MDPI

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 **??** we develop a true-MVDR beamformer with the SSBT, taking into account its features.

2. Frequency and Time-Frequency Transforms

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. We from now on assume that all time-domain signals are real-valued.

For continuous time and frequency domains, the Fourier Transform (FT) is defined as

$$\begin{aligned} X_{\mathcal{F}}(f) &\equiv \mathcal{F}\{x(t)\}(f) \\ &= \int_{-\infty}^{\infty} x(t) e^{-\mathrm{j}2\pi f t} \mathrm{d}t \end{aligned} \tag{1}$$

We define the Real Fourier Transform (RFT) similarly, being cleverly constructed such that its frequency spectrum is real-valued without loss of information, as

$$\begin{aligned} X_{\mathcal{R}}(f) &\equiv \mathcal{R}\{x(t)\}(f) \\ &= \sqrt{2}\mathbb{R}\left\{ \int_{-\infty}^{\infty} x(t) e^{-\mathrm{j}2\pi f t + \mathrm{j}\frac{3\pi}{4}} \mathrm{d}t \right\} \\ &= \int_{-\infty}^{\infty} x(t)[-\cos(2\pi f t) - \sin(2\pi f t)] \mathrm{d}t \end{aligned} \tag{2}$$

and the Inverse Real Fourier Transform (iRFT) as

$$\begin{aligned} x(t) &\equiv \mathcal{R}^{-1}\{X_{\mathcal{R}}(f)\}(t) \\ &= \sqrt{2}\mathbb{R}\left\{\int_{-\infty}^{\infty} X_{\mathcal{R}}(f)e^{\mathrm{j}2\pi ft - \mathrm{j}\frac{3\pi}{4}}\,\mathrm{d}f\right\} \end{aligned} \tag{3}$$

It is easy to see that we can also define the RFT in terms of the FT through a simple substitution of Eq. (1) in Eq. (2), such that

$$\begin{aligned} X_{\mathcal{R}}(f) &= \mathbb{R}\left\{X_{\mathcal{F}}(f)e^{\mathrm{j}\frac{3\pi}{4}}\right\} \\ &= -X^{\mathbb{R}}_{\mathcal{F}}(f) + X^{\mathbb{I}}_{\mathcal{F}}(f) \end{aligned} \tag{4}$$

One can also write the RFT in terms of the FT as

$$X_{\mathcal{R}}(f) = \frac{1}{\sqrt{2}}\left(e^{-\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{F}}(f) + e^{\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{F}}(-f)\right) \tag{5}$$

from which we deduce that

$$X_{\mathcal{F}}(f) = \frac{1}{\sqrt{2}}\left(e^{\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{R}}(f) + e^{-\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{R}}(-f)\right) \tag{6}$$

2.1. Convolution

Given an impulse response $h(t)$ for a LIT system, through the FT it is known that

$$h(t) * x(t) \overset{\mathcal{F}}{\rightleftharpoons} H_{\mathcal{F}}(f)X_{\mathcal{F}}(f) \tag{7}$$

where $\overset{\mathcal{F}}{\rightleftharpoons}$ indicates a Fourier transform pair. It can be shown that this property isn't valid for the RFT (see Appendix A). That is, if $H_{\mathcal{R}}(f)$ and $X_{\mathcal{R}}(f)$ are the RFT's of $h(t)$ and $x(t)$ respectively, then

$$h(t) * x(t) \overset{\mathcal{R}}{\nRightarrow} H_{\mathcal{R}}(f)X_{\mathcal{R}}(f) \tag{8}$$

However, through Eq. (4) we can show that

$$\begin{aligned} h(t) * x(t) \overset{\mathcal{R}}{\rightleftharpoons} &\tfrac{1}{2}X_{\mathcal{R}}(f)[-H_{\mathcal{R}}(-f) - H_{\mathcal{R}}(f)] \\ &+\tfrac{1}{2}X_{\mathcal{R}}(-f)[H_{\mathcal{R}}(-f) - H_{\mathcal{R}}(f)] \end{aligned} \tag{9}$$

in which we see that, for a given frequency f, the convolution's output on the RFT domain depends on both it and its dual frequency $-f$. Using Eq. (4), which is also valid for $H_{\mathcal{R}}(f)$, we easily have that

$$h(t) * x(t) \overset{\mathcal{R}}{\rightleftharpoons} X_{\mathcal{R}}(f)H^{\mathbb{R}}_{\mathcal{F}}(f) - X_{\mathcal{R}}(-f)H^{\mathbb{I}}_{\mathcal{F}}(f) \tag{10}$$

2.2. Discrete time-frequency transforms

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

$$X_{\mathcal{S}}[l,k] = \sum_{n=0}^{K-1} w[n]x[n + l\cdot O]e^{-\mathrm{j}2\pi k\frac{(n+l\cdot O)}{K}} \tag{11}$$

where $w[n]$ is an analysis window of length K; and O is the overlap between windows of the transform, usually $O = \lfloor K/2 \rfloor$. The STFT can be seen as a discretization of the FT, while also applying it over different "snippets" of time.

