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.

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:

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 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 being able to obey 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. In Section 5 we present and discuss the results, comparing the studied methods and beamformers obtained. Section 6 concludes this paper.

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

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

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 generic sensor array within a reverberant environment, it being comprised of M sensors. In this setting there also are a desired and an interfering sources (namely $x[n]$ and $v[n]$), and also uncorrelated noise $r_m[n]$ at each sensor, all traveling with a speed c. We assume that the sources are spatially stationary, and all discrete signals were sampled with the same sampling frequency f_s.

We denote $h_m[n]$ as the room impulse response between the desired source and the m-th sensor ($1 \leq m \leq M$). We similarly define $g_m[n]$ for the interfering source. From this, we write $y_m[n]$ as the observed signal at the m-th sensor as

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

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

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

We similarly define $c_m[n]$ such that $c_m[n] * v_1[n] = g_m[n] * v[n]$. Therefore, Eq. (6) becomes

$$y_m[n] = b_m[n] * x_1[n] + c_m[n] * v_1[n] + r_m[n] \tag{8}$$

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

We can use a time-frequency transform (here the STFT or the SSBT, both exposed in Section 2) with the CTF model [13] to get our time-frequency signal model,

$$Y_m[l,k] = B_m[l,k] * X_1[l,k] + C_m[l,k] * V_1[l,k] + R_m[l,k] \tag{9}$$

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

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

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

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

in which

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

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

and in the same way we define $\mathbf{c}_m[k]$ and $\mathbf{v}_1[l,k]$. Note that $\mathbf{b}_m[k]$ and $\mathbf{c}_m[k]$ don't depend on the index l, since neither the environment nor the sources' positions change over time. With this, Eq. (9) becomes

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

Vectorizing the signals sensor-wise, we finally get

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

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, $\mathbf{B}[k]$ and $\mathbf{C}[k]$ are $M \times L_B$ and $M \times L_C$ matrices respectively; $\mathbf{x}_1[l,k]$ and $\mathbf{v}_1[l,k]$ are $L_B \times 1$ 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-rejecting formulation*

Let the Δ-th column of $\mathbf{B}[k]$ (equivalent to $l = 0$) be the desired-speech frequency response (named $\mathbf{d}_x[k]$), with the rest comprising an undesired component. We therefore write

$$\begin{aligned}\mathbf{B}[k]\mathbf{x}_1[l,k] &= \mathbf{d}_x[k]X_1[l,k] + \sum_{\substack{l'=-\Delta \\ l'\neq 0}}^{L_B-\Delta-1} \mathbf{p}_{B,l'}[k]X_1[l-l',k] \\ &= \mathbf{d}_x[k]X_1[l,k] + \mathbf{q}[l,k]\end{aligned} \tag{15}$$

where $\mathbf{p}_{B,l'}[k]$ is the l'-th column of $\mathbf{B}[k]$. With this, $\mathbf{d}_x[k]X_1[l,k]$ is the desired speech component of $\mathbf{B}[k]\mathbf{x}_1[l,k]$ with $\mathbf{d}_x[k]$ being the desired-speech frequency response; and $\mathbf{q}[l,k]$ (the summation over $l \neq 0$) is the undesired component, or reverberation signal.

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), multiple windows may represent the desired speech. This isn't a problem if $\frac{\delta}{c} < \frac{K}{f_s}$, where δ is the biggest reference-to-sensor distance, and K is the window length.

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

$$\mathbf{w}[l,k] = \mathbf{q}[l,k] + \mathbf{C}[k]\mathbf{v}_1[l,k] + \mathbf{r}[l,k] \tag{16}$$

and therefore

$$\mathbf{y}[l,k] = \mathbf{d}_x[k]X_1[l,k] + \mathbf{w}[l,k] \tag{17}$$

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

$$\begin{aligned} Z[l,k] &= \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{y}[l,k] \\ &\approx X_1[l,k] \end{aligned} \tag{18}$$

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

In order to minimize $\mathbf{w}[l,k]$ the MVDR beamformer [17] will be used, being given by

$$\mathbf{f}_{\mathrm{mvdr}}[l,k] = \min_{\mathbf{f}[l,k]} \mathbf{f}[l,k]^{\mathsf{H}}\mathbf{\Phi}_{\mathbf{y}}[l,k]\mathbf{f}[l,k] \text{ s.t. } \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k] = 1 \tag{19}$$

in which $\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k] = 1$ is the distortionless constraint, and $\mathbf{\Phi}_{\mathbf{y}}[l,k]$ is the correlation matrix of the observed signal $\mathbf{y}[l,k]$.

