A STUDY OF INVERSE SHORT-TIME FOURIER TRANSFORM

Bin Yang

Chair of System Theory and Signal Processing, University of Stuttgart, Germany

ABSTRACT

In this paper, we study the inverse short-time Fourier transform (STFT). We propose a new vector formulation of STFT. We derive a family of inverse STFT estimators and a least squares one. We discuss their relationship and compare their performance with respect to both additive and multiplicative modifications to STFT. The influence of window, overlap, and zero-padding are investigated as well.

Index Terms— inverse short-time Fourier transform, time-frequency analysis, least squares methods

1. INTRODUCTION

The short-time Fourier transform (STFT) is a useful tool to analyze nonstationary signals and time-varying systems. It has been successively applied to a large number of signal processing applications like time-frequency analysis, speech enhancement, echo cancelation, and blind source separation. The problem of the inverse STFT (ISTFT) is to construct the original time domain sequence from a modified STFT. The modification could be additive or multiplicative. Surprisingly, ISTFT has not been studied systematically in the literature. Almost all papers up to now including many recent publications use a heuristic overlap-add method for computing ISTFT [1, 2, 3, 4]. It combines the results of the inverse Fourier transform of different sections by overlap-add. In [5], a weighted overlap-add procedure has been proposed, but without a guideline about the optimum weighting.

STFT is a linear operation. Due to section overlapping and zero-padding, the result of STFT contains a larger number of samples than the original time domain sequence. Clearly, ISTFT is a linear overdetermined problem for which the least squares (LS) approach is well applicable [6].

In this paper, we take a detailed look at ISTFT. In comparison to [6], our contributions are: 1) We derive a novel compact vector formulation of STFT. It facilitates the working with STFT and is also useful for other purposes. 2) We derive a family of heuristic ISTFT estimators including the classical overlap-add method. 3) We also derive a LS estimator. While [6] assumed a continuous-frequency representation, we focus on the discrete-frequency Fourier transform. In addition, we allow zero-padding and non-equally spaced frequencies. 4) We clarify the relationship between different ISTFT estimators. 5) We compare their performance with respect to both additive and multiplicative modifications to STFT. 6) Finally, we also study the influence of window, overlap, and zero-padding on ISTFT.

The following notations are used in the paper. Matrices and column vectors are represented by boldface and underlined characters. The superscript ^{T} and ^{H} denote transpose and Hermitian transpose, respectively. $\|\cdot\|$ is the Euclidean vector norm. $\mathrm{diag}(\cdot)$ describes a diagonal matrix.

2. VECTOR FORMULATION OF STFT

We first derive a new vector formulation of STFT. This will be the basis for further investigations.

A time domain sequence $x(n)$ $(n \geq 0)$ is divided into M overlapping sections. Each section has the length N. The shift length from section to section is $1 \leq S \leq N$. The overlap length between two adjacent sections is $N - S$, see Fig. 1. Let

$$\underline{x}_m = [x(mS), x(mS+1), \ldots, x(mS+N-1)]^T \in \mathbb{C}^N \quad (1)$$

denote the m-th section $(0 \leq m \leq M-1)$ of $x(n)$. M such overlapping sections with a shift length S contain a total number of $J = (M-1)S + N$ samples $x(0), \ldots, x(J-1)$ with $J \leq MN$. Let

$$\underline{x} = [x(0), x(1), \ldots, x(J-1)]^T \in \mathbb{C}^J \quad (2)$$

be the vector of all involved samples. The relationship between $\underline{x}$ and $\underline{x}_m$ in (1) is described by

$$[\underline{x}_0^T, \underline{x}_1^T, \ldots, \underline{x}_{M-1}^T]^T = \mathbf{O}\underline{x} \quad (3)$$

where

$$\mathbf{O} = \begin{bmatrix} \mathbf{I}_N & & & \\ \overleftrightarrow{S}\ \mathbf{I}_N & & & \\ & \ddots & & \\ & & \mathbf{I}_N & \end{bmatrix} \in \mathbb{R}^{MN \times J} \quad (4)$$

is a so called overlap matrix. It consists of M identity matrices $\mathbf{I}_N$ along the main diagonal. Each identity matrix is shifted by S columns to the right with respect to the above one. The left-multiplication of $\underline{x}$ by $\mathbf{O}$ corresponds to its segmentation into M overlapping sections as shown in Fig. 1.

