


On the Single-Sideband Transform for MVDR Beamformers

Vitor Probst Curtarelli^{1,*} , Israel Cohen¹ 

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 applications, 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:

Copyright: © 2024 by the authors. Submitted to *Algorithms* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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

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

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

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

2. STFT and the Single-Sideband Transform

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

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

$$X_{\mathcal{F}}[l, k] = \sum_{n=0}^{K-1} w[n] x[n + l \cdot O] e^{-j2\pi k \frac{(n+l \cdot O)}{K}} \quad (1)$$

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

$$X_S[l, k] = \sqrt{2} \Re \left\{ \sum_{n=0}^{K-1} w[n] x[n + l \cdot O] e^{-j2\pi k \frac{(n+l \cdot O)}{K} + j\frac{3\pi}{4}} \right\} \quad (2)$$

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

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

It is easy to see that¹

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

from which it we deduce

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

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

3. Signal Model and Beamforming

Let there be a device that consists of M sensors and a loudspeaker (LS) in a reverberant environment, in 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 $y_m[n]$ as the signal at the m -th sensor, being defined as

$$y_m[n] = h_m[n] * x[n] + g_m[n] * s[n] + r_m[n] \quad (6)$$

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

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

$$Y_{m,k}[l] = H_{m,k}[l] * X_k[l] + G_{m,k}[l] * S_k[l] + R_{m,k}[l] \quad (7)$$

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

We let m' be the reference sensor's index, for simplicity assume $m' = 1$, and we denote $x_{1,k}[l] = H_{1,k}[l] * X_k[l]$. For each sensor m and each frequency k , we denote $A_{m,k}[l]$ as the relative impulse response for the desired signal (at the reference sensor) and the m -th sensor, such that

$$A_{m,k}[l] * X_{1,k}[l] = H_{m,k}[l] * X_k[l] \quad (8)$$

We similarly define $B_{m,k}[l]$ and $S_{1,k}[l]$ based on $G_{m,k}[l]$ and $S_k[l]$. Therefore, Eq. (7) becomes

$$Y_{m,k}[l] = A_{m,k}[l] * X_{1,k}[l] + B_{m,k}[l] * S_{1,k}[l] + R_{m,k}[l] \quad (9)$$

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

Here, the impulse responses $A_{m,k}[l]$ and $B_{m,k}[l]$ can be non-causal, depending on the direction of arrival and features of the reverberant environment, as well as relative delays between the sources at each sensor. They can also be non-causal on account of the windowing process of the time-frequency transforms. We assume that there are Δ non-causal samples in $A_{m,k}[l]$. It is trivial to see that $A_{1,k}[l] = \delta_{0,l}$, a Kronecker delta at $l = 0$.

We let $A'_{m,k}[l]$ comprise the L_A samples of $A_{m,k}[l]$ of most interest (for example, the L_A causal samples starting on the non-zero value of $A_{1,k}[l]$), and define $Q_{m,k}[l]$ such that

$$Q_{m,k}[l] = A_{m,k}[l] * X_{1,k}[l] - A'_{m,k}[l] * X_{1,k}[l] \quad (10)$$

We thus have

$$A'_{m,k}[l] * X_{1,k}[l] = \mathbf{a}_{m,k}^T \mathbf{x}_{1,k}[l] \quad (11)$$

in which

$$\mathbf{a}_{m,k} = \left[A_{m,k}[\Delta], \dots, A_{m,k}[0], \dots, A_{m,k}[L_A - \Delta - k] \right]^T \quad (12a)$$

$$\mathbf{x}_{1,k}[l] = \left[X_{1,k}[l + \Delta], \dots, X_{1,k}[l], \dots, X_{1,k}[l - L_A + \Delta + k] \right]^T \quad (12b)$$

and in the same way we define $\mathbf{b}_{m,k}$ and $\mathbf{s}_{1,k}[l]$. Note that $\mathbf{a}_{m,k}$ and $\mathbf{b}_{m,k}$ don't depend on the index l , given the spatial stationarity assumption. With this, Eq. (7) becomes

$$Y_{m,k}[l] = \mathbf{a}_{m,k}^T \mathbf{x}_{1,k}[l] + \mathbf{b}_{m,k}^T \mathbf{s}_{1,k}[l] + R_{m,k}[l] + Q_{m,k}[l] \quad (13)$$

We take L_Y samples of our observed signal, and define $\bar{\mathbf{y}}_{m,k}[l]$ as

$$\bar{\mathbf{y}}_{m,k}[l] = \left[Y_{m,k}[l], Y_{m,k}[l - 1], \dots, Y_{m,k}[l - L_Y + 1] \right]^T \quad (14)$$