The solution to Eq. (19) is

$$\mathbf{f}_{\mathrm{mvdr}}[l,k] = \frac{\mathbf{\Phi}_{\mathbf{y}}^{-1}[l,k]\mathbf{d}_x[k]}{\mathbf{d}_x^{\mathsf{H}}[k]\mathbf{\Phi}_{\mathbf{y}}^{-1}[l,k]\mathbf{d}_x[k]} \tag{20}$$

Equivalently, the MVDR beamformer can be written as

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

with $\mathbf{\Phi}_{\mathbf{w}}[l,k]$ being the correlation matrix of $\mathbf{w}[l,k]$. This formulation is advantageous given that $\mathbf{w}[l,k]$ and $X_1[l,k]$ are correlated, since different windows of $\mathbf{B}[k]\mathbf{x}_1[l,k]$ aren't independent due to their overlap.

3.2. Beamformer metrics

The metrics which will be used to observe and study the results will be the gain in Signal-to-Noise Ratio (SNR), where "noise" is $\mathbf{w}[l,k]$, encompassing all undesired signals; Reverberation Signal Reduction Factor ξ_{r}, which shows how much the undesired speech components were reduced; and Desired Signal Reduction Factor ξ_{d}, which explores how much the desired speech components were reduced; in both window-averaged narrowband (Eq. (22)) and window-averaged broadband (Eq. (23)) forms.

$$\mathrm{gSNR}[k] = \frac{\sum_l \phi_{X_1}[l,k]\left|\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k]\right|^2}{\sum_l \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{\Phi}_{\mathbf{w}}[l,k]\mathbf{f}[l,k]} \div \frac{\sum_l \phi_{X_1}[l,k]}{\sum_l \phi_{W_1}[l,k]} \tag{22a}$$

$$\xi_{\mathrm{r}}[k] = \frac{\sum_l \phi_{Q_1}[l,k]}{\sum_l \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{\Phi}_{\mathbf{q}}[l,k]\mathbf{f}[l,k]} \tag{22b}$$

$$\xi_{\mathrm{d}}[k] = \frac{\sum_l \phi_{X_1}[l,k]}{\sum_l \phi_{X_1}[l,k]\left|\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k]\right|^2} \tag{22c}$$

$$\mathrm{gSNR} = \frac{\sum\limits_{l,k} \phi_{X_1}[l,k]\left|\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k]\right|^2}{\sum\limits_{l,k} \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{\Phi}_{\mathbf{w}}[l,k]\mathbf{f}[l,k]} \div \frac{\sum\limits_{l,k} \phi_{X_1}[l,k]}{\sum\limits_{l,k} \phi_{W_1}[l,k]} \tag{23a}$$

$$\xi_{\mathrm{r}} = \frac{\sum\limits_{l,k} \phi_{Q_1}[l,k]}{\sum\limits_{l,k} \mathbf{f}^{\mathsf{H}}[l,k]\boldsymbol{\Phi}_{\mathbf{q}}[l,k]\mathbf{f}[l,k]} \tag{23b}$$

$$\xi_{\mathrm{d}} = \frac{\sum\limits_{l,k} \phi_{X_1}[l,k]}{\sum\limits_{l,k} \phi_{X_1}[l,k]\left|\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k]\right|^2} \tag{23c}$$

where $Q_1[l,k]$ is the first element (corresponding to the reference sensor) of $\mathbf{q}[l,k]$. It is easy to see that if $\xi_{\mathrm{d}}[l,k] = 1$ then the distortionless constraint is satisfied. Also, $\xi_{\mathrm{d}}[l,k] = 1$ implies that $\xi_{\mathrm{d}}[k] = \xi_{\mathrm{d}} = 1$.