Each section $\underline{x}_m$ is now weighted by a real valued window $w(n)$ of the same length. We assume that $w(n)$ is nonzero and thus invertible. Then this windowed sequence is appended by $K - N$ zeros before the K-point Fourier transform $X_m(k)$ is calculated at K discrete frequencies ω_k $(0 \leq k \leq K-1)$ with $K \geq N$. In general, ω_k do not need to

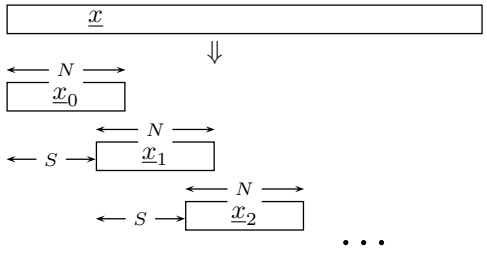

Fig. 1. $\mathbf{O}\underline{x}$: Divide a sequence $\underline{x}$ into overlapping sections

be equally spaced like in discrete Fourier transform (DFT). In matrix-vector-notation, the Fourier transform is described by

$$\underline{X}_m = [X_m(0), \ldots, X_m(K-1)]^T = \mathbf{FPW}\underline{x}_m \in \mathbb{C}^K \quad (5)$$

with

$$\begin{aligned} \mathbf{W} &= \mathrm{diag}(w(0), \ldots, w(N-1)) \in \mathbb{R}^{N\times N}, \\ \mathbf{P} &= \begin{bmatrix} \mathbf{I}_N \\ \mathbf{0}_{(K-N)\times N} \end{bmatrix} \in \mathbb{R}^{K\times N}, \\ \mathbf{F} &= [e^{-j\omega_k n}]_{0\le k,n\le K-1} \in \mathbb{C}^{K\times K}. \end{aligned} \quad (6)$$

$\mathbf{W}$ is a diagonal window matrix. $\mathbf{P}$ consists of an $N \times N$ identity and a $(K-N)\times N$ zero matrix and describes the zero-padding. $\mathbf{F}$ is the square Fourier transform matrix with the element $e^{-j\omega_k n}$ at the k-th row and n-th column, both counted starting from zero.

We stack the Fourier transforms of all M sections into a single column vector

$$\underline{X} = [\underline{X}_0^T, \underline{X}_1^T, \ldots, \underline{X}_{M-1}^T]^T \in \mathbb{C}^{MK}. \quad (7)$$

By using (3), (5), and the notation $\mathbf{A}\otimes\mathbf{B} = [a_{ij}\mathbf{B}]_{i,j}$ for the Kronecker tensor product, we finally obtain the linear relationship between the complete time domain sequence $\underline{x}$ and its windowed zero-padded STFT $\underline{X}$

$$\begin{aligned} \underline{X} &= \begin{bmatrix} \mathbf{FPW} & & \\ & \ddots & \\ & & \mathbf{FPW} \end{bmatrix} \begin{bmatrix} \underline{x}_0 \\ \vdots \\ \underline{x}_{M-1} \end{bmatrix} \\ &= \mathbf{H}\underline{x}, \quad \mathbf{H} = (\mathbf{I}_M \otimes (\mathbf{FPW}))\mathbf{O}. \end{aligned} \quad (8)$$

The meaning of the different matrices in (8) is self-explained:

- $\mathbf{O}$: overlapped segmentation
- $\mathbf{W}$: windowing
- $\mathbf{P}$: zero-padding
- $\mathbf{F}$: discrete-frequency Fourier transform
- $\mathbf{I}_M\otimes$: section-by-section processing

3. A FAMILY OF HEURISTIC ISTFT ESTIMATORS

Starting from (8), we now study how to compute the corresponding ISTFT.