In this new framework, we can write $\mathbf{y}_{m,k}[l]$ as a $L_Y \times 1$ vector

$$\bar{\mathbf{y}}_{m,k}[l] = \bar{\mathbf{A}}_{m,k} \bar{\mathbf{x}}_{1,k}[l] + \bar{\mathbf{B}}_{m,k} \bar{\mathbf{s}}_{1,k}[l] + \bar{\mathbf{r}}_{m,k}[l] + \bar{\mathbf{q}}_{m,k}[l] \quad (15)$$

where $\bar{\mathbf{r}}_{m,k}[l]$ and $\bar{\mathbf{q}}_{m,k}[l]$ are defined similarly to Eq. (14), $\bar{\mathbf{x}}_{1,k}[l]$ is a $L \times 1$ vector, and $\bar{\mathbf{A}}_{m,k}$ is a $L_Y \times L$ matrix, both being given by

$$\bar{\mathbf{x}}_{1,k}[l] = \left[X_{1,k}[l + \Delta], X_{1,k}[l + \Delta - 1], \dots, X_{1,k}[l + \Delta - L] \right]^T \quad (16a)$$

$$\bar{\mathbf{A}}_{m,k} = \begin{bmatrix} \mathbf{a}_{m,k}^T & 0 & \dots & 0 \\ 0 & \mathbf{a}_{m,k}^T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{a}_{m,k}^T \end{bmatrix} \quad (16b)$$

in which $L = L_Y + L_A - 1$, and $\mathbf{a}_{m,k} \equiv \mathbf{a}_m[k]$. $\bar{\mathbf{s}}_{1,k}[l]$ and $\bar{\mathbf{B}}_{m,k}$ are defined similarly, them being a $(L + L_C - 1) \times 1$ vector and a $L_Y \times (L + L_C - 1)$ matrix respectively. We now concatenate the matrices and vectors for the M different sensors, such that

$$\mathbf{y}_k[l] = \mathbf{A}_k \bar{\mathbf{x}}_{1,k}[l] + \mathbf{B}_k \bar{\mathbf{s}}_{1,k}[l] + \mathbf{r}_k[l] + \mathbf{q}_k[l] \quad (17)$$

where

$$\mathbf{y}_k[l] = \begin{bmatrix} \bar{\mathbf{y}}_{1,k}[l] \\ \vdots \\ \bar{\mathbf{y}}_{M,k}[l] \end{bmatrix} \quad (18a)$$

$$\mathbf{A}_k = \begin{bmatrix} \bar{\mathbf{A}}_{1,k} \\ \vdots \\ \bar{\mathbf{A}}_{M,k} \end{bmatrix} \quad (18b)$$

$\mathbf{y}_k[l]$ is a $ML_Y \times 1$ vector, and \mathbf{A}_k is a $ML_Y \times L$ matrix. $\mathbf{r}_k[l]$ and $\mathbf{q}_k[l]$ are defined in the same way as $\mathbf{y}_k[l]$, and \mathbf{B}_k as \mathbf{A}_k .

3.1. Filtering and the MPDR beamformer

We denote $Z_k[l]$ as an estimate the desired signal at reference $X_{1,k}[l]$, via a filter $\mathbf{f}_k[l]$ of length ML_Y , such that

$$\begin{aligned} Z_k[l] &\approx X_{1,k}[l] \\ &= \mathbf{f}_k^H[l] \mathbf{y}_k[l] \end{aligned} \quad (19)$$

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

$$\begin{aligned} Z_k[l] &= \sum_m \bar{\mathbf{f}}_{m,k}^H[l] \bar{\mathbf{y}}_{m,k}[l] \\ &= \sum_m F_{m,k}^*[l] * Y_{m,k}[l] \end{aligned} \quad (20)$$

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. (19), with Eq. (17) we can write

$$Z_k[l] = \mathbf{f}_k^H[l] \mathbf{A}_k \bar{\mathbf{x}}_{1,k}[l] + \mathbf{f}_k^H[l] \mathbf{B}_k \bar{\mathbf{s}}_{1,k}[l] + \mathbf{f}_k^H[l] \mathbf{r}_k[l] + \mathbf{f}_k^H[l] \mathbf{q}_k[l] \quad (21)$$

From this, we easily see that to achieve a distortionless response from the desired signal, we must have that $\mathbf{f}_k^H[l] \mathbf{A}_k \bar{\mathbf{x}}_{1,k}[l] = X_{1,k}[l]$, and therefore the distortionless constraint is given by

$$\mathbf{f}_k^H[l] \mathbf{A}_k = \mathbf{i}_\Delta^T \quad (22)$$

where \mathbf{i}_Δ is a $L \times 1$ vector of zeroes, except for the Δ -th entry which is a 1.