4. True-MVDR 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 both bins k and $K-k$ simultaneously, since from Eq. (5) they both contribute to each the k-th bin in the STFT domain. We define $\mathbf{w}'[l,k]$ as

$$\mathbf{w}'[l,k] = \begin{bmatrix} \mathbf{w}[l,k] \\ \mathbf{w}[l,K-k] \end{bmatrix}_{2M\times 1} \tag{24}$$

from which we define $\boldsymbol{\Phi}_{\mathbf{w}'}[l,k]$ 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

$$\mathbf{f}'_{\mathcal{F}}[l,k] = \mathbf{A}\mathbf{f}'[l,k] \tag{25}$$

in which

$$\mathbf{A} = \frac{1}{\sqrt{2}} \begin{bmatrix} e^{-\mathrm{j}\frac{3\pi}{4}} & 0 & \cdots & 0 & e^{\mathrm{j}\frac{3\pi}{4}} & 0 & \cdots & 0 \\ 0 & e^{-\mathrm{j}\frac{3\pi}{4}} & \cdots & 0 & 0 & e^{\mathrm{j}\frac{3\pi}{4}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \cdots & e^{-\mathrm{j}\frac{3\pi}{4}} & 0 & 0 & \cdots & e^{\mathrm{j}\frac{3\pi}{4}} \end{bmatrix}_{M\times 2M} \tag{26}$$

4.1. *Reverb-aware SSBT formulation*

Continuing the results of Section 3.1, from Eq. (25) the distortionless constraint for the STFT, within the SSBT domain, is

$$\mathbf{f}'^{\mathsf{H}}[l,k]\hat{\mathbf{d}}_x[k] = 1 \tag{27a}$$

$$\hat{\mathbf{d}}_x[k] = \mathbf{A}^{\mathsf{H}}\mathbf{d}_{\mathcal{F};x}[k] \tag{27b}$$

where $\mathbf{d}_{\mathcal{F};x}[l,k]$ is the desired-speech frequency response in the STFT domain. In this scheme, our minimization problem becomes

$$\mathbf{f}'_{\mathrm{mvdr}}[l,k] = \min_{\mathbf{f}'[l,k]} \mathbf{f}'^{\mathsf{H}}[l,k]\boldsymbol{\Phi}_{\mathbf{w}'}[l,k]\mathbf{f}'[l,k] \text{ s.t. } \mathbf{f}'^{\mathsf{H}}[l,k]\hat{\mathbf{d}}_x[k] = 1 \tag{28}$$

Although $\mathbf{\Phi}_{\mathbf{w}'}[l,k]$ is a matrix with real entries, $\hat{\mathbf{d}}_x$ is complex-valued, and thus is the solution to Eq. (28), contradicting the purpose of utilizing the SSBT.

4.2. Real-valued true-MVDR 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. (27a) we trivially have that

$$\mathbf{f}'^{\mathsf{T}}[l,k]\Re\left\{\hat{\mathbf{d}}_x[k]\right\} = 1 \tag{29a}$$

$$\mathbf{f}'^{\mathsf{T}}[l,k]\Im\left\{\hat{\mathbf{d}}_x[k]\right\} = 0 \tag{29b}$$

which can be put in matricial form as $\mathbf{f}'^{\mathsf{T}}[l,k]\mathbf{D}_x[k] = \mathbf{i}^{\mathsf{T}}$, with

$$\mathbf{D}_x[k] = \left[\ \Re\left\{\hat{\mathbf{d}}_x[k]\right\},\ \Im\left\{\hat{\mathbf{d}}_x[k]\right\}\ \right]_{2M\times 2} \tag{30a}$$

$$\mathbf{i} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{30b}$$

Therefore, the minimization problem from Eq. (28) becomes

$$\mathbf{f}'_{\mathrm{mvdr}}[l,k] = \min_{\mathbf{f}'[l,k]} \mathbf{f}'^{\mathsf{T}}[l,k]\mathbf{\Phi}_{\mathbf{w}'}[l,k]\mathbf{f}'[l,k] \text{ s.t. } \mathbf{f}'^{\mathsf{T}}[l,k]\mathbf{D}_x[k] = \mathbf{i}^{\mathsf{T}} \tag{31}$$

whose formulation is again similar to that of the LCMV beamformer, and therefore

$$\mathbf{f}'_{\mathrm{mvdr}}[l,k] = \mathbf{\Phi}_{\mathbf{w}'}^{-1}[l,k]\mathbf{D}_x[k]\left(\mathbf{D}_x^{\mathsf{T}}[k]\mathbf{\Phi}_{\mathbf{w}'}^{-1}[k]\mathbf{D}_x[k]\right)^{-1}\mathbf{i} \tag{32}$$

Using Eq. (25), we can obtain the desired beamformer $\mathbf{f}'_{\mathcal{F};\mathrm{mvdr}}[l,k]$, in the STFT domain.