If $\underline{X}$ denotes the exact STFT of a given time domain sequence $\underline{x}$, we can uniquely determine $\underline{x}$. In practical applications, however, the STFT is often modified before it is transformed back into the time domain. In this case, there is in general no time domain sequence $\underline{x}$ whose STFT matches exactly $\underline{X}$ because $\underline{X}$ contains more elements than $\underline{x}$ if $S < N$ (true overlapping) or $K > N$ (true zero-padding). The problem is overdetermined. The question is how to compute a reasonable estimate $\hat{\underline{x}}$ for $\underline{x}$ from $\underline{X}$?

A simple but heuristic idea is based on the following observation: If we multiply both sides of (8) by the $J \times MK$ matrix $\mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{A}_p)$ from left with $\mathbf{A}_p = \mathbf{W}^{p-1}\mathbf{P}^H\mathbf{F}^{-1} \in \mathbb{C}^{N\times K}$, we obtain according to $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{AC}\otimes\mathbf{BD}$ [7] the following result

$$\begin{aligned} &\mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{A}_p)\underline{X} \\ &\quad = \mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{A}_p)(\mathbf{I}_M \otimes (\mathbf{FPW}))\mathbf{O}\underline{x} \\ &\quad = \mathbf{O}^H(\mathbf{I}_M \otimes (\mathbf{A}_p\mathbf{FPW}))\mathbf{O}\underline{x} \\ &\quad = \mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{W}^p)\mathbf{O}\underline{x}, \quad (p = 0, 1, 2 \ldots). \end{aligned} \quad (9)$$

$\mathbf{W}^p$ is a diagonal matrix containing the diagonal elements $w^p(n)$. This motivates the following family of estimates

$$\hat{\underline{x}}_p = \mathbf{D}_p^{-1}\mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{A}_p)\underline{X}, \quad \mathbf{D}_p = \mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{W}^p)\mathbf{O}. \quad (10)$$

We call it the p-ISTFT estimate.

Step	Meaning
1) $\mathbf{F}^{-1}$	inverse Fourier transform of $\underline{X}_m$
2) $\mathbf{P}^H$	keep only the first N samples of $\mathbf{F}^{-1}\underline{X}_m$
3) $\mathbf{W}^{p-1}$	weight these N samples by $w^{p-1}(n)$
4) $\mathbf{I}_M\otimes$	do the above computations for all M sections
5) $\mathbf{O}^H$	overlap-add of the results of all sections
6) $\mathbf{D}_p^{-1}$	final normalization

Table 1. Steps of p-ISTFT

Note that Eq. (10) has a simple interpretation. Table 1 summarizes all steps of p-ISTFT. The first four steps are easy to understand. According to the definition of the overlap matrix $\mathbf{O}$ in (4), the operation $\mathbf{O}^H\underline{z}$ in the 5-th step with $\underline{z} = [\underline{z}_0^T, \ldots, \underline{z}_{M-1}^T]^T \in \mathbb{C}^{MN}$ describes the well known overlap-add of the M sections $\underline{z}_m$. The shadowed areas in Fig. 2 represent the overlap-add regions. Fig. 3 illustrates the matrix overlap-add operation $\mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{U})\mathbf{O}$ where $\mathbf{U}$ is any $N \times N$ matrix. The result is a $J \times J$ matrix in which adjacent matrices $\mathbf{U}$ along the main diagonal overlap and add in an $(N-S)\times(N-S)$ area. Clearly, if $\mathbf{U} = \mathbf{W}^p = \mathrm{diag}(w^p(0), \ldots, w^p(N-1))$ is diagonal, then the matrix $\mathbf{D}_p = \mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{W}^p)\mathbf{O}$ is diagonal as well

$$\mathbf{D}_p = \mathrm{diag}(d_p(0), \ldots, d_p(J-1)) \in \mathbb{R}^{J\times J}. \quad (11)$$

The diagonal elements $d_p(n)$ are obtained by overlap-adding M sections of $w^p(0), \ldots, w^p(N-1)$ as in Fig. 2. The last normalization step in Table 1 involves thus only J scalar divisions.

