therefore, the minimization problem from Eq. (26) becomes

$$\mathbf{f}_{\mathrm{mpdr}}[l,k] = \min_{\mathbf{f}[l,k]} \mathbf{f}^{\mathsf{T}}[l,k]\mathbf{\Phi}_{\mathbf{y}'}[l,k]\mathbf{f}[l,k] \text{ s.t. } \mathbf{f}^{\mathsf{T}}[l,k]\mathbf{D}_x[k] = \mathbf{i}_2^{\mathsf{T}} \tag{29}$$

whose formulation is the same as of the Linearly-Constrained Minimum Power (LCMP) beamformer, and thus its solution is

$$\mathbf{f}_{\mathrm{mpdr}}[l,k] = \mathbf{\Phi}_{\mathbf{y}'}^{-1}[l,k]\mathbf{D}_x[k]\left(\mathbf{D}_x^{\top}[k]\mathbf{\Phi}_{\mathbf{y}'}^{-1}[l,k]\mathbf{D}_x[k]\right)^{-1}\mathbf{i}_2 \tag{30}$$

Using Eq. (23), we can obtain the desired beamformer $\mathbf{f}_{\mathcal{F};\mathrm{mpdr}}[l,k]$, transformed to the STFT domain.

Here onward we will omit the $[k]$ and $[l,k]$ indices for clarity, except for definitions. Using Eq. (28), we can write

$$\mathbf{D}_x^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{D}_x = \begin{bmatrix} \mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Re} & \mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} \\ \mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} & \mathbf{d}_{x;\Im}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} \end{bmatrix} \tag{31}$$

With this,

$$\left(\mathbf{D}_x^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{D}_x\right)^{-1} = \begin{bmatrix} \mathbf{d}_{x;\Im}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} & -\mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} \\ -\mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} & \mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Re} \end{bmatrix} \cdot \frac{1}{\det\left(\mathbf{D}_x^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{D}_x\right)} \tag{32a}$$

$$\det\left(\mathbf{D}_x^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{D}_x\right) = \mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Re}\mathbf{d}_{x;\Im}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} - \mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im}\mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} \tag{32b}$$

Going one step further,

$$\left(\mathbf{D}_x^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{D}_x\right)^{-1}\mathbf{i}_2 = \begin{bmatrix} \mathbf{d}_{x;\Im}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} \\ -\mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} \end{bmatrix} \cdot \frac{1}{\det\left(\mathbf{D}_x^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{D}_x\right)} \tag{33}$$

and thus

$$\mathbf{D}_x\left(\mathbf{D}_x^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{D}_x\right)^{-1}\mathbf{i}_2 = \frac{\mathbf{d}_{x;\Re}\mathbf{d}_{x;\Im}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} - \mathbf{d}_{x;\Im}\mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im}}{\det\left(\mathbf{D}_x^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{D}_x\right)} \tag{34}$$

Finally,

$$\mathbf{f}_{\mathrm{mpdr}} = \frac{\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Re}\mathbf{d}_{x;\Im}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} - \mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im}\mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im}}{\mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Re}\mathbf{d}_{x;\Im}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im} - \mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im}\mathbf{d}_{x;\Re}^{\top}\mathbf{\Phi}_{\mathbf{y}'}^{-1}\mathbf{d}_{x;\Im}} \tag{35}$$

Denoting $\mathbf{\Omega} \equiv \mathbf{\Omega}[l,k]$ as

$$\mathbf{\Omega} \triangleq \mathbf{\Phi}_{\mathbf{y}'}^{-1}\left(\mathbf{d}_{x;\Re}\mathbf{d}_{x;\Im}^{\top} - \mathbf{d}_{x;\Im}\mathbf{d}_{x;\Re}^{\top}\right)\mathbf{\Phi}_{\mathbf{y}'}^{-1} \tag{36}$$

then we finally arrive at the final version of our MPDR beamformer with the SSBT,

$$\mathbf{f}_{\mathrm{mpdr}}[l,k] = \frac{\mathbf{\Omega}[l,k]\,\mathbf{d}_{x;\Im}[k]}{\mathbf{d}_{x;\Re}[k]^{\top}\,\mathbf{\Omega}[l,k]\,\mathbf{d}_{x;\Im}[k]} \tag{37}$$

This equation is advantageous since it is similar to the MPDR's formulation for the STFT, given in Eq. (21). This allows us to study both beamformers similarly, given that their expressions are analogous.

4.2. Comparison between beamformers

In the simulations[2], we employ a sampling frequency of 16kHz. Room impulse responses were generated using Habets' RIR generator [20], and signals were selected from the SMARD database [21].

The room's dimensions are 4m $\times$ 6m $\times$ 3m (width $\times$ length $\times$ height), with a reverberation time of 0.3s. The device composed of the loudspeaker + sensors is centered at $(2\text{m}, 2\text{m}, 1\text{m})$. Its sensors are located in a circular array with radius of 8cm, all omnidirectional and of flat frequency response. The positions and signals used for the sources are in Table 1. Although both interfering and noise signals are the same, they were taken starting on different timestamps, to ensure they are uncorrelated. All signals were resampled to the desired sampling frequency of 16kHz.