The Single-Sideband Transform (SSBT) [11] is similarly defined, being the RFT's windowed discrete-time adaptation. The SSB transform of $x[n]$ is defined as

$$X_{\mathcal{S}}[l,k] = \sqrt{2}\mathbb{R}\left\{\sum_{n=0}^{K-1} w[n]x[n+l\cdot O]e^{-\mathrm{j}2\pi k\frac{(n+l\cdot O)}{K}+\mathrm{j}\frac{3\pi}{4}}\right\} \tag{12}$$

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, like with Eqs. (4) to (6) we have[1]

$$\begin{aligned} X_{\mathcal{S}}[l,k] &= \sqrt{2}\mathbb{R}\left\{X_{\mathcal{F}}[l,k]e^{\mathrm{j}\frac{3\pi}{4}}\right\} \\ &= -\mathbb{R}\{X_{\mathcal{F}}[l,k]\} + \mathbb{I}\{X_{\mathcal{F}}[l,k]\} \end{aligned} \tag{13}$$

$$X_{\mathcal{S}}[l,k] = \frac{1}{\sqrt{2}}\left(e^{-\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{F}}[l,k] + e^{\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{F}}[l,K-k]\right) \tag{14}$$

$$X_{\mathcal{F}}[l,k] = \frac{1}{\sqrt{2}}\left(e^{\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{S}}[l,k] + e^{-\mathrm{j}\frac{3\pi}{4}}X_{\mathcal{S}}[l,K-k]\right) \tag{15}$$

As was the case for the RFT, the SSBT also doesn't hold the convolution theorem the same way as the STFT does. However, similarly to what was shown in Eq. (9), we can write the convolution on the SSBT domain as

$$h[n] * x[n] \overset{\mathcal{S}}{\rightleftharpoons} X_{\mathcal{S}}[l,k]H^{\mathbb{R}}_{\mathcal{F}}[k] - X_{\mathcal{S}}[l,K-k]H^{\mathbb{I}}_{\mathcal{F}}[k] \tag{16}$$

or, with the convolutive transfer function (CTF) model [13],

$$h[n] * x[n] \overset{\mathcal{S}}{\rightleftharpoons} X_{\mathcal{S}}[l,k] * H^{\mathbb{R}}_{\mathcal{F}}[l,k] - X_{\mathcal{S}}[l,K-k] * H^{\mathbb{I}}_{\mathcal{F}}[l,k] \tag{17}$$

in which this convolution is done over the frames l.

2.3. Relative transfer functions

Given two systems that share an input $x[n]$ each with an impulse response $h_1[n]$ and $h_2[n]$, on the STFT domain $H_1[l,k]$ and $H_2[l,k]$, we can calculate their relative transfer functions (RTF's), respective to a common input. We denote these RTF's $A_1[l,k]$ and $A_2[l,k]$, respective for each system.

Let us denote $Y_1[l,k] = H_1[k]X[l,k]$, under the MTF model, and similarly for $Y_2[l,k]$. With the STFT, we write $X_1[l,k] = H_1[k]X[l,k]$, and thus $Y_1[l,k] = A_1[k]X_1[l,k]$ with $A_1[k] = 1$. We can obtain $A_2[k]$ as

$$A_2[k] = \frac{H_2[k]}{H_1[k]} \tag{18}$$

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

which trivially satisfies that $A_2[k]X_1[l,k] = H_2[k]X[l,k]$. These RTF's can be calculated as

$$A_m[k] = \frac{\mathbb{E}\{X_m[l,k]X_1^*[l,k]\}}{\mathbb{E}\{X_1[l,k]X_1^*[l,k]\}} \tag{19}$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator.

This isn't as straight-forward with the SSBT, since after the convolution each frequency depends on its conjugate as well. However, by considering each system to have two inputs $X'[l,k] = X[l,k]$ and $X''[l,k] = X[l,K-k]$ and two transfer functions $H'_m[k]$ and $H''_m[k]$ (where m represents the system's index), then our outputs can be described as

$$Y_1[l,k] = H'_1[k]X'[l,k] + H''_1[k]X''[l,k] \tag{20a}$$
$$Y_2[l,k] = H'_2[k]X'[l,k] + H''_2[k]X''[l,k] \tag{20b}$$

From Eq. (16), we easily see that

$$H'_m[k] = H'_m[K-k] = H^{\mathbb{R}}_{\mathcal{F};m}[k] \tag{21a}$$
$$H''_m[k] = -H''_m[K-k] = -H^{\mathbb{I}}_{\mathcal{F};m}[k] \tag{21b}$$

We let

$$X'_1[l,k] = H'_1[k]X'[l,k] \tag{22a}$$
$$X''_1[l,k] = H''_1[k]X''[l,k] \tag{22b}$$

and

$$\begin{aligned} X_1[l,k] &= H'_1[k]X'[l,k] + H''_1[k]X''[l,k] \\ &= X'_1[l,k] + X''_1[l,k] \end{aligned} \tag{23}$$

in which $X'_1[l,k]$ and $X''_1[l,k]$ are the inputs processed by the first system, and $X_1[l,k]$ is the observable input signal. Through these, we get that

$$Y_1[l,k] = A'_1[k]X'_1[l,k] + A''_1[k]X''_1[l,k] \tag{24a}$$
$$Y_2[l,k] = A'_2[k]X'_1[l,k] + A''_2[k]X''_1[l,k] \tag{24b}$$