3.1. LS inverse STFT estimator

Since ISTFT is a linear overdetermined problem, it is natural to apply the least squares (LS) approach to the signal model

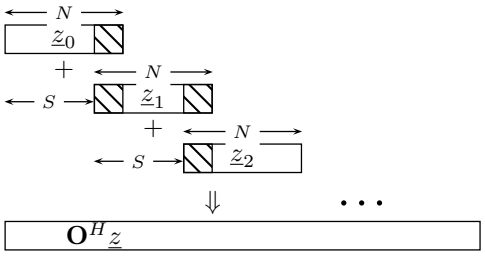

Fig. 2. $\mathbf{O}^H\underline{z}$: Overlap-add of vectors

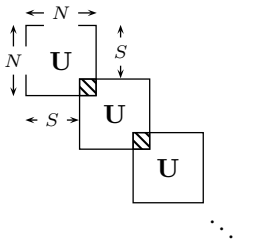

Fig. 3. $\mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{U})\mathbf{O}$: Overlap-add of matrices

(8). After some calculations, we obtain the LS estimate

$$\begin{aligned} \underline{\hat{x}}_{\mathrm{LS}} &= (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\underline{X} \\ &= \mathbf{D}_{\mathrm{LS}}^{-1}\mathbf{O}^H(\mathbf{I}_M \otimes (\mathbf{W}\mathbf{P}^H\mathbf{F}^H))\underline{X}, \\ \mathbf{D}_{\mathrm{LS}} &= \mathbf{H}^H\mathbf{H} = \mathbf{O}^H(\mathbf{I}_M \otimes \mathbf{U})\mathbf{O}, \\ \mathbf{U} &= \mathbf{W}\mathbf{P}^H\mathbf{F}^H\mathbf{F}\mathbf{P}\mathbf{W}. \end{aligned} \tag{12}$$

It is referred to as LS-ISTFT.

4. DISCUSSIONS

Below we discuss the relationship between different ISTFT:

- The classical overlap-add method [1, 2, 3, 4] is a special case of our p-ISTFT with $p = 1$.
- The LS solution from [6] is identical to p-ISTFT with $p = 2$. It assumed, however, a continuous-frequency Fourier transform.
- Our LS-ISTFT is derived for the discrete-frequency case. In general, i.e. for arbitrary discrete frequencies ω_k, $\mathbf{U}$ and thus $\mathbf{D}_{\mathrm{LS}}$ in (12) are not diagonal. In this case, the matrix inversion $\mathbf{D}_{\mathrm{LS}}^{-1}$ is expansive and $\underline{\hat{x}}_{\mathrm{LS}}$ differs from $\underline{\hat{x}}_p$ and the LS solution from [6].
- In the special case of DFT with equally spaced discrete frequencies $\omega_k = \frac{2\pi k}{K}$ $(0 \le k \le K-1)$, $\mathbf{F}^H\mathbf{F} = K\mathbf{I}_K, \mathbf{U} = K\mathbf{W}^2, \mathbf{F}^H = K\mathbf{F}^{-1}$, and $\underline{\hat{x}}_{\mathrm{LS}}$ simplifies to $\underline{\hat{x}}_2$.

Besides the choice of the estimator, the inverse STFT also depends on a number of other factors like the modification of the STFT, the choice of the window, the shift length S, and the number of appended zeros $K - N$. Below we study the influence of these factors on ISTFT.

- If $\underline{X}$ denotes the exact STFT of a time domain sequence $\underline{x}$, then both $\underline{\hat{x}}_p$ and $\underline{\hat{x}}_{\mathrm{LS}}$ return the same $\underline{x}$. This can be easily shown be combining (8) and (10) as well as (12).
- If we use the rectangular window $w(n) = 1$, then all estimators $\underline{\hat{x}}_p$ return the same result. In this case, $\mathbf{W} = \mathbf{I}_N$ and $\underline{\hat{x}}_p$ in (10) does not depend on p.
- If there is no overlap between adjacent sections ($S = N$), then all estimators $\underline{\hat{x}}_p$ return the same result as well. In this case, $\mathbf{O} = \mathbf{I}_{MN}$ and $\underline{\hat{x}}_p$ simplifies to $(\mathbf{I}_M \otimes (\mathbf{W}^{-1}\mathbf{P}^H\mathbf{F}^{-1}))\underline{X}$.

We expect a small difference between various p-ISTFT estimators if the modification to STFT or the deviation of $w(n)$ from the rectangular window or the overlap length is small.