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

$$\mathbf{f}_{\text{mpdr};k}[l] = \min_{\mathbf{f}_k[l]} \mathbf{f}_k^H[l] \Phi_{\mathbf{y}_k}[l] \mathbf{f}_k[l] \text{ s.t. } \mathbf{f}_k^H[l] \mathbf{A}_k = \mathbf{i}_\Delta^T \quad (23)$$

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

$$\mathbf{f}_{\text{mpdr};k}[l] = \Phi_{\mathbf{y};k}^{-1}[l] \mathbf{A}_k \left[\mathbf{A}_k^H \Phi_{\mathbf{y};k}^{-1}[l] \mathbf{A}_k \right]^{-1} \mathbf{i}_\Delta \quad (24)$$

4. True-MPDR with the Single-Sideband Transform

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

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

$$\mathbf{y}'_k[l] = \begin{bmatrix} \mathbf{y}_k[l] \\ \mathbf{y}_{K-k}[l] \end{bmatrix}_{2ML_Y \times 1} \quad (25)$$

from which we define $\Phi_{\mathbf{y}'_k}[l]$ as its correlation matrix. Under this idea, our filter $\mathbf{f}_k[l]$ is a $2ML_Y \times 1$ 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 $\mathbf{f}_k[l]$ be $\mathbf{f}_{k,\mathcal{F}}[l]$, given by

$$\mathbf{f}_{\mathcal{F},k}[l] = \Lambda \mathbf{f}_k[l] \quad (26)$$

in which Λ is a $ML_Y \times 2ML_Y$ matrix,

$$\Lambda = \frac{1}{\sqrt{2}} \begin{bmatrix} e^{-j\frac{3\pi}{4}} & 0 & \dots & 0 & e^{j\frac{3\pi}{4}} & 0 & \dots & 0 \\ 0 & e^{-j\frac{3\pi}{4}} & \dots & 0 & 0 & e^{j\frac{3\pi}{4}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & e^{-j\frac{3\pi}{4}} & 0 & 0 & \dots & e^{j\frac{3\pi}{4}} \end{bmatrix} \quad (27)$$

From Eq. (26) the distortionless constraint from the STFT, within the SSBT domain, becomes

$$\mathbf{f}_k^H[l] \mathbf{A}'_k = \mathbf{i}_\Delta^T \quad (28a)$$

$$\mathbf{A}'_k = \Lambda^H \mathbf{A}_{\mathcal{F},k} \quad (28b)$$

where $\mathbf{A}_{\mathcal{F},k}$ is the constraint matrix within the STFT domain, and \mathbf{A}'_k is the new constraint matrix within the SSBT domain.

In this scheme, our minimization problem becomes

$$\mathbf{f}_{\text{mpdr},k}[l] = \min_{\mathbf{f}_k[l]} \mathbf{f}_k^H[l] \Phi_{\mathbf{y}'_k}[l] \mathbf{f}_k[l] \text{ s.t. } \mathbf{f}_k^H[l] \mathbf{A}'_k = \mathbf{i}_\Delta^T \quad (29)$$

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

4.1. Real-valued true-MPDR beamformer with SSBT

To ensure the desired behavior of $\mathbf{f}_k[l]$ being real-valued, an additional constraint is necessary. By forcing $\mathbf{f}_k[l]$ to have real entries, from Eq. (28a) we trivially have that

$$\mathbf{f}_k^T[l] \Re\{\mathbf{A}'_k\} = 1 \quad (30a)$$

$$\mathbf{f}_k^T[l] \Im\{\mathbf{A}'_k\} = 0 \quad (30b)$$

which can be put in matricial form as $\mathbf{f}_k^T[l]\mathbf{A}_k = \mathbf{i}_\Delta^T$, with

$$\mathbf{A}_k = \begin{bmatrix} \Re\{\mathbf{A}'_k\}, & \Im\{\mathbf{A}'_k\} \end{bmatrix}_{2ML_Y \times 2L} \quad (31a)$$

$$\mathbf{i}_\Delta = [0, \dots, 0, 1, 0, \dots, 0]_{2ML_Y \times 1}^T \quad (31b)$$

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

$$\mathbf{f}_{\text{mpdr};k}[l] = \min_{\mathbf{f}_k[l]} \mathbf{f}_k^T[l] \Phi_{\mathbf{y}'_k}[l] \mathbf{f}_k[l, k] \text{ s.t. } \mathbf{f}_k^T[l] \mathbf{A}_k = \mathbf{i}_\Delta^T \quad (32)$$

whose solution is

$$\mathbf{f}_{\text{mpdr};k}[l] = \Phi_{\mathbf{y}'_k}^{-1}[l] \mathbf{A}_k \left[\mathbf{A}_k^T \Phi_{\mathbf{y}'_k}^{-1}[l] \mathbf{A}_k \right]^{-1} \mathbf{i}_\Delta \quad (33)$$

Using Eq. (26), we can obtain the desired beamformer $\mathbf{f}_{\mathcal{F},\text{mpdr};k}[l]$, transformed to the STFT domain.