Source	Position	Signal
$x[n]$	$(2\text{m}, 1\text{m}, 1\text{m})$	`50_male_speech_english_ch8_OmniPower4296.flac`
$s[n]$	$(2\text{m}, 2\text{m}, 1\text{m})$	`69_abba_ch8_OmniPower4296.flac`
$w[n]$	$(2\text{m}, 5.8\text{m}, 2.8\text{m})$	`wgn_48kHz_ch8_OmniPower4296.flac`
$r[n]$	$\sim$	`wgn_48kHz_ch8_OmniPower4296.flac`

Table 1. Source information for the simulations.

At the reference sensor (the one at $(2\text{m}, 2\text{m}, 1.08\text{m})$), the SNR for the loudspeaker's, interfering, and noise signals are respectively -15dB, 10dB and 30dB. These will be referred as Signal-to-Echo, Signal-to-Interference and Signal-to-White-Noise Ratios (SER, SIR and SWR) respectively, with SNR still being the ratio between signal and the sum of the noises.

For the transforms, Hann windows were used, with a length of 32 samples/window and an overlap of 50%. The beamformers were calculated every 200 frames (equivalent to every 0.2s), and used up to the previous 1000 samples to estimate the correlation matrices.

We will compare one beamformer for the STFT with and two for the SSBT. The STFT one will be based on Section 3.2 and Eq. (21), as well as the first with the SSBT (which will be called "Single-Frequency SSBT", or "SF-SSBT" for short). The second one based on the SSBT will be called "Dual-Frequency SSBT" (or "DF-SSBT"), as derived in Section 4, building up to Eq. (37). These names are logical, given that the one proposed in Section 4 uses two frequencies (namely the "dual-frequencies" from the STFT) at each moment.

In line plots, STFT is presented in red with continuous lines, SF-SSBT in green with dashed lines, and DF-SSBT in blue with dotted lines. The output metrics were averaged over 200 frames and presented every 100 windows, for a better visualization.

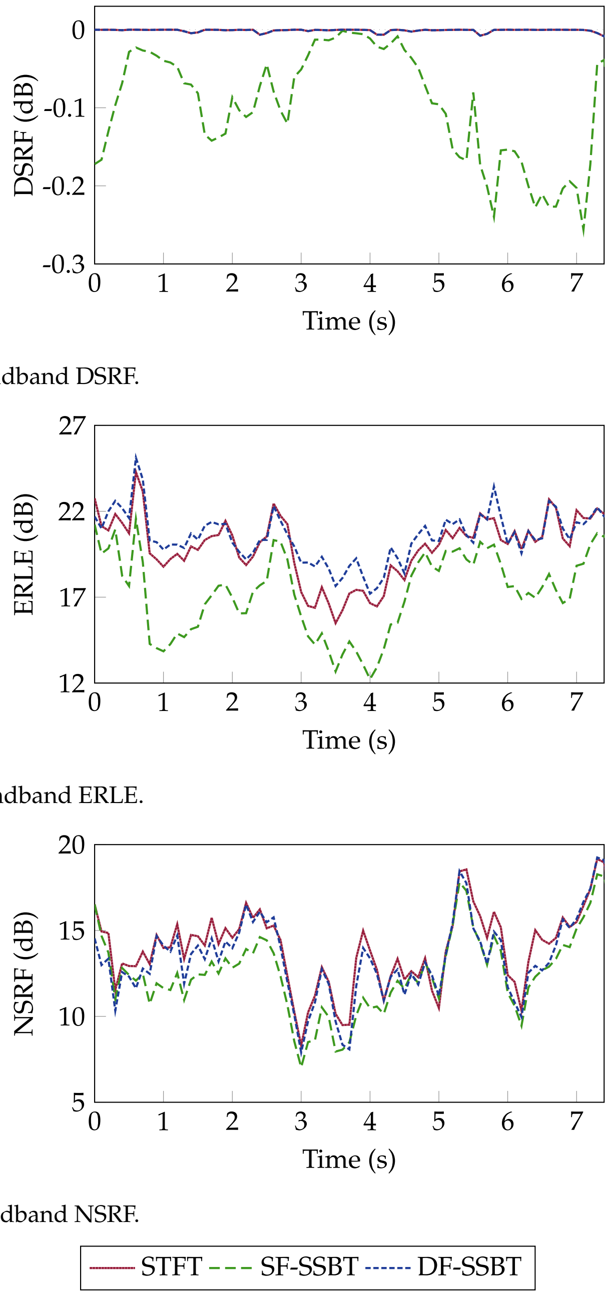
In these simulations, we are interested in three main metrics: desired signal reduction factor (DSRF), echo-return loss enhancement (ERLE), and noise signal reduction factor (NSRF). Their time-dependent broadband formulations are

$$\xi_x[l] = \frac{\sum_k |X_1[l,k]|^2}{\sum_k \left|X_f[l,k]\right|^2} \tag{38a}$$

$$\xi_s[l] = \frac{\sum_k |S_1[l,k]|^2}{\sum_k \left|S_f[l,k]\right|^2} \tag{38b}$$

$$\xi_v[l] = \frac{\sum_k |V_1[l,k]|^2}{\sum_k \left|V_f[l,k]\right|^2} \tag{38c}$$

[2] Code is available at https://github.com/VCurtarelli/py-ssb-ctf-bf.