where

$$A^n_m[k] = \frac{H^n_m[k]}{H^n_1[k]} \tag{25}$$

are the RTF's for the n-th input, between the m-th system and the reference (assumed to be $m = 1$). Trivially, $A'_1[k] = A''_1[k] = 1$. Note that, for this, we must be able to estimate each $G^n_m[k]$ separately, which may not be easy. For example, the technique in Eq. (19) used for the STFT isn't applicable here. Using it directly on the observable input $X_m[l,k]$ would yield

$$\frac{\mathbb{E}\{X_m[l,k]X_1^*[l,k]\}}{\mathbb{E}\{X_1[l,k]X_1^*[l,k]\}} = \frac{H'_m H'_1 + H''_m H''_1}{{H'_1}^2 + {H''_1}^2} \tag{26}$$

where we used that different frequency bins are independent and all are zero-mean. The exact same would be obtained for $K-k$, thus this tool isn't useful to give us any insight into the desired RTF's of the form from Eq. (25), opposite to with the STFT, where this is possible. Also, since we don't have access to each input $X'_1[l,k]$ and $X''_1[l,k]$ separately, we can't do this practically on each of them, following Eq. (19).

This same formulation can be used with the CTF model. That is, if we consider $A_m^n[l,k]$ as a convolutive gain, then

$$A_m^n[l,k] \approx \frac{H_m^n[l,k]}{H_1^n[0,k]} \tag{27}$$

this being valid for both the STFT (where $n = 1$ is the only option), and for the SSBT. This is an approximation, since different windows in a time-frequency transform aren't independent and thus can contribute to the gain of one-another.

We will from now on assume that we have the RTF's according to Eqs. (24) and (25), allowing the mathematical continuation of the problem.

3. Signal Model and Beamforming

Let there be a device that consists of M sensors and a loudspeaker (LS) in a reverberant environment, in which there also is a desired source, both traveling with a speed c. We also assume the presence of undesired noise at each sensor. For simplicity we assume that all sources are spatially stationary, although this condition can be easily removed.

We denote $Y_{m,k}[l]$ as the signal at the m-th sensor on the time-frequency domain, being represented by

$$Y_{m,k}[l] = X_{m,k}[l] + S_{m,k}[l] + R_{m,k}[l] \tag{28}$$

where $X_{m,k}[l]$ is a desired signal component, $S_{m,k}[l]$ is the undesired loudspeaker signal that is captured by the sensors, and $R_{m,k}[l]$ is uncorrelated white noise present in the sensors. m is the sensor index ($1 \leq m \leq M$), k is the frequency bin index ($0 \leq k < K$), and l is the decimated-time index. We will use a different notation to the one previously used, for ease of reading.

Treating the path and reverberations it takes between the desired source and the m-th sensor as a system, and with the CTF model, in the STFT domain $X_{m,k}[l]$ can be represented as

$$X_{m,k}[l] = A_{m,k}[l] * X'_{1,k}[l] \tag{29}$$

where $X'_{1,k}[l] = H_{1,k}[l] * X_k[l]$ is the desired signal at the reference sensor ($m = 1$), $A_{m,k}[l]$ is the RTF for each sensor, and $H_{1,k}[l]$ is the desired signal's transfer function between the source and the reference sensor. Meanwhile, in the SSBT domain we can represent it as

$$X_{m,k}[l] = A'_{m,k}[l] * X'_{1,k}[l] + A''_{m,k}[l] * X''_{1,k}[l] \tag{30}$$

$$X'_{1,k}[l] = H'_{1,k}[l] * X_k[l] \tag{31a}$$

$$X''_{1,k}[l] = H''_{1,k}[l] * X_{K-k}[l] \tag{31b}$$

where we now have two desired inputs for each sensor, and two RTF's as well. Using Eqs. (21) to (23), we have that $X'_{1,k}[l] \neq X''_{1,K-k}[l]$, even though both originate from $X_k[l]$. Likewise, $X''_{1,k}[l] \neq X'_{1,K-k}[l]$, for the same reason.

Also, since the STFT formulation can also be represented by the SSBT one, with $A'_{m,k}[l] = A_{m,k}[l]$ and $A''_{m,k}[l] = 0$. Therefore, the mathematical developments will be done the SSBT formulation in mind, as it is a less restricting model.

Note that $A_{m,k}[l]$ isn't strictly a causal response, depending on the direction of arrival and features of the reverberant environment, as well as relative delays between the sources at each sensor. We will assume that there are Δ non-causal samples in $A_{m,k}[l]$. It is trivial to see that, for Eq. (30) to be respected, $A'_{1,k}[l] = A''_{1,k}[l] = \delta_{0,l}$, a Kronecker delta at $l = 0$.