Concerning the computational complexity, all p-ISTFT estimators and the DFT-based version of LS-ISTFT are comparable to the classical overlap-add method. For each section, one inverse Fourier transform has to be computed. The only difference is the use of different windows $w^{p-1}(n)$ for the scaling of the inverse Fourier transform and the computation of the final normalization sequence $d_p(n)$ in (11).

5. EXPERIMENTS

In this section, we compare the performance of different ISTFT estimators. We use clean speech signals of roughly 6 second duration sampled at 16 kHz in our experiments. For each speech signal $\underline{x}$, we compute its STFT and modify it by additive noise or multiplicative masking. Then we compute the signal estimate $\underline{\hat{x}}$ for different ISTFT estimators. Since DFT (FFT) is used, $\underline{\hat{x}}_{\mathrm{LS}}$ is identical to $\underline{\hat{x}}_2$ and will not be considered separately. The performance measure is the signal-to-distortion ratio (SDR) in the time domain after the signal reconstruction

$$\mathrm{SDR}_p = 10\log_{10}(\|\underline{x}\|^2/\|\underline{x} - \underline{\hat{x}}_p\|^2)\ \mathrm{dB}. \tag{13}$$

In particular, we focus on SDR_2 of the LS estimator and the performance loss of other estimators $\Delta\mathrm{SDR}_p = \mathrm{SDR}_2 - \mathrm{SDR}_p$ with respect to that. For statistical averaging, we use 8 different utterances from a male and a female speaker and calculate the average values of SDR_2 and $\Delta\mathrm{SDR}_p$ over these 8 speech signals. The default parameter set for the Fourier transform is hamming window, window length $N = 512$, shift length $S = N/2$, and FFT length $K = N$ unless otherwise stated.

Additive distortion

First we consider additive distortion. $\underline{X}$ is modeled as $\mathbf{H}\underline{x} + \underline{N}$. $\mathbf{H}\underline{x}$ is the exact STFT of $\underline{x}$ and $\underline{N}$ contains realizations of zero-mean uncorrelated random variables with equal variance σ^2. The variance is chosen to achieve a certain signal-to-noise ratio (SNR) in the time-frequency domain $\mathrm{SNR} = 10\log_{10}(\|\mathbf{H}\underline{x}\|^2/\|\underline{N}\|^2)$. For this particular signal model, it is well known that the LS estimator $\underline{\hat{x}}_{\mathrm{LS}} = \underline{\hat{x}}_2$ achieves the smallest variance among all linear unbiased estimators like $\underline{\hat{x}}_p$. In addition, the variance of $\underline{\hat{x}}_p$ increases linearly with σ^2.

The default value of SNR is 10 dB. In Table 2, we use different invertible windows. The window names are taken from MATLAB. In Table 3, we vary the overlap length $N-S$. In Table 4, we change the FFT length K.

Multiplicative distortion

In a second series of experiments, we multiply the exact STFT

with a mask, a typical operation in underdetermined blind source separation based on the spareness of speech signals in the time-frequency domain [4]. Due to limited space, we only consider a binary mask. The mask is determined such that p_{mask} percentage of the time-frequency points with the highest amplitude of $\mathbf{H}\underline{x}$ pass the mask. In all other time-frequency points, the binary mask is zero.

In Table 5, we choose different values for p_{mask}. The other parameters are identical to the default choice in the previous subsection. In Table 6 to 8, we keep $p_{\mathrm{mask}} = 30\%$ and repeat the same performance study with respect to window, overlap, and zero-padding as previously.

window	SDR_2	$\Delta\mathrm{SDR}_0$	$\Delta\mathrm{SDR}_1$	$\Delta\mathrm{SDR}_3$
rectwin	13.00	0	0	0
kaiser (β=0)	13.00	0.01	0	0
hamming	12.47	9.26	1.08	0.10
triang	7.22	20.30	0.33	0.03

Table 2. Additive distortion: Varying window

$N-S$	SDR_2	$\Delta\mathrm{SDR}_0$	$\Delta\mathrm{SDR}_1$	$\Delta\mathrm{SDR}_3$
0	0.23	0	0	0
$N/8$	5.00	0.82	0.17	0.07
$N/4$	8.50	3.97	0.55	0.09
$N/2$	12.47	9.26	1.08	0.10
$3N/4$	15.27	9.13	1.15	0.20