5. Perturbation Robustness Analysis

Until now, we assumed an appropriate knowledge of all signals and their impulse responses. However, in a real application these would be estimated, and thus prone to error. Given our beamformers from Eqs. (24) and (33) and their dependence on \mathbf{d} , they are directly influenced by impulse response estimation errors.

Going back to the time domain in Eq. (6), we can write the observed/measured room impulse response $h_m[n]$ for each sensor as

$$h_m[n] = h_m^*[n] + \Delta h_m[n] \quad (34)$$

with h_m^* being the accurate room impulse response, and $\Delta h_m[n]$ is a perturbation (or error) on the measurement. Through the same derivations that were presented in Section 3, we find that

$$\mathbf{A}_k = \mathbf{A}_k^* + \Delta \mathbf{A}_k \quad (35)$$

This will also have an effect on the beamformers designed with an inaccurate steering matrix \mathbf{A}_k , leading to errors in the signal estimation and beamforming output. Note that $\Delta \mathbf{A}_k$ will depend on both $h_m^*[n]$ and $\Delta h_m[n]$ given the deconvolution process necessary to obtain the relative transfer functions. From now on we will omit the $[l]$ dependence on the variables, unless strictly necessary.

The solutions obtained at the end of both Sections 3 and 4 have the same form. Expanding the inverse in the beamformer's solution from Eq. (24), we get

$$\begin{aligned} \left[\mathbf{A}_k^H \Phi_{\mathbf{y}'_k}^{-1} \mathbf{A}_k \right]^{-1} &= \left[(\mathbf{A}_k^* + \Delta \mathbf{A}_k)^H \Phi_{\mathbf{y}'_k}^{-1} (\mathbf{A}_k^* + \Delta \mathbf{A}_k) \right]^{-1} \\ &= \left[\mathbf{A}_k^{*H} \Phi_{\mathbf{y}'_k}^{-1} \mathbf{A}_k^* + \Delta \mathbf{A}_k^H \Phi_{\mathbf{y}'_k}^{-1} \mathbf{A}_k^* + \mathbf{A}_k^{*H} \Phi_{\mathbf{y}'_k}^{-1} \Delta \mathbf{A}_k + \Delta \mathbf{A}_k^H \Phi_{\mathbf{y}'_k}^{-1} \Delta \mathbf{A}_k \right]^{-1} \end{aligned} \quad (36)$$

We will denote $\Omega = \mathbf{A}_k^{*H} \Phi_{\mathbf{y}'_k}^{-1} \mathbf{A}_k^*$ and Ψ as the rest. Using the Woodbury identity, we have that

$$[\Omega + \Psi]^{-1} = \Omega^{-1} - \Omega^{-1} [\Omega^{-1} + \Psi^{-1}]^{-1} \Omega^{-1} \quad (37)$$

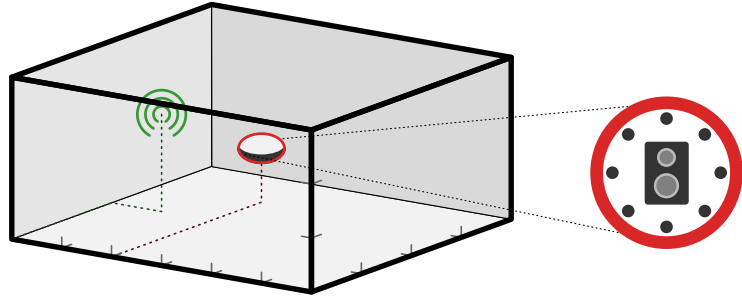


Figure 1. Room layout for simulations.