We consider the delayed signal $X_{m,k}[l+\lambda]$, such that, when expanding the convolutions, we have

$$X_{m,k}[l+\lambda] = \sum_{\tau}\Big(A'_{m,k}[\tau]X'_{1,k}[l+\lambda-\tau] + A''_{m,k}[\tau]X''_{1,k}[l+\lambda-\tau]\Big) \tag{32}$$

Now we explicit the contribution of $X^n_{1,k}[l]$ on the summation, leading us to

$$\begin{aligned} X_{m,k}[l+\lambda] &= A'_{m,k}[\lambda]X'_{1,k}[l] + A''_{m,k}[\lambda]X''_{1,k}[l] \\ &\quad + \sum_{\tau\neq\lambda}\Big(A'_{m,k}[\tau]X'_{1,k}[l+\lambda-\tau] + A''_{m,k}[\tau]X''_{1,k}[l+\lambda-\tau]\Big) \\ &= A'_{m,k}[\lambda]X'_{1,k}[l] + A''_{m,k}[\lambda]X''_{1,k}[l] + Q_{m,k}[l+\lambda] \end{aligned} \tag{33}$$

where the first two terms are the contributions of the desired signals at the time of interest l, and $Q_{m,k}[l+\lambda]$ are the remaining terms of the convolution, which can be regarded as only reverberation. Using this on Eq. (28) with Eq. (30), we have that

$$Y_{m,k}[l+\lambda] = A'_{m,k}[\lambda]X'_{1,k}[l] + A''_{m,k}[\lambda]X''_{1,k}[l] + Q_{m,k}[l+\lambda] + S_{m,k}[l+\lambda] + R_{m,k}[l+\lambda] \tag{34}$$

We consider L_Y previous and $\bar{L}_Y$ future samples of $Y_{m,k}[l]$ that are influenced by $X^n_{1,k}[l]$ (that is, $A'_{m,k}[\lambda] \neq 0$ or $A''_{m,k}[\lambda] \neq 0$), and define $L = L_Y + \bar{L}_Y + 1$. With this, we get

$$\mathbf{Y}_{m,k}[l] = \mathbf{A}'_{m,k}X'_{1,k}[l] + \mathbf{A}''_{m,k}X''_{1,k}[l] + \mathbf{Q}_{m,k}[l] + \mathbf{S}_{m,k}[l] + \mathbf{R}_{m,k}[l] \tag{35}$$

in which

$$\mathbf{Y}_{m,k}[l] = \Big[\; Y_{m,k}[l+\bar{L}_Y]\,,\; \cdots,\; Y_{m,k}[l]\,,\; \cdots,\; Y_{m,k}[l-L_Y] \;\Big]^{\mathsf{T}} \tag{36}$$

and similarly for all other signal vectors, and

$$\mathbf{A}^n_{m,k} = \Big[\; A_{m,k}[\bar{L}_Y]\,,\; \cdots,\; A_{m,k}[0]\,,\; \cdots,\; A_{m,k}[-L_Y] \;\Big]^{\mathsf{T}} \tag{37}$$

Now stacking all sensors into a vector, we get

$$\mathbf{y}_k[l] = \mathbf{a}'_k X'_{1,k}[l] + \mathbf{a}''_k X''_{1,k}[l] + \mathbf{q}_k[l] + \mathbf{s}_k[l] + \mathbf{r}_k[l] \tag{38}$$

with

$$\mathbf{y}_k[l] = \Big[\; \mathbf{Y}^{\mathsf{T}}_{1,k}[l]\,,\; \cdots,\; \mathbf{Y}^{\mathsf{T}}_{M,k}[l] \;\Big]^{\mathsf{T}} \tag{39}$$

and the same for $\mathbf{a}^n_k$, $\mathbf{q}_k[l]$, $\mathbf{s}_k[l]$ and $\mathbf{r}_k[l]$, with them all being $ML \times 1$ vectors.

3.1. Filtering and the MPDR beamformer

We want to recover the desired signal at the reference sensor, $X_{1,k}[l] = X'_{1,k}[l] + X''_{1,k}[l]$ (see Eq. (30) with $m = 1$), without any distortion. For this, a linear filter $\mathbf{f}_k[l]$ will be employed, producing an estimate $Z_k[l]$ of our desired signal, such that

$$\begin{aligned} Z^n_k[l] &\approx X_{1,k}[l] \\ &= \mathbf{f}^{\mathsf{H}}_k[l]\mathbf{y}_k[l] \end{aligned} \tag{40}$$

with $(\cdot)^{\mathsf{H}}$ being the transposed-complex-conjugate operator. This process can also be interpreted as

$$\begin{aligned} Z_k[l] &= \sum_m \bar{\mathbf{f}}^{\mathsf{H}}_{m,k}[l]\bar{\mathbf{y}}_{m,k}[l] \\ &= \sum_m F^*_{m,k}[l] * Y_{m,k}[l] \end{aligned} \tag{41}$$

where $\bar{\mathbf{f}}_{m,k}[l]$ is the $L_Y \times 1$ part of $\mathbf{f}_k[l]$ that filters the m-th sensor, and $F_{m,k}[l]$ is its signal-form counterpart. In this sense, the filtering process can be interpreted as the sum across all sensors of the convolution between the signal and the observations.