5. Simulations

In the simulations[2], we employ a sampling frequency of 16kHz. The sensor array consists of a uniform linear array with 10 sensors spaced at 2cm. Room impulse responses were generated using Habets' RIR generator [19], and signals were selected from the SMARD [20] and LINSE [21] databases.

The room's dimensions are 4m × 6m × 3m (width × length × height), with a reverberation time of 0.11s. The desired source is located at $(2\text{m},\ 1\text{m},\ 1\text{m})$, it being a male voice (SMARD, `50_male_speech_english_ch8_OmniPower4296.flac`). The interfering source, simulating an open door, is located simultaneously at $(0.5\text{m},\ 5\text{m},\ [0.3:0.3:2.7]\text{m})$, with a babble sound signal (LINSE database, `babble.mat`). The noise signal is white Gaussian noise (SMARD database, `wgn_48kHz_ch8_OmniPower4296.flac`). All signals were resampled to the desired frequency.

The sensor array is positioned at $(2\text{m},\ [4.02:0.02:4.2]\text{m},\ 1\text{m})$, with omnidirectional sensors of flat frequency response. The input SNR between desired and noise signals is 30dB, while the input SNR between desired and interference signals (or input SIR) varies per simulation. Filters are calculated every 25 windows, considering the previous 25 windows to calculate correlation matrices.

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

We compare filters obtained through the STFT and SSBT transforms. N-SSBT uses Eq. (20) to (naively) calculate the SSBT beamformer, and T-SSBT will denote the beamformer obtained via the true-distortionless MVDR from Section 4. Performance analysis is conducted via the STFT domain, with the SSBT beamformers being converted into it. In line plots, STFT is presented in red, N-SSBT in green, and T-SSBT in blue.

5.1. *Simulations with* iSIR = 5dB

In this scenario, we assume that the input SIR is 5dB, and we use analysis windows with: 32 samples in Figs. 1 and 2; or 64 samples in Figs. 3 and 4. Figs. 1 and 3 show the window-wise averaged narrowband gain in SNR, and Figs. 2 and 4 the window-averaged narrowband DSRF, for all presented methods.

From Figs. 1 and 3 it is clear that both beamformers derived through the SSBT outperformed the STFT beamformer in terms of SNR gain, with the N-SSBT one having a slightly better yield than the T-SSBT one. Also, both Figs. 2 and 4 show that the T-SSBT and the STFT beamformers ensured a distortionless response, a feature that wasn't achieved by the N-SSBT beamformer. This was expected, since the T-SSBT was appropriately designed to achieve this quality, while the N-SSBT wasn't.

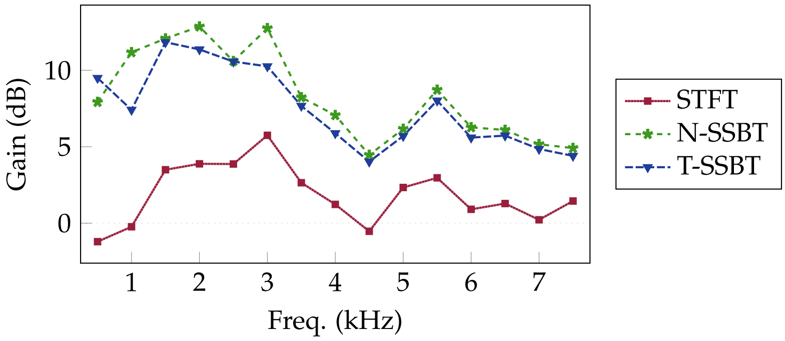

Figure 1. Window-average SNR gain for $K = 32$.