Going back to Eq. (24), we have that

$$\begin{aligned} \mathbf{f}_{\text{mpdr};k} &= \Phi_{y;k}^{-1}(\mathbf{A}_k^* + \Delta \mathbf{A}_k) \left(\Omega^{-1} - \Omega^{-1} \left[\Omega^{-1} + \Psi^{-1} \right]^{-1} \Omega^{-1} \right) \mathbf{i}_\Delta \\ &= \mathbf{f}_{\text{mpdr};k}^* + \mathbf{f}_{\delta;k} \end{aligned} \quad (38)$$

in which

$$\mathbf{f}_{\text{mpdr};k}^* = \Phi_{y;k}^{-1} \mathbf{A}_k^* \left[\mathbf{A}_k^{*H} \Phi_{y;k}^{-1} \mathbf{A}_k^* \right]^{-1} \mathbf{i}_\Delta \quad (39a)$$

$$\mathbf{f}_{\delta;k} = \Phi_{y;k}^{-1} \left[\Delta \mathbf{A}_k \Omega^{-1} - (\mathbf{A}_k^* + \Delta \mathbf{A}_k) \left(\Omega^{-1} \left[\Omega^{-1} + \Psi^{-1} \right]^{-1} \Omega^{-1} \right) \right] \mathbf{i}_\Delta \quad (39b)$$

Thus, the resulting beamformer is composed of the accurate beamformer $\mathbf{f}_{\text{mpdr};k}^*$ and one that is dependent on both the accurate and the perturbation steering matrices, $\mathbf{f}_{\delta;k}$, both being dependent on $[l]$. Of course, when $\Delta \mathbf{A}_k \rightarrow \mathbf{0}$, then $\mathbf{f}_{\delta;k} \rightarrow \mathbf{0}$, and $\mathbf{f}_{\text{mpdr};k} \rightarrow \mathbf{f}_{\text{mpdr};k}^*$.

This same process is valid for the beamformer in Eq. (33), with $\underline{\mathbf{A}}_k$, $\Phi_{y';k}$ and $\underline{\mathbf{i}}_\Delta$ instead of \mathbf{A}_k , $\Phi_{y;k}$ and \mathbf{i}_Δ , respectively, and also taking the transpose instead of the hermitian-transpose.

6. Comparisons and simulations

In the simulations², 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 $4\text{m} \times 6\text{m} \times 3\text{m}$ (width \times length \times height), with a reverberation time of 0.3s. The device composed of the loudspeaker + sensors is centered at (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_0mniPower4296.flac
$s[n]$	(3m, 2m, 1m)	69_abba_ch8_0mniPower4296.flac
$r[n]$	~	wgn_48kHz_ch8_0mniPower4296.flac

Table 1. Source information for the simulations.

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

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. (24), as well as the first with the SSBT (which will be called "Single-Frequency SSBT", or "SF-SSBT" for short). The second one based on the SSBT will be called "Dual-Frequency SSBT" (or "DF-SSBT"), as derived in Section 4, which led to Eq. (33). These names were chosen given that the one proposed in Section 4 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.

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

$$\zeta_x[l] = \frac{\sum_k |X_1[l, k]|^2}{\sum_k |X_f[l, k]|^2} \quad (40a)$$

$$\zeta_s[l] = \frac{\sum_k |S_1[l, k]|^2}{\sum_k |S_f[l, k]|^2} \quad (40b)$$

$$\zeta_r[l] = \frac{\sum_k |V_1[l, k]|^2}{\sum_k |V_f[l, k]|^2} \quad (40c)$$

where $S_f[l, k] = \mathbf{f}^H[l, k]\mathbf{s}_1[l, k]$, $X_f[l, k] = \mathbf{f}^H[l, k]\mathbf{x}_1[l, k]$ and $R_f[l, k] = \mathbf{f}^H[l, k]\mathbf{r}_1[l, k]$ 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.