(a) Per-window broadband DSRF.

(b) Per-window broadband ERLE.

(c) Per-window broadband NSRF.

Figure 1. Output metrics for the beamformers.

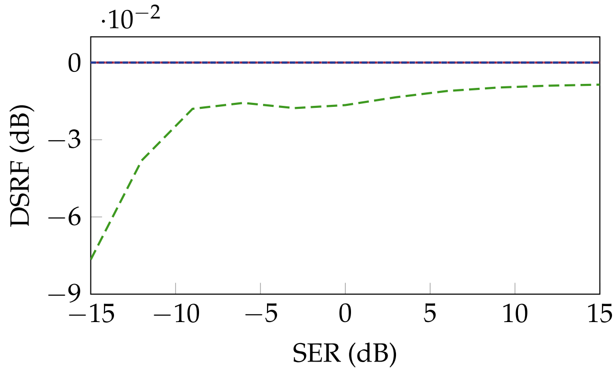
and their time-average broadband definitions are 209

$$\xi_x = \frac{\sum_{l,k} |X_1[l,k]|^2}{\sum_{l,k} \left|X_f[l,k]\right|^2} \tag{39a}$$

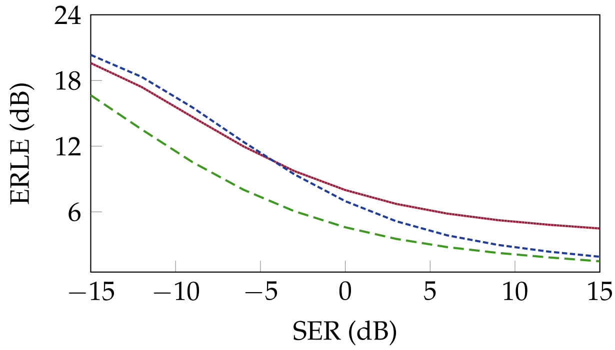
$$\xi_s = \frac{\sum_{l,k} |S_1[l,k]|^2}{\sum_{l,k} \left|S_f[l,k]\right|^2} \tag{39b}$$

$$\xi_v = \frac{\sum_{l,k} |V_1[l,k]|^2}{\sum_{l,k} \left|V_f[l,k]\right|^2} \tag{39c}$$

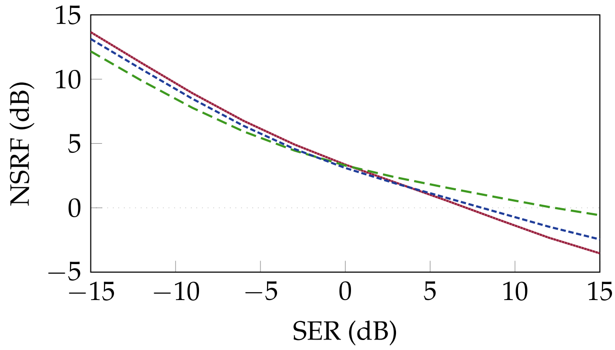
were $S_f[l,k] = \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{s}_1[l,k]$, $X_f[l,k] = \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{x}_1[l,k]$ and $V_f[l,k] = \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{v}_1[l,k]$ as the 210
filtered-LS, filtered-desired and filtered-undesired signals, respectively. 211



(a) Time-average broadband DSRF.



(b) Time-average broadband ERLE.



(c) Time-average broadband NSRF.

Figure 2. Output metrics for the beamformers.

These metrics represent, respectively, how much distortion is being caused in the desired signal, how much the echo (loudspeaker signal) is being reduced, and how much the general noise is being reduced.

From Fig. 1a, we see that all beamformers had a negligible distortion of less than 0.5dB, however the distortion for the STFT and DF-SSBT beamformers was still closer to zero.

From the ERLE results in Fig. 1b it is noticeable that the STFT and SF-SSBT beamformers had a similar performance, with the DF-SSBT one being slightly better, both outperforming the SF-SSBT beamformer's results by approx. 4dB. A similar result can be seen in Fig. 1c for the NSRF, however in this regard the STFT had a minimal margin over the DF-SSBT, and their gap compared to the SF-SSBT wasn't as prominent, although still visible.

We will also compare the results with a varying input SERs, to assess the beamformer's performances for different loudspeaker signal levels. For such, we will use the time-average metrics as in Eq. (39). The other parameters and variables are maintained from the previous simulations, with only the SER being changed.

As seen in Fig. 2a, the SF-SSBT beamformer caused some distortion on the desired signal. However, this distortion is minimal, and decreases as the loudspeaker SNR increases.

In terms of ERLE, we see that the SF-SSBT is the worse for all SER's. The DF-SSBT beamformer is slightly better than the STFT, for SER $\lessapprox -5$dB, and the STFT beamformer is considerably better otherwise.

For the NSRF, the opposite is true regarding these two beamformers, with the STFT one being better for lower input SERs in terms of overall SNR gain, and the SSBT being better for higher input SERs. Interestingly, for this metric the SF-SSBT beamformer manages to outperform the other two for higher input SERs.