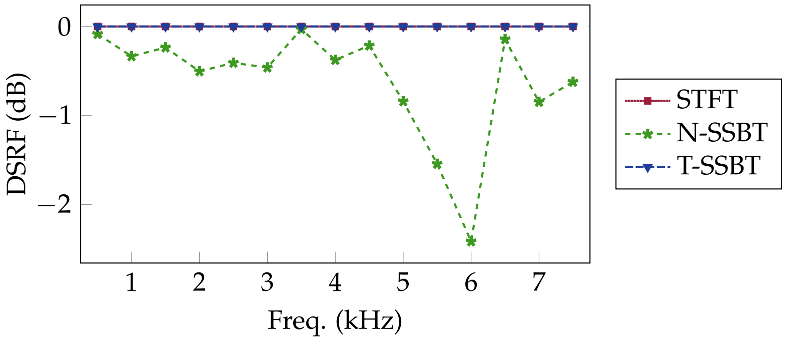

Figure 2. Window-average DSRF for $K = 32$.

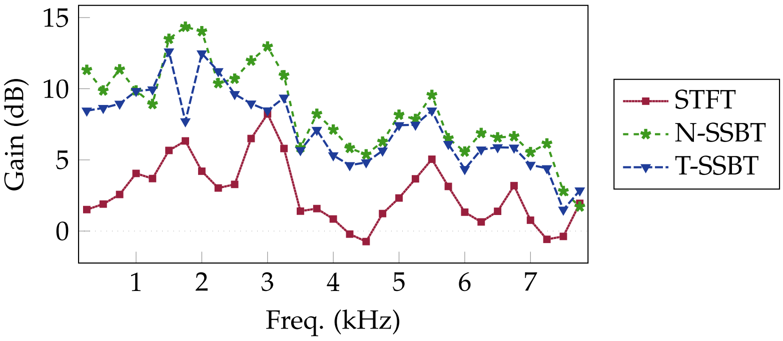

Figure 3. Window-average SNR gain for $K = 64$.

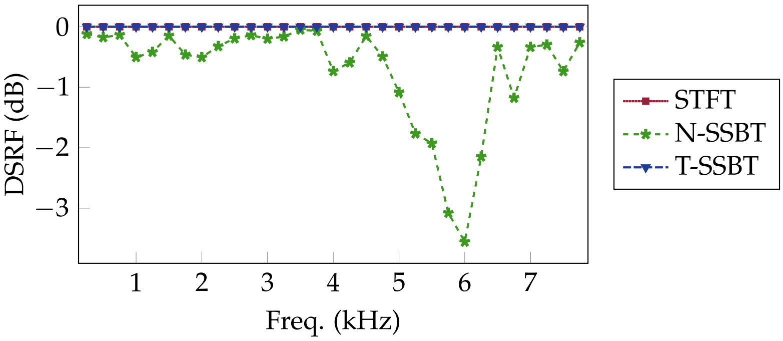

Figure 4. Window-average DSRF for $K = 64$.

5.2. Average results per iSIR 202

Here, we now take the broadband metrics for the beamformers, allowing us to compare them for different values of input SIR. We again test for both $K = 32$ and $K = 64$. 203 204

From Figs. 5 and 6, as expected we see that the STFT and T-SSBT beamformers didn't cause distortion on the desired signal, while the N-SSBT did, due to its nature. In terms of SNR gain (Figs. 7 and 8), we again see that the SSBT beamformers led to a strictly better result than the STFT one, the later being 6-7 dB worse than the former ones for all input SIRs. 205 206 207 208 209

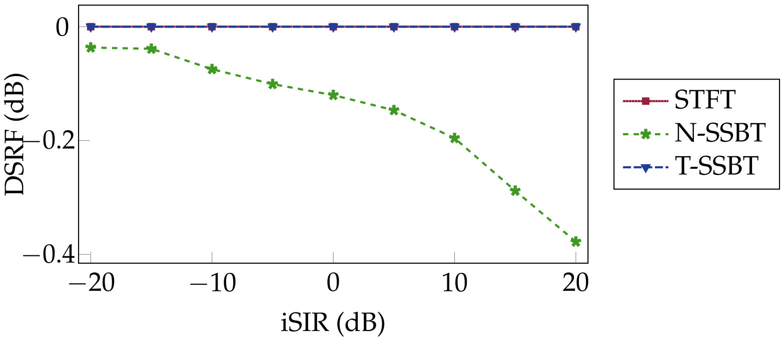

Figure 5. Broadband DSRF for $K = 32$.