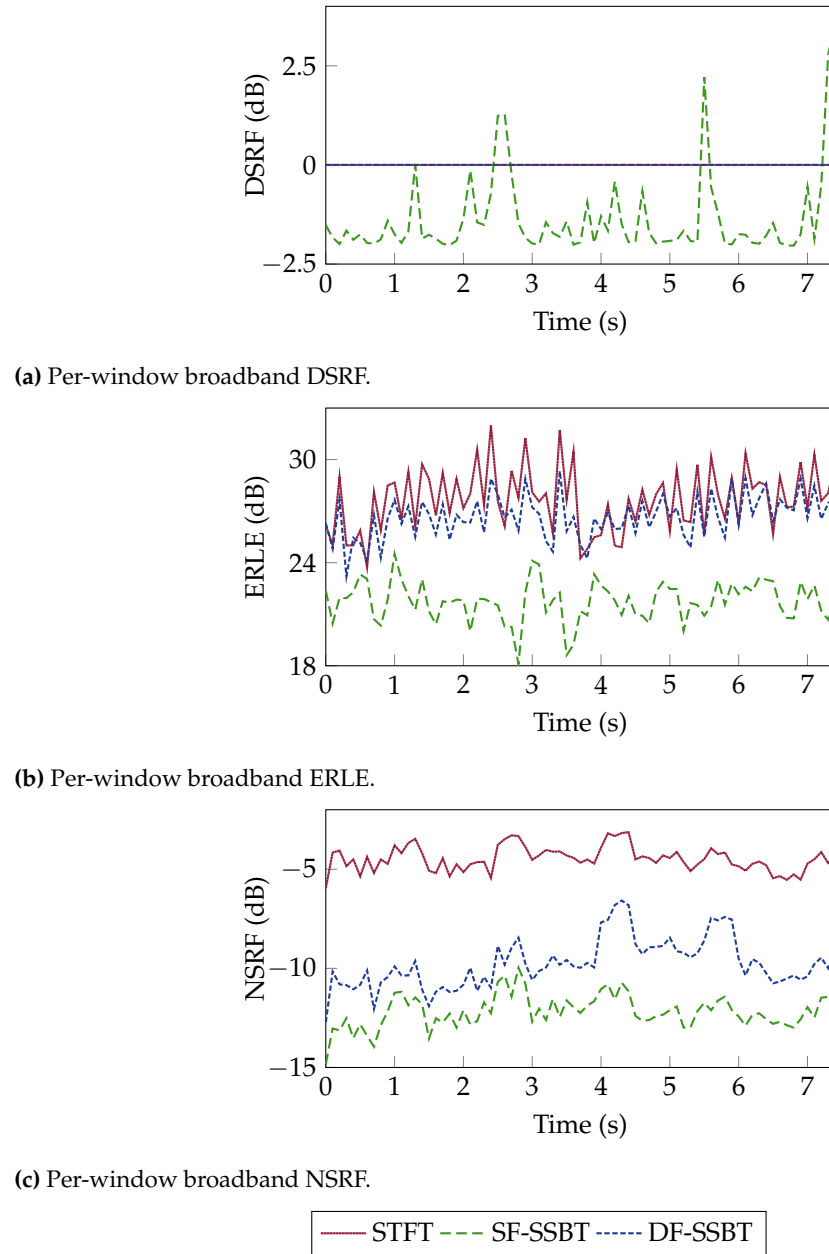


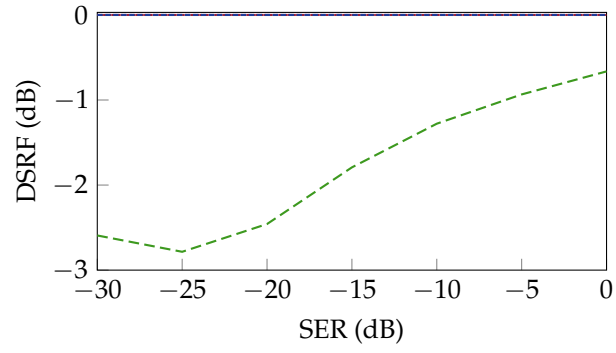
Figure 2. Output metrics for the beamformers over time, in the base scenario.

6.2. Comparison over different input SERs

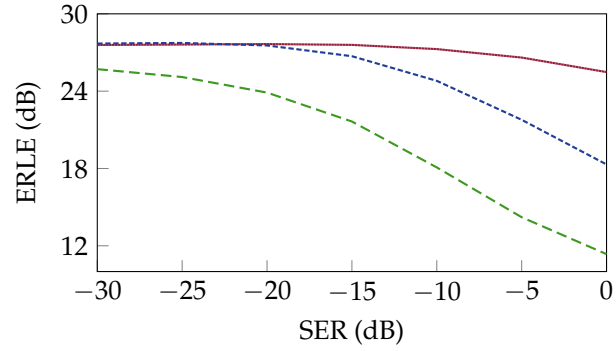
We now examine the results with a varying input SERs, to assess the beamformer's performances for different loudspeaker signal levels. For such, we will use the time-average metrics, as presented below. The other parameters and variables are maintained from the previous simulations, with only the SER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

$$\xi_x = \frac{\sum_{l,k} |X_1[l,k]|^2}{\sum_{l,k} |X_f[l,k]|^2} \quad (41a)$$

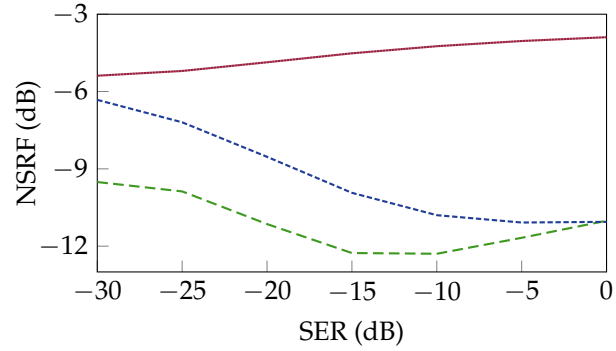
$$\xi_s = \frac{\sum_{l,k} |S_1[l,k]|^2}{\sum_{l,k} |S_f[l,k]|^2} \quad (41b)$$



(a) Time-average broadband DSRF.