5. Perturbation Robustness Analysis

Until now, we assumed an appropriate knowledge of all signals and their impulse responses. However, in a real application these would be estimated, and thus prone to error. Given our beamformers from Eqs. (21) and (30) and their dependence on $\mathbf{d}$, they are directly influenced by impulse response estimation errors.

We can write

$$\mathbf{d}_x[k] = \bar{\mathbf{d}}_x[k] + \delta_x[k] \tag{40}$$

where $\bar{\mathbf{d}}_x[k]$ is the accurate steering vector (SV), $\mathbf{d}_x[k]$ is the measured SV, $\delta_x[k]$ is a perturbation (or error) on the SV measurement. With this, the MPDR beamformer with the STFT (assuming the knowledge of $\mathbf{d}_x[k]$) from Eq. (21) is

$$\begin{aligned}
\mathbf{f}_{\text{mpdr}} &= \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\mathbf{d}_x}{\left(\mathbf{d}_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\mathbf{d}_x\right)} \\
&= \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(\bar{\mathbf{d}}_x + \delta_x)}{\left(\bar{\mathbf{d}}_x^{\mathsf{H}} + \delta_x^{\mathsf{H}}\right)\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(\bar{\mathbf{d}}_x + \delta_x)} \\
&= \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x + \boldsymbol{\Phi}_{\mathbf{y}}^{-1}\delta_x}{\bar{\mathbf{d}}_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x + \bar{\mathbf{d}}_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\delta_x + \delta_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x + \delta_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\delta_x} \\
&= \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x}{\bar{\mathbf{d}}_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x + e} + \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\delta_x}{\bar{\mathbf{d}}_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x + e}
\end{aligned} \tag{41}$$

$$\mathbf{f}_{\text{mpdr}}[l,k] = g_e[l,k]\bar{\mathbf{f}}_{\text{mpdr}}[l,k] + \mathbf{f}_\delta[l,k] \tag{42}$$

in which $\bar{\mathbf{f}}_{\text{mpdr}}[l,k]$ is the beamformer for the accurate steering vector $\bar{\mathbf{d}}_x[k]$, and $g_e[l,k]$ and $\mathbf{f}_\delta[l,k]$ are

$$g_e = \frac{\bar{\mathbf{d}}_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x}{\bar{\mathbf{d}}_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x + e} \tag{43a}$$

$$\mathbf{f}_\delta = \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\delta_x}{\bar{\mathbf{d}}_x^{\mathsf{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\bar{\mathbf{d}}_x + e} \tag{43b}$$

Thus, the beamformer with an estimated steering vector is an affine transformation of $\bar{\mathbf{f}}_{\mathrm{mpdr}}[l,k]$. It is trivial that $\delta_x \to \mathbf{0} \implies g_e = 1,\ \mathbf{f}_\delta = \mathbf{0}$.

Applying the same procedure to the MPDR beamformer with the SSBT from Eq. (37), we get a similar result as obtained in Eq. (42), but in which e, g_e, and $\mathbf{f}_\delta$ now are

$$e = \mathbf{d}_{x;\Re}^{\mathsf{T}}\,\boldsymbol{\Omega}\,\delta_{x;\Im} + \delta_{x;\Re}^{\mathsf{T}}\,\boldsymbol{\Omega}\,\mathbf{d}_{x;\Im} + \delta_{x;\Re}^{\mathsf{T}}\,\boldsymbol{\Omega}\,\delta_{x;\Im} \tag{44a}$$

$$g_e = \frac{\mathbf{d}_{x;\Re}^{\mathsf{T}}\,\boldsymbol{\Omega}\,\mathbf{d}_{x;\Im}}{\mathbf{d}_{x;\Re}^{\mathsf{T}}\,\boldsymbol{\Omega}\,\mathbf{d}_{x;\Im} + e} \tag{44b}$$

$$\mathbf{f}_\delta = \frac{\boldsymbol{\Omega}\,\delta_{x;\Im}}{\mathbf{d}_{x;\Re}^{\mathsf{T}}\,\boldsymbol{\Omega}\,\mathbf{d}_{x;\Im} + e} \tag{44c}$$

where $\delta_{x;\Re}$ and $\delta_{x;\Im}$ are the perturbation vectors for $\mathbf{d}_{x;\Re}$ and $\mathbf{d}_{x;\Im}$ respectively.

5.1. Tests

In these tests, we repeated the simulation as before, but now with a slightly miscalculated steering vector. This was done by simulating $\delta_x[k]$ as a zero-mean random additive noise, whose variance is the controllable variable. In the plots of Fig. 3, the x-axis indicates the standard deviation of $\delta_x[k]$ relative to that of $\mathbf{d}_x[k]$. Trivially, the case where the error is 0% is when the steering vector is accurately evaluated. The input SER in this scenario is of -15dB.