Going back to Eq. (40), with Eq. (38) we can write

$$Z_k[l] = \mathbf{f}^{\mathsf{H}}_k[l]\mathbf{a}'_k X'_{1,k}[l] + \mathbf{f}^{\mathsf{H}}_k[l]\mathbf{a}''_k X''_{1,k}[l] + \mathbf{f}^{\mathsf{H}}_k[l]\mathbf{q}_k[l] + \mathbf{f}^{\mathsf{H}}_k[l]\mathbf{s}_k[l] + \mathbf{f}^{\mathsf{H}}_k[l]\mathbf{r}_k[l] \tag{42}$$

From this, we easily see that to achieve a distortionless response from the desired signal, we must have that

$$\mathbf{f}^{\mathsf{H}}_k[l]\mathbf{a}'_k = 1 \tag{43a}$$

$$\mathbf{f}^{\mathsf{H}}_k[l]\mathbf{a}''_k = 1 \tag{43b}$$

which will ensure that both components of the desired signal are undistorted. For the STFT, only the first constraint is considered, since in this case we have that $\mathbf{a}''_k = \mathbf{0}$ and thus the second condition is impossible. With this, we write our constraint matrix as

$$\mathbf{f}^{\mathsf{H}}_k[l]\mathbf{C}_k = \mathbf{i}^{\mathsf{T}} \tag{44}$$

where, for the STFT, $\mathbf{C}_k = \mathbf{a}_k$ and $\mathbf{i} = 1$; and, for the SSBT, we have

$$\mathbf{C}_k = \begin{bmatrix} \mathbf{a}'_k\,, & \mathbf{a}''_k \end{bmatrix}_{ML\times 2} \tag{45a}$$

$$\mathbf{i} = \begin{bmatrix} 1\,, & 1 \end{bmatrix} \tag{45b}$$

To minimize the variance of the output signal while obeying the distortionless constraint, a Minimum-Power Distortionless Response (MPDR) beamformer will be used, it being defined as

$$\mathbf{f}_{\text{mpdr};k}[l] = \min_{\mathbf{f}_k[l]} \mathbf{f}^{\mathsf{H}}_k[l]\mathbf{\Phi}_{\mathbf{y}_k}[l]\mathbf{f}_k[l] \text{ s.t. } \mathbf{f}^{\mathsf{H}}_k[l]\mathbf{C}_k = \mathbf{i}^{\mathsf{T}} \tag{46}$$

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

$$\mathbf{f}_{\text{mpdr};k}[l] = \mathbf{\Phi}^{-1}_{\mathbf{y}_k}[l]\mathbf{C}_k\left[\mathbf{C}^{\mathsf{H}}_k\mathbf{\Phi}^{-1}_{\mathbf{y}_k}[l]\mathbf{A}_k\right]^{-1}\mathbf{i} \tag{47}$$

4. Comparisons and simulations

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 × 6m × 3m (width × length × height), with a reverberation time of 0.3s. The device composed of the loudspeaker + sensors is centered at

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

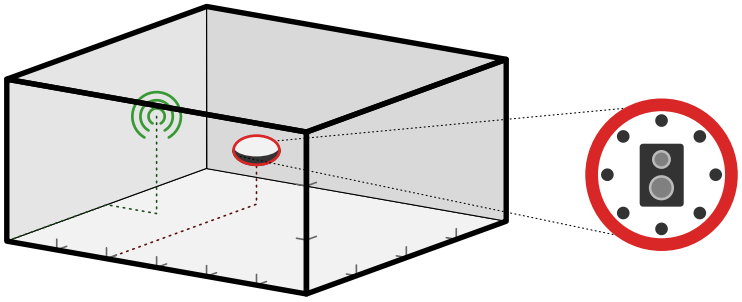


Figure 1. Room layout for simulations.

(3m, 4m, 1m), being comprised of $M = 8$ sensors. They are arranged in a circular array with radius of 8cm, and all are omnidirectional of flat frequency response. The positions and signals used for the sources are in Table 1. The room's layout is in Fig. 1, where in green we have the desired source (assumed to be omnidirectional), and in red the device, with the 8 sensors and the loudspeaker on the center.

Source	Position	Signal
$x[n]$	(2m, 1m, 1.8m)	`50_male_speech_english_ch8_OmniPower4296.flac`
$s[n]$	(3m, 2m, 1m)	`69_abba_ch8_OmniPower4296.flac`
$r[n]$	$\sim$	`wgn_48kHz_ch8_OmniPower4296.flac`

Table 1. Source information for the simulations.

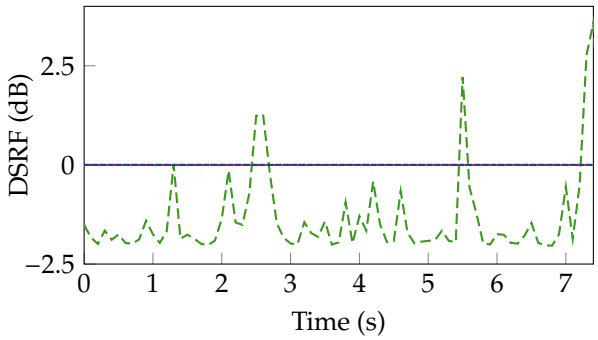
All signals were resampled to the desired sampling frequency of 16kHz. For the transforms, Hamming windows were used, with a length of 32 samples/window and an overlap of 50%. The beamformers were calculated once for the whole signal, for faster processing and ease to compare the results. We will compare one beamformer for the STFT with and two for the SSBT. The STFT one will be based on Eq. (47), 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 **??**, which led to **??**. These names were chosen given that the one proposed in **??** uses two frequencies (namely the "dual-frequencies" from the STFT) at each moment, while the SF-SSBT beamformer only calculates one frequency at a time.