(b) Time-average broadband ERLE.



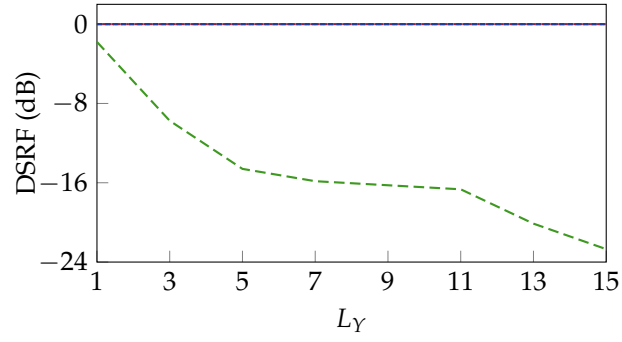
(c) Time-average broadband NSRF.

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

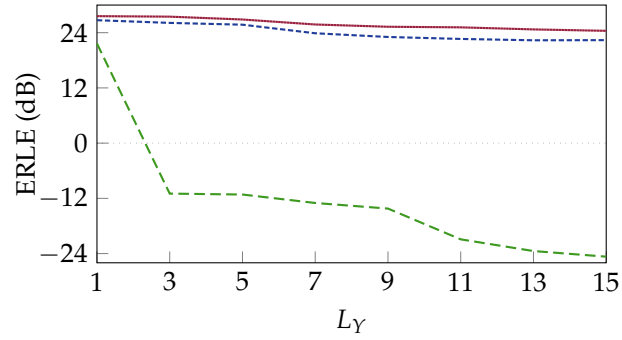
$$\xi_r = \frac{\sum_{l,k} |V_1[l, k]|^2}{\sum_{l,k} |V_f[l, k]|^2} \quad (41c)$$

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.

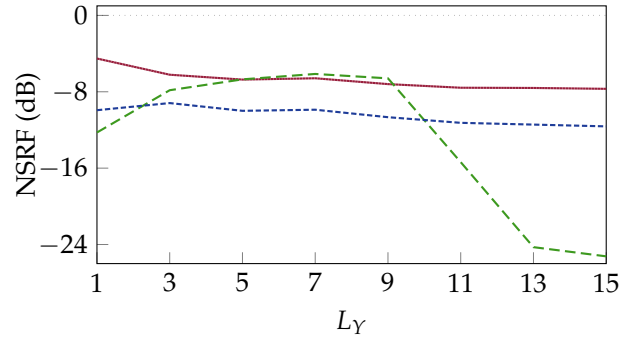
In terms of ERLE, we see that the SF-SSBT is strictly worse than the two other beamformers, for all SER's. The DF-SSBT beamformer has a similar performance to that of the STFT beamformer for iSER $\lesssim -20$ dB, but is outperformed for higher iSER's.



(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.

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.

6.3. 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 behavior, and the ERLE. This is likely due to the mathematical results exposed in Appendix A, and since the SSBT transform doesn't respect