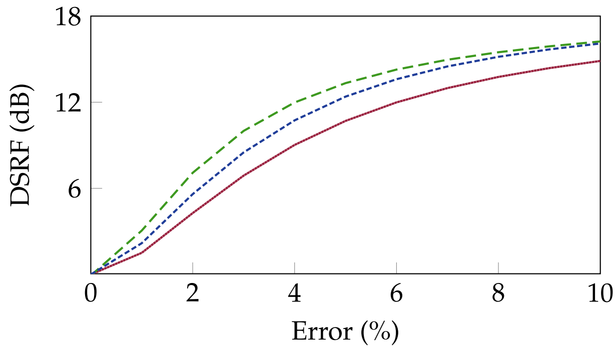
We are still interested in the same three results as before, but the (possible) desired signal distortion, the Signal-to-Echo Ratio gain (gSER) and Signal-to-Noise Ratio gain (gSNR) will be used, which will normalize the decrease in the undesired signals, by how much the desired signal was reduced. These are defined as

$$\mathrm{gSER} = \frac{\xi_s}{\xi_x} \tag{45a}$$

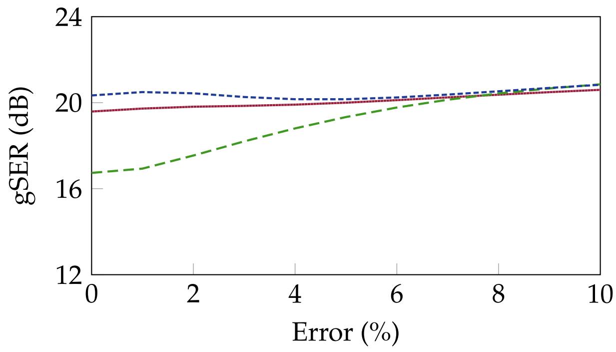
$$\mathrm{gSNR} = \frac{\xi_v}{\xi_x} \tag{45b}$$

Analysing the results, we see that all methods follow a similar pattern, in all observed metrics. In the DSRF, it is noticeable that the signal distortion increases as the error in the steering vector increases. Interestingly, the gain in SER increases slightly as the error increases, however in general the overall gain in SNR decreases, so even though we reduce more of the echo signal, we would end up increasing the other undesired sources.

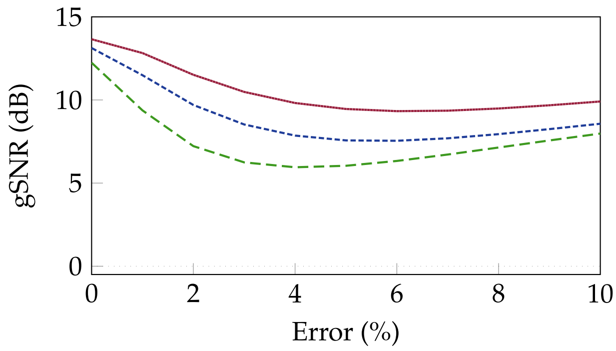
In both Figs. 3a and 3c we see that the STFT beamformer was the more robust one, followed by the DF-SSBT and then the SF-SSBT. The results in Fig. 3b differ from this, with the DF-SSBT marginally outperforming the STFT beamformer.



(a) Time-average broadband DSRF.



(b) Time-average broadband gSER.



(c) Time-average broadband gSNR.

Figure 3. Output metrics for the beamformers.

Author Contributions: Conceptualization, I. Cohen and V. Curtarelli; Methodology, V. Curtarelli; Software, V. Curtarelli; Writing—original draft: V. Curtarelli; Writing—review and editing, I. Cohen and V. Curtarelli; Supervision, V. Curtarelli. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Pazy Research Foundation, and the Israel Science Foundation (grant no. 1449/23).

Data Availability Statement: The source-code for the simulations developed here is available at https://github.com/VCurtarelli/py-ssb-ctf-bf.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CTF	Convolutive Transfer Function
DSRF	Desired Signal Reduction Factor
MPDR	Minimum-Power Distortionless-Response
MTF	Multiplicative Transfer Function
SNR	Signal-to-Noise Ratio
SSBT	Single-Sideband Transform
STFT	Short-Time Fourier Transform

Appendix A. SSBT Convolution

Let $x[n]$ be a time domain signal, with $X_{\mathcal{F}}[l,k]$ being its STFT equivalent, and $X_{\mathcal{S}}[l,k]$ its SSBT equivalent. We here assume that both the STFT and the SSBT have K frequency bins. $X_{\mathcal{S}}[l,k]$ can be obtained using $X_{\mathcal{F}}[l,k]$, through

$$X_{\mathcal{S}}[l,k] = -X_{\mathcal{F}}{}^{\Re}[l,k] - X_{\mathcal{F}}{}^{\Im}[l,k] \tag{A.1}$$

in which $(\cdot)^{\Re}$ and $(\cdot)^{\Im}$ represent the real and imaginary components of their argument, respectively.

It is easy to see that

$$X_{\mathcal{F}}[l,k] = \frac{1}{\sqrt{2}}\left(e^{-\mathrm{j}\frac{3\pi}{4}} X_{\mathcal{S}}[l,k] + e^{\mathrm{j}\frac{3\pi}{4}} X_{\mathcal{S}}[l,K-k]\right) \tag{A.2}$$

As stated before, in this formulation we abuse the notation by letting $X_{\mathcal{S}}[l,K] = X_{\mathcal{S}}[l,0]$ to simplify the mathematical operations.