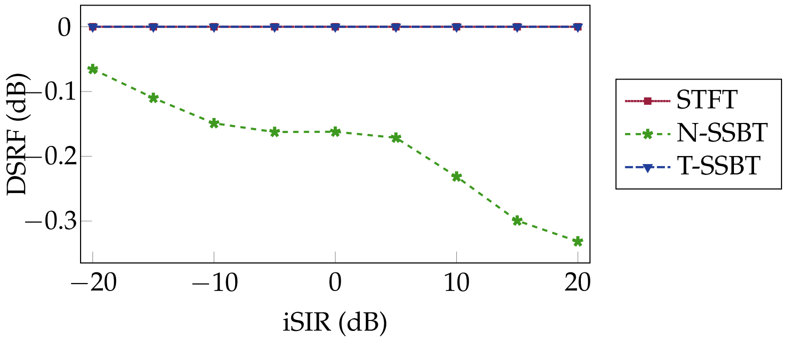

Figure 6. Broadband DSRF for $K = 64$.

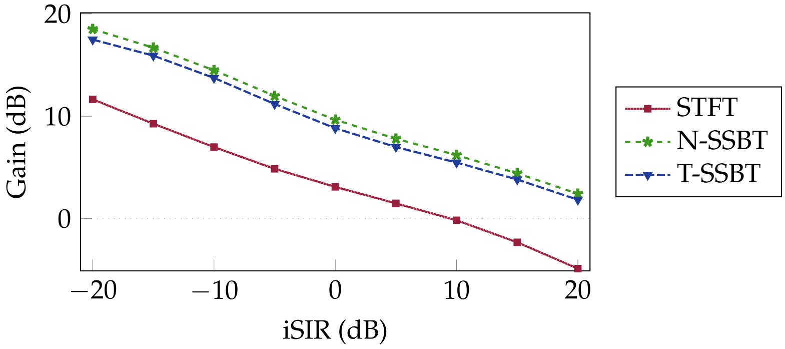

Figure 7. Broadband SNR gain for $K = 32$.

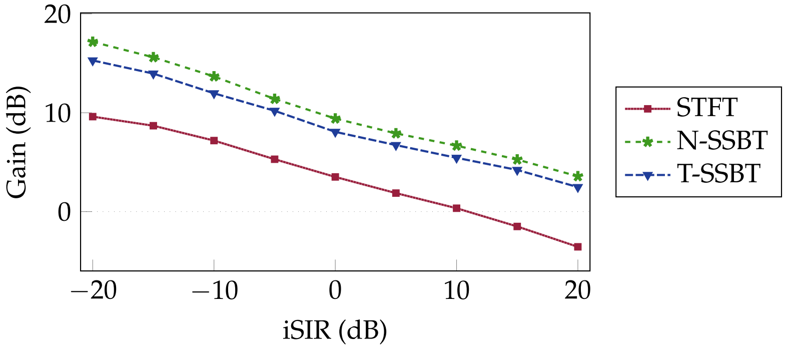

Figure 8. Broadband SNR gain for $K = 64$.

5.3. Overall result

In all situations evaluated, the proposed true-MVDR SSBT beamformer consistently outperformed the beamformer through the traditional STFT in terms of SNR gain, both achieving the appropriate DSRF necessary for the desired distortionless behavior of the MVDR filter. Meanwhile, the naive SSBT beamformer had a (slightly) better overall than that of the T-SSBT beamformer for both evaluated K, while also incurring some undesired distortion on the desired signal seen via the DSRF. It is interesting to note that for all beamformers the gain in SNR decreases as the input SIR increases, which was expected as in this scenario the weight of both the undesired speech components and uncorrelated noise over the input SNR increase.

