Article

[DRAFT] On the Single-Sideband Transform for MVDR Beamformers

Vitor 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: 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. Our study aims to optimize the 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 results in a similar performance to when using 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, having a plethora of applications from hearing aids [1] to source localization [2] to imaging [3,4]. One of the possible ways to use beamformers is in the time-frequency domain [5], allowing the exploitation of frequency-related information while also dynamically adapting to signal changes over time. While the Short-Time Fourier Transform (STFT) is widely used for time-frequency analysis [6,7], alternative transforms [8–10] can also be employed, offering unique perspectives on signal analysis.

Among these, the Single-Sideband Transform (SSBT) [11,12] stands out for its real-valued frequency spectrum. It has been shown that the SSBT works particularly well with short analysis windows [11], lending itself useful when working with the convolutive transfer function (CTF) model [13] and filter-banks [14,15] for signal analysis.

Two of the most important goals in beamforming are output noise minimization 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 of

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

Received:
Revised:
Accepted:
Published:

interest 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 has a similar performance as an 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

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

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) it's easy to see that

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

From Eq. (3) it is possible to show[1] that

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

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 first 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 ($1 \le m \le M$). 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

$$y_m[n] = h_m[n] * x[n] + g_m[n] * v[n] + r_m[n] \tag{5}$$

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

$$b_m[n] * x_1[n] = h_m[n] * x[n] \tag{6}$$

We similarly define $c_m[n]$ such that $c_m[n] * v_1[n] = g_m[n] * v[n]$. Therefore, Eq. (5) becomes

$$y_m[n] = b_m[n] * x_1[n] + c_m[n] * v_1[n] + r_m[n] \tag{7}$$

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,

$$Y_m[l,k] = B_m[l,k] * X_1[l,k] + C_m[l,k] * V_1[l,k] + R_m[l,k] \tag{8}$$

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 $0 \le k \le K-1$; 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

$$B_m[l,k] * X_1[l,k] = \mathbf{b}_m^{\mathsf{T}}[k]\mathbf{x}_1[l,k] \tag{9}$$

in which

$$\mathbf{b}_m[k] = \Big[\, B_m[-\Delta,k]\,,\ \cdots\,,\ B_m[0,k]\,,\ \cdots\,,\ B_m[L_B-\Delta-1,k] \,\Big]^{\mathsf{T}} \tag{10a}$$

[1] Eq. (4) (and its derivations) is invalid for $k = 0$, and $k = K/2$ if K is even. However in these cases $X_{\mathcal{F}}[l,k] = X_{\mathcal{S}}[l,k]$, and the "naive" SSBT beamformer from Section 3 works.

$$\mathbf{x}_1[l,k] = \begin{bmatrix} X_1[l+\Delta,k]\,, & \cdots, & X_1[l,k]\,, & \cdots, & X_1[l-L_B+\Delta+1,k] \end{bmatrix}^{\mathsf{T}} \tag{10b}$$

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. (8) becomes

$$Y_m[l,k] = \mathbf{b}_m^{\mathsf{T}}[k]\mathbf{x}_1[l,k] + \mathbf{c}_m^{\mathsf{T}}[k]\mathbf{v}_1[l,k] + R_m[l,k] \tag{11}$$

Vectorizing the signals sensor-wise, we finally get

$$\mathbf{y}[l,k] = \mathbf{B}^{\mathsf{T}}[k]\mathbf{x}_1[l,k] + \mathbf{C}^{\mathsf{T}}[k]\mathbf{v}_1[l,k] + \mathbf{r}[l,k] \tag{12}$$

where

$$\mathbf{y}[l,k] = \begin{bmatrix} y_1[l,k]\,, & \cdots, & y_M[l,k] \end{bmatrix}^{\mathsf{T}} \tag{13}$$

and similarly for the other variables. In this situation, $\mathbf{B}[k]$ and $\mathbf{C}[k]$ are $L_B \times M$ and $L_C \times M$ 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-aware formulation*

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

$$\mathbf{B}^{\mathsf{T}}[k]\mathbf{x}_1[l,k] = \mathbf{d}_x[k]X_1[l,k] + \sum_{\substack{l'=-\Delta \\ l'\neq 0}}^{L_B-\Delta-1} \mathbf{p}_{B,l'}[k]X_1[l-l',k] \tag{14}$$

where $\mathbf{p}_{B,l'}[k]$ is the l'-th row of $\mathbf{B}[k]$. With this, $\mathbf{d}_x[k]X_1[l,k]$ is the desired speech component of $\mathbf{B}^{\mathsf{T}}[k]\mathbf{x}_1[l,k]$, and the summation over l' is the undesired component. We will call $\mathbf{d}_x[k]$ the desired-speech frequency response.