Now, let there also be $h[n]$, $H_{\mathcal{F}}[k]$ and $H_{\mathcal{S}}[k]$, with the same assumptions as before. We define $Y_{\mathcal{F}}[l,k]$ and $Y_{\mathcal{S}}[l,k]$ as the output of an LTI system with impulse response $h[n]$, such that

$$Y_{\mathcal{F}}[l,k] = H_{\mathcal{F}}[k] X_{\mathcal{F}}[l,k] \tag{A.3a}$$
$$Y_{\mathcal{S}}[l,k] = H_{\mathcal{S}}[k] X_{\mathcal{S}}[l,k] \tag{A.3b}$$

We will assume that the MTF model [13] correctly models the convolution here. This waas used instead of the CTF for simplicity, as these derivations would work exactly the same for the CTF, but with window-wise summations as well, which would pollute the notation.

Applying Eq. (A.1) in Eq. (A.3b), and knowing that $X_{\mathcal{F}}[l,k] = X^{*}_{\mathcal{F}}[l,K-k]$ (same for $H_{\mathcal{F}}[k]$), with $(\cdot)^{*}$ representing the complex-conjugate; we get that

$$\begin{aligned} Y_{\mathcal{S}}[l,k] &= X_{\mathcal{F}}{}^{\Re}[l,k] H_{\mathcal{F}}{}^{\Re}[l,k] + X_{\mathcal{F}}{}^{\Re}[l,k] H_{\mathcal{F}}{}^{\Im}[l,k] \\ &\quad + X_{\mathcal{F}}{}^{\Im}[l,k] H_{\mathcal{F}}{}^{\Re}[l,k] + X_{\mathcal{F}}{}^{\Im}[l,k] H_{\mathcal{F}}{}^{\Im}[l,k] \\ Y_{\mathcal{S}}[l,K-k] &= X_{\mathcal{F}}{}^{\Re}[l,k] H_{\mathcal{F}}{}^{\Re}[l,k] - X_{\mathcal{F}}{}^{\Re}[l,k] H_{\mathcal{F}}{}^{\Im}[l,k] \\ &\quad - X_{\mathcal{F}}{}^{\Im}[l,k] H_{\mathcal{F}}{}^{\Re}[l,k] + X_{\mathcal{F}}{}^{\Im}[l,k] H_{\mathcal{F}}{}^{\Im}[l,k] \end{aligned} \tag{A.4}$$

Passing this through Eq. (A.2),

$$\begin{aligned} Y'_{\mathcal{F}}[l,k] = &- X_{\mathcal{F}}{}^{\Re}[l,k]H_{\mathcal{F}}{}^{\Re}[l,k] + \mathrm{j}X_{\mathcal{F}}{}^{\Re}[l,k]H_{\mathcal{F}}{}^{\Re}[l,k] \\ &+ \mathrm{j}X_{\mathcal{F}}{}^{\Im}[l,k]H_{\mathcal{F}}{}^{\Re}[l,k] - X_{\mathcal{F}}{}^{\Im}[l,k]H_{\mathcal{F}}{}^{\Im}[l,k] \end{aligned} \tag{A.5}$$

where $Y'_{\mathcal{F}}[l,k]$ is the STFT-equivalent of $Y_{\mathcal{S}}[l,k]$.

Expanding Eq. (A.3a) in terms of real and imaginary components,

$$\begin{aligned} Y_{\mathcal{F}}[l,k] = &X_{\mathcal{F}}{}^{\Re}[l,k]H_{\mathcal{F}}{}^{\Re}[l,k] + \mathrm{j}X_{\mathcal{F}}{}^{\Re}[l,k]H_{\mathcal{F}}{}^{\Im}[l,k] \\ &+ \mathrm{j}X_{\mathcal{F}}{}^{\Im}[l,k]H_{\mathcal{F}}{}^{\Re}[l,k] - X_{\mathcal{F}}{}^{\Im}[l,k]H_{\mathcal{F}}{}^{\Im}[l,k] \end{aligned} \tag{A.6}$$

Comparing Eq. (A.5) and Eq. (A.6), trivially $Y'_{\mathcal{F}}[l,k] \neq Y_{\mathcal{F}}[l,k]$. This proves that the SSBT doesn't appropriately models the convolution, and therefore the convolution theorem doesn't hold when applying this transform.

Appendix B. Correct separation of desired signal

Let $X_m[l,k]$ be such that

$$\begin{aligned} X_m[l,k] &= A_m[l,k] * X_1[l,k] \\ &= \mathbf{a}_m^{\mathsf{T}}[k]\mathbf{x}_1[l,k] \end{aligned} \tag{B.1}$$

as in Eqs. (9) and (10). We can separate $X_m[l,k]$ as

$$X_m[l,k] = d_m[l,k]X_1[l,k] + X'[l,k] \tag{B.2}$$

where $d_m[l,k]$ is the steering vector for the desired speech portion $X_1[l,k]$, and $X'[l,k]$ is the undesired speech component, in such a way that $X_1[l,k]$ and $X'[l,k]$ are uncorrelated.