6. Conclusion

In this study, we investigated the application of the Single-Sideband Transform in beamforming within a reverberant environment, utilizing the convolutive transfer function model for filter bank (i.e., the beamformer) estimation. We implemented a Minimum-Variance Distortionless-Response beamformer to enhance signals in a real-life-like scenario, elucidating the process to achieve a truly-distortionless MVDR beamformer when employing the SSB transform.

The true-MVDR SSBT beamformer strictly and consistently outperformed the beamformer obtained with the traditional STFT in terms of SNR gain, matching it regarding the distortionlessness. The naive-MVDR SSBT beamformer's SNR gain performance was slightly better than that of the true-MVDR one, and in all situations causing distortion in the desired signal.

Future research avenues may explore the integration of this transform into different beamformers, or undertake further comparisons of the proposed SSBT beamformer (following the considerations exposed in here) against the established and reliable STFT methodology.

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
LCMV	Linearly-Constrained Minimum-Variance
MVDR	Minimum-Variance 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

$$X_{\mathcal{S}}[l,k] = -X_{\mathcal{F}}{}^{\Re}[l,k] - X_{\mathcal{F}}{}^{\Im}[l,k] \tag{A1}$$

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

It is easy to see that

$$X_{\mathcal{F}}[l,k] = \frac{1}{\sqrt{2}}\left(e^{-\mathrm{j}\frac{3\pi}{4}} X_{\mathcal{S}}[l,k] + e^{\mathrm{j}\frac{3\pi}{4}} X_{\mathcal{S}}[l,K-k]\right) \tag{A2}$$

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

$$Y_{\mathcal{F}}[l,k] = H_{\mathcal{F}}[k] X_{\mathcal{F}}[l,k] \tag{A3a}$$

$$Y_{\mathcal{S}}[l,k] = H_{\mathcal{S}}[k] X_{\mathcal{S}}[l,k] \tag{A3b}$$

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. (A1) in Eq. (A3b), 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

$$\begin{aligned}
Y_{\mathcal{S}}[l,k] &= {X_{\mathcal{F}}}^{\Re}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] + {X_{\mathcal{F}}}^{\Re}[l,k]{H_{\mathcal{F}}}^{\Im}[l,k] \\
&\quad + {X_{\mathcal{F}}}^{\Im}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] + {X_{\mathcal{F}}}^{\Im}[l,k]{H_{\mathcal{F}}}^{\Im}[l,k] \\
Y_{\mathcal{S}}[l,K-k] &= {X_{\mathcal{F}}}^{\Re}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] - {X_{\mathcal{F}}}^{\Re}[l,k]{H_{\mathcal{F}}}^{\Im}[l,k] \\
&\quad - {X_{\mathcal{F}}}^{\Im}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] + {X_{\mathcal{F}}}^{\Im}[l,k]{H_{\mathcal{F}}}^{\Im}[l,k]
\end{aligned} \tag{A4}$$

Passing this through Eq. (A2),

$$\begin{aligned}
Y'_{\mathcal{F}}[l,k] = &- {X_{\mathcal{F}}}^{\Re}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] + \mathrm{j}{X_{\mathcal{F}}}^{\Re}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] \\
&+ \mathrm{j}{X_{\mathcal{F}}}^{\Im}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] - {X_{\mathcal{F}}}^{\Im}[l,k]{H_{\mathcal{F}}}^{\Im}[l,k]
\end{aligned} \tag{A5}$$

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

Expanding Eq. (A3a) in terms of real and imaginary components,

$$\begin{aligned}
Y_{\mathcal{F}}[l,k] &= {X_{\mathcal{F}}}^{\Re}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] + \mathrm{j}{X_{\mathcal{F}}}^{\Re}[l,k]{H_{\mathcal{F}}}^{\Im}[l,k] \\
&\quad + \mathrm{j}{X_{\mathcal{F}}}^{\Im}[l,k]{H_{\mathcal{F}}}^{\Re}[l,k] - {X_{\mathcal{F}}}^{\Im}[l,k]{H_{\mathcal{F}}}^{\Im}[l,k]
\end{aligned} \tag{A6}$$

Comparing Eq. (A5) and Eq. (A6), 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.

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. Habets, E. RIR Generator, 2020.
20. 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.
21. Johnson, D.H. Signal Processing Information Database, 2013.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.