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

$$\mathbf{w}[l,k] = \sum_{\substack{l'=-\Delta \\ l'\neq 0}}^{L_B-\Delta-1} \mathbf{p}_{B,l'}[k]X_1[l-l',k] + \mathbf{C}^{\mathsf{T}}[k]\mathbf{v}_1[l,k] + \mathbf{r}[l,k] \tag{15}$$

and therefore

$$\mathbf{y}[l,k] = \mathbf{d}_x[k]X_1[l,k] + \mathbf{w}[l,k] \tag{16}$$

For simplicity we assume that all windows of $X_1[l,k]$ are independent of one-another, and that $\mathbf{w}[l,k]$ is independent of $\mathbf{d}_x[k]X_1[l,k]$. This isn't strictly true, given both the time-frequency windowing process and the reverberant behavior of the environment.

3.2. MVDR beamformer

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

$$\begin{aligned} Z[l,k] &= \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{y}[l,k] \\ &\approx X_1[l,k] \end{aligned} \tag{17}$$

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

$$\mathbf{f}_{\mathrm{mvdr}}[l,k] = \min_{\mathbf{f}[l,k]} \mathbf{f}[l,k]^{\mathsf{H}}\mathbf{\Phi}_{\mathbf{y}}[l,k]\mathbf{f}[l,k] \text{ s.t. } \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k] = 1 \tag{18}$$

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 $\mathbf{y}[l,k]$. This formulation is preferred over the one using $\mathbf{\Phi}_{\mathbf{w}}[l,k]$ given the observed signal's availability.

The solution to Eq. (18) is

$$\mathbf{f}_{\mathrm{mvdr}}[l,k] = \frac{\mathbf{\Phi}_{\mathbf{y}}^{-1}[l,k]\mathbf{d}_x[k]}{\mathbf{d}_x^{\mathsf{H}}[k]\mathbf{\Phi}_{\mathbf{y}}^{-1}[l,k]\mathbf{d}_x[k]} \tag{19}$$

3.3. Beamformer metrics

Considering the problem, the relevant metrics are the narrowband gain in signal-to-noise ratio (SNR) and narrowband desired signal distortion index (DSDI), respectively given by

$$\Delta\mathrm{SNR}[l,k] = \phi_{V_1}[l,k]\frac{\left|\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k]\right|^2}{\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{\Phi}_{\mathbf{w}}[l,k]\mathbf{f}[l,k]} \tag{20}$$

$$v[l,k] = \left|\mathbf{f}^{\mathsf{H}}[l,k]\mathbf{d}_x[k] - 1\right|^2 \tag{21}$$

We can also define the window-averaged gain in SNR and DSDI as

$$\Delta\mathrm{SNR}[k] = \frac{1}{L_Z}\sum_{l=0}^{L_Z-1} \mathrm{gSNR}[l,k] \tag{22}$$

$$v[k] = \frac{1}{L_Z}\sum_{l=0}^{L_Z-1} v[l,k] \tag{23}$$

with L_Z being the number of windows of $Z[l,k]$.

4. True-MVDR with the Single-Sideband Transform

When carelessly using Eqs. (18) and (19) with the SSBT, the distortionless constraint ensures that the beamformer avoids causing distortion exclusively within the SSBT domain. However, as explained towards the end of Section 2, the SSBT beamformer must undergo conversion into the STFT domain (via Eq. (4)) before filtering due to the transform's limitations. To construct an SSBT beamformer that correctly adheres to the distortionless constraint, it is essential to consider this conversion step.

We thus propose a framework for the SSBT in which we consider both bins k and $K-k$ simultaneously, since from Eq. (3) they aren't independent. We define $\mathbf{y}'[l,k]$ as

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

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. With Eq. (4) it is easy to see that

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

where $\hat{\mathbf{f}}_{\mathcal{F}}[l,k]$ is the STFT-equivalent beamformer for $\mathbf{f}[l,k]$, and $\hat{\mathbf{A}}$ is

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

From Eq. (25), the distortionless constraint for the STFT, within the SSBT domain, is

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

where $\mathbf{d}_{\mathcal{F};x}[l,k]$ is the desired-speech frequency response in the STFT domain; and

$$\mathbf{D}_x[k] = \hat{\mathbf{A}}^{\mathsf{H}}\mathbf{d}_{\mathcal{F};x}[k] \tag{28}$$

In this scheme, our minimization problem becomes

$$\mathbf{f}'_{\mathrm{mvdr}}[l,k] = \min_{\mathbf{f}'[l,k]} \mathbf{f}'^{\mathsf{H}}[l,k]\mathbf{\Phi}_{\mathbf{y}'}[l,k]\mathbf{f}'[l,k] \text{ s.t. } \mathbf{f}'^{\mathsf{H}}[l,k]\mathbf{D}_x[k] = 1 \tag{29}$$

Although $\mathbf{\Phi}_{\mathbf{y}'}[l,k]$ is a matrix with real entries, $\mathbf{D}_x$ is complex-valued, and thus is the solution to Eq. (29), contradicting the purpose of utilizing the SSBT.