This seems trivial in a first glance, by adopting $d_m[k]$ as the 0-th element of $\mathbf{a}_m[k]$, and $X'[l,k]$ as the rest of the summation. However, this doesn't take into account that $X[l,k]$ and $X[l',k]$ may be correlated if $|l - l'| \leq S$, where $S = \lfloor (K-1)/\mathcal{O} \rfloor$. That is, if l-th and l'-th transform window share samples, then there is some correlation between them.

We define

$$\begin{aligned} d_m[l,k] &= \frac{\mathsf{E}\{X_1^*[l,k]X_m[l,k]\}}{\mathsf{E}\left\{|X_1[l,k]|^2\right\}} \\ &= \frac{\sum_i A_m[i,k]\mathsf{E}\{X_1^*[l,k]X_1[l-i,k]\}}{\mathsf{E}\left\{|X_1[l,k]|^2\right\}} \end{aligned} \tag{B.3}$$

Focusing on each expectation in the numerator,

$$E_i = \mathsf{E}\{X_1^*[l,k]X_1[l-i,k]\} \tag{B.4}$$

Through the definition of the STFT,

$$\begin{aligned} E_i &= \mathsf{E}\left\{\sum_{n=0}^{K-1} w(n)x_1(n+l\mathcal{O})e^{\mathrm{j}2\pi\frac{k}{K}(n+l\mathcal{O})} \sum_{\nu=0}^{K-1} w(\nu)x_1(\nu+(l-i)\mathcal{O})e^{-\mathrm{j}2\pi\frac{k}{K}(\nu+(l-i)\mathcal{O})}\right\} \\ &= \sum_{n=0}^{K-1}\sum_{\nu=0}^{K-1} w(n)w(\nu)e^{-\mathrm{j}2\pi\frac{k}{K}(\nu-n+i\mathcal{O})}\mathsf{E}\{x_1(n+l\mathcal{O})x_1(\nu+(l-i)\mathcal{O})\} \end{aligned} \tag{B.5}$$

Using the substitutions $\tilde{n} = n + l\mathcal{O}$ and $\tilde{\nu} = \nu + (l-i)\mathcal{O}$,

$$E_i = \sum_{\tilde{\nu}=l\mathcal{O}}^{K-1+l\mathcal{O}} \sum_{\tilde{n}=(l-i)\mathcal{O}}^{K-1+(l-i)\mathcal{O}} w(\tilde{n}-l\mathcal{O})w(\tilde{\nu}-(l-i)\mathcal{O})\mathsf{E}\{x_1(\tilde{n})x_1(\tilde{\nu})\}e^{-\mathrm{j}2\pi\frac{k}{K}(\tilde{\nu}-\tilde{n})} \tag{B.6}$$

We now assume $x_1(n)$ is the result of a zero-mean white process, and therefore $x_1(\tilde{n})$ and $x_1(\tilde{\nu})$ are independent if $\tilde{n} \neq \tilde{\nu}$ (which isn't true, but will allow us to continue the derivations). Then

$$E_i = \sum_{\tilde{n}} w(\tilde{n}-l\mathcal{O})w(\tilde{n}-(l-i)\mathcal{O})\mathsf{E}\left\{x_1(\tilde{n})^2\right\} \tag{B.7}$$

Rolling back the substitution $n = \tilde{n} - l\mathcal{O}$,

$$E_i = \sum_{n=0}^{K-1-|i\mathcal{O}|} w(n)w(n+i\mathcal{O})\mathsf{E}\left\{x_1(n+l\mathcal{O})^2\right\} \tag{B.8}$$

where the summation takes into account that both windows are finite. We at last assume that $\mathsf{E}\left\{x_1(n+l\mathcal{O})^2\right\} \approx \mathsf{E}\left\{x_1(l\mathcal{O})^2\right\}$ for small values of n (such as $0 \leq n < K$), such that

$$E_i = \phi_{x_1}(l\mathcal{O}) \sum_{n=0}^{K-1-|i\mathcal{O}|} w(n)w(n+i\mathcal{O}) \tag{B.9}$$

Going back to Eq. (B.3), and applying the derivations also to the numerator (with $i = 0$),

$$\begin{aligned} d_m[l,k] &= \frac{\sum_i A_m[i,k]\phi_{x_1}(l\mathcal{O}) \sum_{n=0}^{K-1-|i\mathcal{O}|} w(n)w(n+|i\mathcal{O}|)}{\phi_{x_1}(l\mathcal{O}) \sum_{n'=0}^{K-1} w(n')^2} \\ &= \frac{\phi_{x_1}(l\mathcal{O}) \sum_i A_m[i,k] \sum_{n=0}^{K-1-|i\mathcal{O}|} w(n)w(n+|i\mathcal{O}|)}{\phi_{x_1}(l\mathcal{O}) \sum_{n'=0}^{K-1} w(n')^2} \\ &= \frac{\sum_i \sum_{n=0}^{K-1-|i\mathcal{O}|} A_m[i,k]w(n)w(n+|i\mathcal{O}|)}{\sum_{n'=0}^{K-1} w(n')^2} \end{aligned} \tag{B.10}$$

This way, we see that $d_m[l,k] \equiv d_m[k]$ since it doesn't depend on l. However, since for the summation over n we need that $K - 1 - |i\mathcal{O}| \geq 0$, this means that $-S \leq i \leq S$ (where $S = \lfloor (K-1)/\mathcal{O} \rfloor$), and thus our desired speech signal $X_1[l,k]$ depends on up to the previous (or next) S windows.