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.

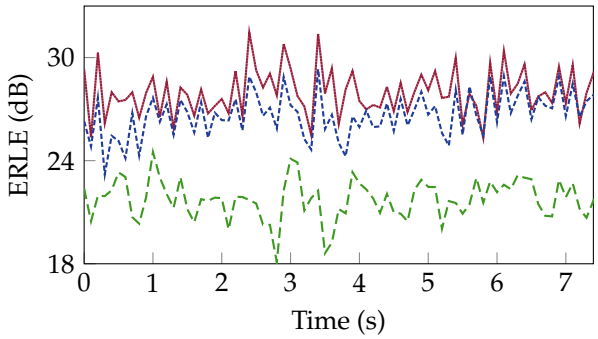
4.1. Basic comparison

At the reference sensor (assumed to be the one at (3m, 1.92m, 1m)), the SNR for the loudspeaker's and noise signals are respectively −15dB and 30dB. These will be referred as Signal-to-Echo and Signal-to-Noise Ratios (SER and SNR) in that order. We will use $L_Y = 1$ in this first scenario; other cases will be studied later.

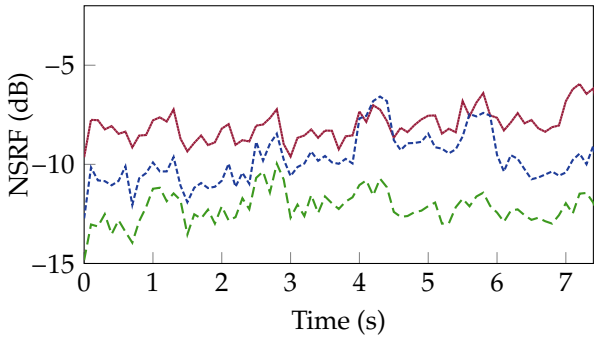
In these simulations, we are interested in three metrics results: maintenance/no-distortion of the desired signal; decrease in the loudspeaker's signal; and reduction/minimal enhancement of the white noise. In order, these will be measured by the desired signal reduction factor (DSRF), echo-return loss enhancement (ERLE), and noise signal reduction factor (NSRF), the later being used for the white noise given that the only other undesired



(a) Per-window broadband DSRF.



(b) Per-window broadband ERLE.



(c) Per-window broadband NSRF.

Figure 2. Output metrics for the beamformers over time, in the base scenario.

signal at the sensors is white uncorrelated. Their time-dependent broadband formulations are

$$\xi_x[l] = \frac{\sum_k |X_1[l,k]|^2}{\sum_k \left|X_f[l,k]\right|^2} \tag{48a}$$

$$\xi_s[l] = \frac{\sum_k |S_1[l,k]|^2}{\sum_k \left|S_f[l,k]\right|^2} \tag{48b}$$

$$\xi_r[l] = \frac{\sum_k |V_1[l,k]|^2}{\sum_k \left|V_f[l,k]\right|^2} \tag{48c}$$

where $S_f[l,k] = \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{s}_1[l,k]$, $X_f[l,k] = \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{x}_1[l,k]$ and $R_f[l,k] = \mathbf{f}^{\mathsf{H}}[l,k]\mathbf{r}_1[l,k]$ as the filtered-LS, filtered-desired and filtered-noise signals, respectively.

From Fig. 2a, we see that the STFT and DF-SSBT beamformers had a null distortion of the desired signal, while the SF-SSBT had errors of (on average) 2dB.

From the ERLE results in Fig. 2b it is noticeable that the STFT and SF-SSBT beamformers had a similar performance, with the STFT one being marginally better, both outperforming the SF-SSBT beamformer's results by about 3dB. In terms of the NSRF, we see that the STFT beamformer had a much better performance than the two SSBT-based beamformers, therefore increasing the white-noise less on the output.

4.2. Comparison over different input SERs

We now examine 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 presented below. The other parameters and variables are maintained from the previous simulations, with only the SER being changed.

$$\xi_x = \frac{\sum_{l,k}|X_1[l,k]|^2}{\sum_{l,k}\left|X_f[l,k]\right|^2} \tag{49a}$$

$$\xi_s = \frac{\sum_{l,k}|S_1[l,k]|^2}{\sum_{l,k}\left|S_f[l,k]\right|^2} \tag{49b}$$

$$\xi_r = \frac{\sum_{l,k}|V_1[l,k]|^2}{\sum_{l,k}\left|V_f[l,k]\right|^2} \tag{49c}$$

As seen in Fig. 3a, the STFT and DF-SSBT beamformers caused zero distortion, and the SF-SSBT beamformer led to some. We can also see that this distortion decreases as the loudspeaker SNR increases.

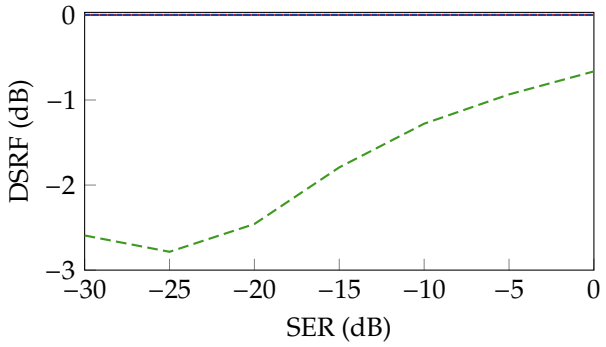
In terms of ERLE, we see that the SF-SSBT is strictly worse than the two other beamformers, for all SER's. The DF-SSBT beamformer has a similar performance to that of the STFT beamformer for iSER $\lessapprox -20$dB, but is outperformed for higher iSER's.