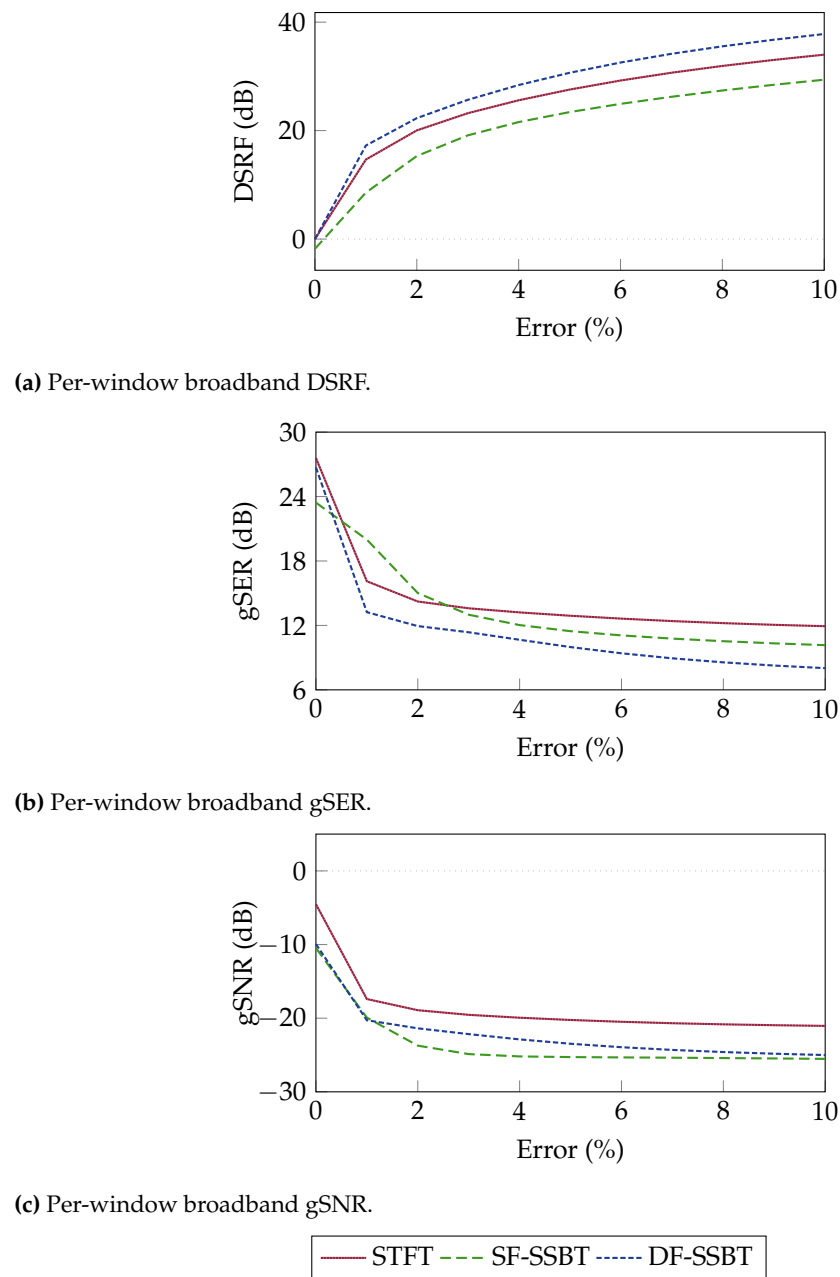


Figure 5. Output metrics for the beamformers with error in the steering vectors.

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".

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.

6.4. Comparison with perturbation

As exposed in Section 5, 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

$$\underline{\mathbf{A}}_k = \underline{\mathbf{A}}_k^* + \Delta \underline{\mathbf{A}}_k \quad (42)$$

where $\underline{\mathbf{A}}_k^*$ 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

$$\text{gSER} = \frac{\xi_s}{\xi_x} \quad (43a)$$

$$\text{gSNR} = \frac{\xi_r}{\xi_x} \quad (43b)$$

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 $\underline{\mathbf{A}}_k^*$.

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	Convulsive 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

295

Appendix A. SSBT Convolution

296

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

297

298

299

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

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

300

301

It is easy to see that

302

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

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

303

304

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

305

306

307

$$Y_{\mathcal{F}}[l, k] = H_{\mathcal{F}}[k] X_{\mathcal{F}}[l, k] \quad (\text{A.3a})$$

$$Y_{\mathcal{S}}[l, k] = H_{\mathcal{S}}[k] X_{\mathcal{S}}[l, k] \quad (\text{A.3b})$$

We will assume that the MTF model [13] correctly models the convolution here. This was 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.

308

309

310

311

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

312

313

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

Passing this through Eq. (A.2),

314

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

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

315

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

$$\begin{aligned} Y_{\mathcal{F}}[l, k] = & X_{\mathcal{F}}^{\Re}[l, k]H_{\mathcal{F}}^{\Re}[l, k] + jX_{\mathcal{F}}^{\Re}[l, k]H_{\mathcal{F}}^{\Im}[l, k] \\ & + jX_{\mathcal{F}}^{\Im}[l, k]H_{\mathcal{F}}^{\Re}[l, k] - X_{\mathcal{F}}^{\Im}[l, k]H_{\mathcal{F}}^{\Im}[l, k] \end{aligned} \quad (\text{A.6})$$

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](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.; Akin, 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, NJ, 1983.
12. Ozyerman, 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 Convolutional 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>. 367
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>. 368
20. Habets, E. RIR Generator, 2020. 369
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. 370

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. 371