4.1. *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. (27) we trivially have that

$$\mathbf{f}'^{\mathsf{T}}[l,k]\Re\{\mathbf{D}_x[k]\} = 1 \tag{30a}$$
$$\mathbf{f}'^{\mathsf{T}}[l,k]\Im\{\mathbf{D}_x[k]\} = 0 \tag{30b}$$

which can be put in matricial form,

$$\mathbf{f}'^{\mathsf{T}}[l,k]\mathbf{Q}_x[k] = \mathbf{i}^{\mathsf{T}} \tag{31}$$

with

$$\mathbf{Q}_x[k] = \Big[\, \Re\{\mathbf{D}_x[k]\},\ \Im\{\mathbf{D}_x[k]\} \,\Big]_{2M\times 2} \tag{32a}$$
$$\mathbf{i} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{32b}$$

Therefore, the minimization problem from Eq. (29) becomes

$$\mathbf{f}'_{\mathrm{mvdr}}[l,k] = \min_{\mathbf{f}'[l,k]} \mathbf{f}'^{\mathsf{T}}[l,k]\mathbf{\Phi}_{\mathbf{y}'}[l,k]\mathbf{f}'[l,k] \text{ s.t. } \mathbf{f}'^{\mathsf{T}}[l,k]\mathbf{Q}_x[k] = \mathbf{i}^{\mathsf{T}} \tag{33}$$

whose formulation is the same as the linearly-constrained minimum variance (LCMV) [19] beamformer, and thus has the solution

$$\mathbf{f}'_{\mathrm{mvdr}}[l,k] = \mathbf{\Phi}^{-1}_{\mathbf{y}'}[l,k]\mathbf{Q}_x[k]\left(\mathbf{Q}^{\mathsf{T}}_x[k]\mathbf{\Phi}^{-1}_{\mathbf{y}'}[k]\mathbf{Q}_x[k]\right)^{-1}\mathbf{i} \tag{34}$$

Using Eq. (25), we can obtain the desired beamformer $\hat{\mathbf{f}}_{\mathcal{F};\mathrm{mvdr}}[l,k]$, 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 [20], and signals were selected from the SMARD [21] and LINSE [22] databases.

The room's dimensions are 4m $\times$ 6m $\times$ 3m (width $\times$ length $\times$ 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 interfering signals is 5dB, and between desired and noise signals is 30dB. Filters are calculated every 25 windows, considering the previous 25 windows to calculate correlation matrices.

We compare filters obtained through the STFT and SSBT transforms. N-SSBT uses Eq. (19) 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. Results - 32 samples/window

In this scenario, we assume that the analysis windows have 32 samples. Fig. 1 shows the window-wise averaged gain in SNR, and in Fig. 2 we have the window-averaged DSDI, for all three methods.

Fig. 1 shows that all three beamformers had a similar performance, with no clear advantage of one to the other on the overall spectrum. Also, Fig. 2 shows that both 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.

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

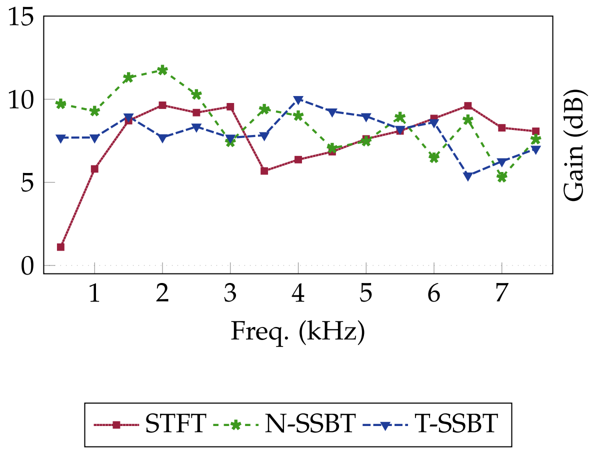

Figure 1. Window-average SNR gain for 32 samples/window.