For the NSRF we see the same results as were obtained previously, with the white-noise increase for the STFT beamformer being considerably lower than that for both SSBT-based beamformers. It is interesting that the performance of both SSBT beamformers worsens for higher iSER's, for both the ERLE and NSRF metrics.

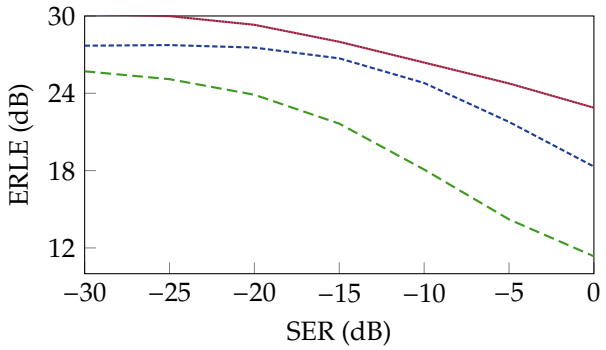
4.3. Comparison for different L_Y

Going back to observing only the case for iSER = -15dB, we will now investigate the effects of varying L_Y on the beamformers' performances.

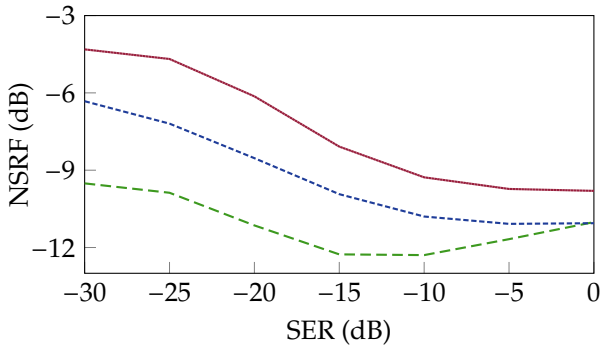
In this comparison, we have two different relevant effects that can be seen: firstly, we see that the SF-SSBT beamformer's performance deteriorates drastically, for both the desired signal's distortionless-ness behavior, and the ERLE. This is likely due to the mathematical results exposed in Appendix A, and since the SSBT transform doesn't respect the convolution, a convolutive filter doesn't work well with it. Note that this is not the case for the DF-SSBT beamformer, since these were appropriately designed to bypass this complication. The increase in the SF-SSBT's NSRF performance is due to it no not working correctly, and thus "randomly".



(a) Time-average broadband DSRF.



(b) Time-average broadband ERLE.



(c) Time-average broadband NSRF.

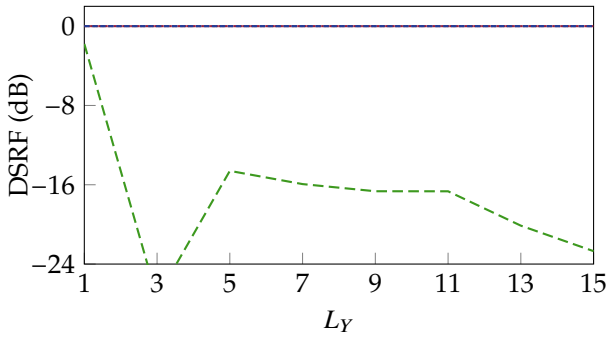
Figure 3. Output metrics for the beamformers for varying input SERs.

Another relevant result from this comparison is that increasing L_Y doesn't seem to have much of an effect on neither the ERLE nor the NSRF, for both the STFT and DF-SSBT beamformers. We also see, in accordance with the previous results, that the STFT beamformer strictly outperforms the DF-SSBT one for both metrics.

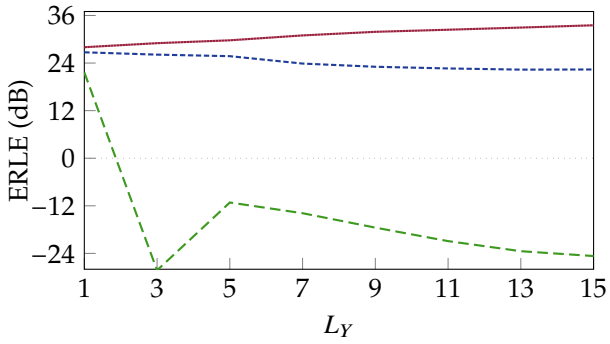
4.4. Comparison with perturbation

As exposed in **??**, it is also of interest to compare how robust the derived beamformers are, when the information regarding the desired signal's RIR isn't accurate. For such, we model the matrix $\underline{\mathbf{A}}_k$ as

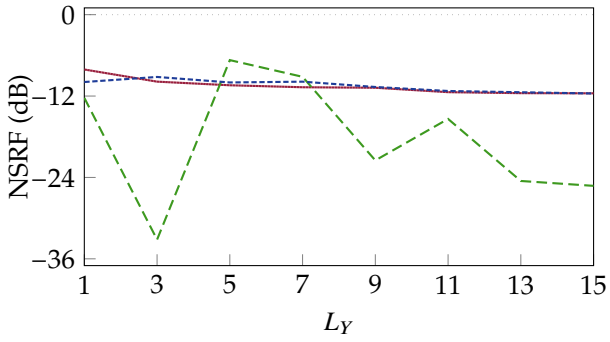
$$\underline{\mathbf{A}}_k = \underline{\mathbf{A}}_k^{\star} + \Delta\underline{\mathbf{A}}_k \tag{50}$$



(a) Per-window broadband DSRF.