Table 3. Additive distortion: Varying overlap length

K	SDR_2	$\Delta\mathrm{SDR}_0$	$\Delta\mathrm{SDR}_1$	$\Delta\mathrm{SDR}_3$
N	12.47	9.28	1.08	0.10
$1.5N$	14.25	9.30	1.10	0.10
$2N$	15.50	9.34	1.10	0.10

Table 4. Additive distortion: Varying zero-padding

Observations

We draw the following conclusions from the above experiments:

- In additive distortion, $\underline{\hat{x}}_{\mathrm{LS}} = \underline{\hat{x}}_2$ is the best one as expected. In binary masking, $\underline{\hat{x}}_{\mathrm{LS}}$ is almost the best one among the tested linear estimators, but not always because of $\Delta\mathrm{SDR}_3 < 0$ in Table 6 and 8.
- $\underline{\hat{x}}_0$ has a poor performance. The weighting of the inverse Fourier transform with $w^{-1}(n)$ amplifies the distortion if $w(n)$ is close to zero. This happens at both end of each section, particularly for the window "triang".
- The classical overlap-add method $\underline{\hat{x}}_1$ is always worse than the LS one with a SDR loss of up to 1 dB. Interestingly, it is also worse than $\underline{\hat{x}}_3$.
- There is almost no performance difference between $\underline{\hat{x}}_2$ and $\underline{\hat{x}}_3$.
- As expected, the larger the overlap length is, the larger the SDR improvement of $\underline{\hat{x}}_2$ is.
- Zero-padding has a fairly small impact to the performance difference.
- In order to achieve a good absolute performance SDR_2, a large overlap is highly desirable resulting in a larger number of noisy samples. Also zero-padding is advantageous though the improvement is much smaller.
- For additive distortion, flat windows like "rectwin, kaiser, hamming" are better. For multiplicative distortion, "hamming, triang" windows are preferred. Hence the hamming window seems to be a good compromise.

p_{mask}	SDR_2	$\Delta\mathrm{SDR}_0$	$\Delta\mathrm{SDR}_1$	$\Delta\mathrm{SDR}_3$
10%	21.32	5.48	0.29	0.07
20%	28.16	6.31	0.43	0.06
30%	33.60	7.41	0.60	0.06
40%	38.74	8.75	0.82	0.06

Table 5. Multiplicative distortion: Varying mask

window	SDR_2	$\Delta\mathrm{SDR}_0$	$\Delta\mathrm{SDR}_1$	$\Delta\mathrm{SDR}_3$
rectwin	29.18	0	0	0
kaiser (β=0)	29.40	0.18	0.09	-0.09
hamming	33.60	7.41	0.60	0.06
triang	33.52	21.74	0.36	0.06

Table 6. Multiplicative distortion: Varying window

$N-S$	SDR_2	$\Delta\mathrm{SDR}_0$	$\Delta\mathrm{SDR}_1$	$\Delta\mathrm{SDR}_3$
0	23.51	0	0	0
$N/8$	28.85	1.59	0.45	-0.03
$N/4$	31.33	3.88	0.43	0.09
$N/2$	33.60	7.41	0.60	0.06
$3N/4$	34.49	5.87	0.51	0.02

Table 7. Multiplicative distortion: Varying overlap length

K	SDR_2	$\Delta\mathrm{SDR}_0$	$\Delta\mathrm{SDR}_1$	$\Delta\mathrm{SDR}_3$
N	33.60	7.41	0.60	0.06
$1.5N$	33.90	5.95	0.38	0.06
$2N$	34.02	5.72	0.35	0.06

Table 8. Multiplicative distortion: Varying zero-padding

6. REFERENCES

[1] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, pp. 1558–1564, 1977.

[2] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, Prentice-Hall, 1978.

[3] "Inverse short-time FFT," MATLAB Signal Processing Blockset Documentation.

[4] S. Araki, H. Sawada, et al., "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, pp. 1833–1847, 2007.

[5] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, pp. 99–102, 1980.

[6] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–243, 1984.

[7] A. Graham, Ed., *Kronecker Products and Matrix Calculus with Applications*, John Wiley & Sons, 1981.