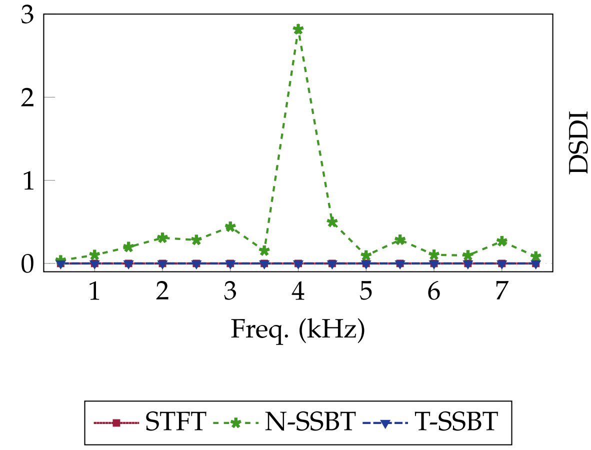

Figure 2. Window-average DSDI for 32 samples/window.

5.2. Results - 64 samples/window

In this simulation, we changed the number of samples per window from 32 to 64, keeping everything else the same.

Differently than the last situation, from Fig. 3 we see that the N-SSBT had a worse overall performance, driven by the higher frequencies. The STFT and T-SSBT beamformers' performances were similar, with one being more advantageous than the other in specific frequencies, but with overall similar output. In Fig. 4 we have an akin result to that of Fig. 2, with the N-SSBT being the only one to distort the desired signal.

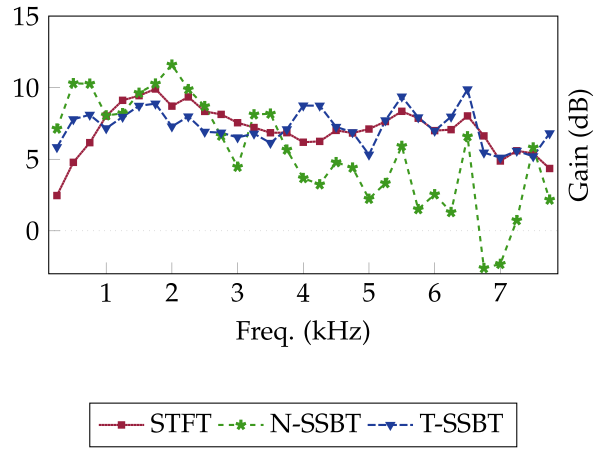

Figure 3. Window-average SNR gain for 64 samples/window.

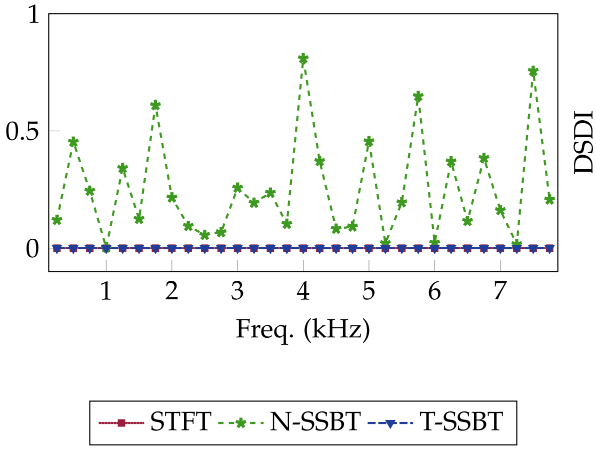

Figure 4. Window-average DSDI for 64 samples/window.

5.3. Overall result

In both situations simulated, the proposed T-SSBT and the traditional STFT beamformers had a similar performance in terms of gain

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 true-distortionless MVDR beamformer when employing the SSB transform. The true-MVDR SSBT beamformer matched the performance of the beamformer for the traditional STFT, in terms of SNR gain and distortionless response. The naive-MVDR SSBT beamformer's SNR gain performance varied with the number of samples per transform window, matching the others for fewer samples, but being outperformed for more. In both cases, it caused distortion in the desired signal in both cases.

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
DSDI	Desired Signal Distortion Index
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

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

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

It is easy to see that

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

As stated before, this formulation isn't valid for $k=0$ and $k={}^{K}/_{2}$ if K is even, but in these cases $X_{\mathcal{F}}[l,k] = X_{\mathcal{S}}[l,k]$.

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. Here we use the MTF model instead of the CTF for simplicity, as these derivations would work exactly the same for the CTF.

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,

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

Comparing Eq. (A5) and Eq. (A6), it is trivial to see that $Y'_{\mathcal{F}}[l,k] \neq Y_{\mathcal{F}}[l,k]$. 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.A.P.; Benesty, J.; Gannot, S.; Naylor, P.A.; Cohen, I. On the application of the LCMV beamformer to speech enhancement. In Proceedings of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, October 2009; pp. 141–144. https://doi.org/10.1109/ASPAA.2009.5346463.
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.
22. Johnson, D.H. Signal Processing Information Database, 2013.