(b) Per-window broadband ERLE.



(c) Per-window broadband NSRF.

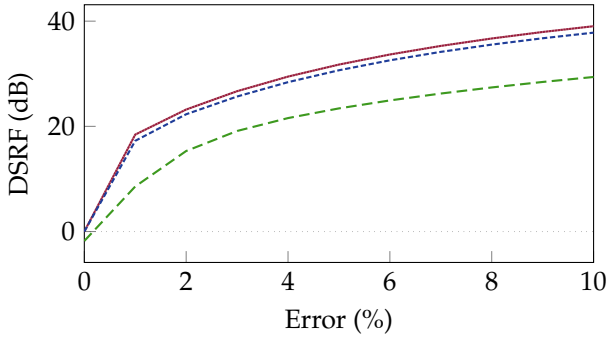
Figure 4. Output metrics for the beamformers for varying input L_Y's.

where $\underline{\mathbf{A}}_k^\star$ is the accurate steering matrix, and $\Delta\underline{\mathbf{A}}_k$ is a perturbation on it, which we assume is a zero-mean uniform white noise, with an adjustable variance.

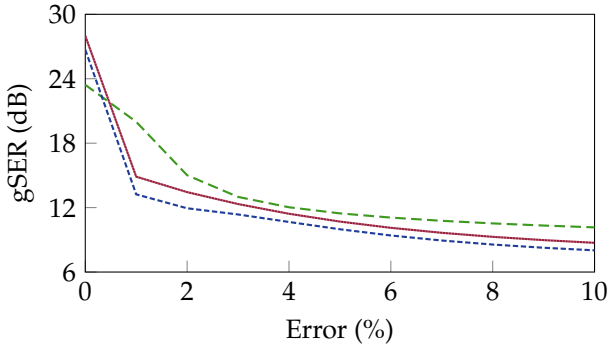
Since in this scenario the desired signal can suffer some distortion (given that its steering matrix isn't appropriately estimated), we will use the gain in SER and gain in SNR metrics instead of ERLE and NSRF, to take this distortion into account. These are defined as

$$\text{gSER} = \frac{\xi_s}{\xi_x} \tag{51a}$$

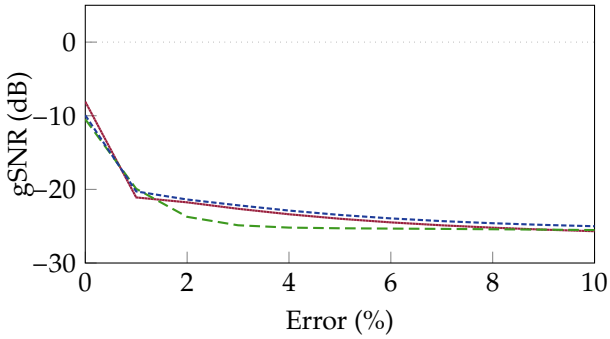
$$\text{gSNR} = \frac{\xi_r}{\xi_x} \tag{51b}$$



(a) Per-window broadband DSRF.



(b) Per-window broadband gSER.



(c) Per-window broadband gSNR.

Figure 5. Output metrics for the beamformers with error in the steering vectors.

The DSRF will still be showed, to give a sense of proportion on how much the beamformer distorts the desired signal. In the results of Fig. 5, the x-axis represents the standard deviation of $\Delta\underline{\mathbf{A}}_k$, as a percentage of the standard deviation of $\underline{\mathbf{A}}_k^\star$.

Each metric showed a different result: differently than before, the SF-SSBT beamformer led to the least distortion on the desired signal, out of all three beamformers, and the DF-SSBT led to the most distortion. Meanwhile, the gain in SER showed the STFT beamformer to be the (overall) more robust, and the one that led to the best results, with the DF-SSBT again being the worse one. A similar result can be seen for the gain in SNR, with the STFT beamformer being the best, but in this regard the SF-SSBT beamformer led to the worst results, although marginally.

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 $X_{\S}[l,k]$ its SSBT equivalent. We here assume that both the STFT and the SSBT have K frequency bins. $X_{\S}[l,k]$ can be obtained using $X_{\mathcal{F}}[l,k]$, through

$$X_{\mathcal{R}}[l,k] = -X_{\mathcal{F}}{}^{\mathbb{R}}[l,k] - X_{\mathcal{F}}{}^{\mathbb{I}}[l,k] \tag{A.1}$$

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

It is easy to see that

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

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

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

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

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

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

Passing this through Eq. (A.2),

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

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

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

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

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.

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.