References

1. Lobato, W.; Costa, M.H. Worst-Case-Optimization Robust-MVDR Beamformer for Stereo Noise Reduction in Hearing Aids. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2020**, *28*, 2224–2237. https://doi.org/10.1109/TASLP.2020.3009831.
2. Chen, J.; Kung Yao.; Hudson, R. Source localization and beamforming. *IEEE Signal Processing Magazine* **2002**, *19*, 30–39. https://doi.org/10.1109/79.985676.
3. Lu, J.Y.; Zou, H.; Greenleaf, J.F. Biomedical ultrasound beam forming. *Ultrasound in Medicine & Biology* **1994**, *20*, 403–428. https://doi.org/10.1016/0301-5629(94)90097-3.
4. Nguyen, N.Q.; Prager, R.W. Minimum Variance Approaches to Ultrasound Pixel-Based Beamforming. *IEEE Transactions on Medical Imaging* **2017**, *36*, 374–384. https://doi.org/10.1109/TMI.2016.2609889.
5. Benesty, J.; Cohen, I.; Chen, J. *Fundamentals of signal enhancement and array signal processing*; John Wiley & Sons: Hoboken, NJ, 2017.

6. Kıymık, M.; Güler, İ.; Dizibüyük, A.; Akın, M. Comparison of STFT and wavelet transform methods in determining epileptic seizure activity in EEG signals for real-time application. *Computers in Biology and Medicine* **2005**, *35*, 603–616. https://doi.org/10.1016/j.compbiomed.2004.05.001.
7. Pan, C.; Chen, J.; Shi, G.; Benesty, J. On microphone array beamforming and insights into the underlying signal models in the short-time-Fourier-transform domain. *The Journal of the Acoustical Society of America* **2021**, *149*, 660–672. https://doi.org/10.1121/10.0003335.
8. Chen, W.; Huang, X. Wavelet-Based Beamforming for High-Speed Rotating Acoustic Source. *IEEE Access* **2018**, *6*, 10231–10239. https://doi.org/10.1109/ACCESS.2018.2795538.
9. Yang, Y.; Peng, Z.K.; Dong, X.J.; Zhang, W.M.; Meng, G. General Parameterized Time-Frequency Transform. *IEEE Transactions on Signal Processing* **2014**, *62*, 2751–2764. https://doi.org/10.1109/TSP.2014.2314061.
10. Almeida, L. The fractional Fourier transform and time-frequency representations. *IEEE Transactions on Signal Processing* **1994**, *42*, 3084–3091. https://doi.org/10.1109/78.330368.
11. Crochiere, R.E.; Rabiner, L.R. *Multirate digital signal processing*; Prentice-Hall signal processing series, Prentice-Hall: Englewood Cliffs, N.J, 1983.
12. Oyzerman, A. Speech Dereverberation in the Time-Frequency Domain. Master's thesis, Technion - Israel Institute of Technology, Haifa, Israel, 2012.
13. Talmon, R.; Cohen, I.; Gannot, S. Relative Transfer Function Identification Using Convolutive Transfer Function Approximation. *IEEE Transactions on Audio, Speech, and Language Processing* **2009**, *17*, 546–555. https://doi.org/10.1109/TASL.2008.2009576.
14. Kumatani, K.; McDonough, J.; Schacht, S.; Klakow, D.; Garner, P.N.; Li, W. Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, March 2008; pp. 1609–1612. https://doi.org/10.1109/ICASSP.2008.4517933.
15. Gopinath, R.; Burrus, C. A tutorial overview of filter banks, wavelets and interrelations. In Proceedings of the 1993 IEEE International Symposium on Circuits and Systems, Chicago, IL, USA, 1993; pp. 104–107. https://doi.org/10.1109/ISCAS.1993.393668.
16. Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE* **1969**, *57*, 1408–1418. https://doi.org/10.1109/PROC.1969.7278.
17. Erdogan, H.; Hershey, J.R.; Watanabe, S.; Mandel, M.I.; Roux, J.L. Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks. In Proceedings of the Interspeech 2016. ISCA, September 2016, pp. 1981–1985. https://doi.org/10.21437/Interspeech.2016-552.
18. DeMuth, G. Frequency domain beamforming techniques. In Proceedings of the ICASSP '77. IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, CT, USA, 1977; Vol. 2, pp. 713–715. https://doi.org/10.1109/ICASSP.1977.1170316.
19. Bai, M.R.; Ih, J.G.; Benesty, J. *Acoustic Array Systems: Theory, Implementation, and Application*, 1 ed.; Wiley, 2013. https://doi.org/10.1002/9780470827253.
20. Habets, E. RIR Generator, 2020.
21. Nielsen, J.K.; Jensen, J.R.; Jensen, S.H.; Christensen, M.G. The Single- and Multichannel Audio Recordings Database (SMARD). In Proceedings of the Int. Workshop Acoustic Signal Enhancement, Sep. 2014.