Time-Frequency Signal Reconstruction of Nonsparse Audio Signals

Isidora Stanković, Miloš Daković
Faculty of Electrical Engineering
University of Montenegro
81000 Podgorica, Montenegro
Email: {isidoras,milos}@ac.me

Cornel Ioana
GIPSA Lab, INP Grenoble
University of Grenoble Alpes
38400 Saint-Martin-d'Hères, France
Email: cornel.ioana@gipsa-lab.grenoble-inp.fr

***Abstract*—In this paper, the reconstruction of non-stationary audio signals is considered. Audio signals are approximately sparse in the joint time-frequency representation domain. The reconstruction is based on a reduced set of samples, and it is considered that the signals are sparse. The short-time Fourier transform (STFT) is considered as the representation domain where the audio signals are sparse. The formula for error caused by the reconstruction of approximately sparse signals under the sparsity assumption is derived. The results are numerically illustrated on three audio signals.**

***Keywords*—audio signals, compressive sensing, reconstruction, sparsity, time-frequency analysis**

I. Introduction

A non-stationary signal, that covers most of the time and frequency domain, may be well localised in the joint time-frequency domain. If time and frequency are considered separately, these signals are dense in both domains. An example of a non-stationary signal is the audio signal. The audio signals could be located within much smaller regions in the time-frequency domain using appropriate time-frequency representations [1–6]. The most basic time-frequency representation is the short-time Fourier transform (STFT). If only few coefficients in time-frequency domain are nonzero, compared to the total number of coefficients, then the signal is sparse in this transformation domain.

By compressive sensing (CS) theory, a signal that is sparse in a certain domain can be reconstructed with a reduced set measurements (signal samples) than required by the standard sampling theorem [7–11]. Reducing the number of available samples in the analysis manifests as a noise, whose properties in the DFT are studied in [12]. These results will be used to define noise properties in the STFT of audio signals considered here. The results presented here are based on the noise analysis in the two-dimensional DFT done in [13]. Let us consider the case when a nonsparse signal is reconstructed with a reduced set of available samples. Then the non-reconstructed components will behave as the noise related to the number of the missing samples. It will be treated as an additive input noise in the reconstructed signal. The relation for the mean square error is derived for the case of the STFT matrix. The main result is the error in reconstructed signal related to the energy of non-reconstructed components, the number of missing samples, and signal sparsity. The results are tested on audio signals, as one of the most common applications of the STFT representation.

The paper is organised as follows. The theoretical background of compressive sensing and time-frequency signal analysis is shown in Section II. The formula for the influence of nonsparsity on the reconstructed signal is presented in Section III. The numerical results are given in Section IV.

II. Theoretical Background

Consider a general form of a multicomponent signal

$$x(n) = \sum_{l=1}^{C} x_l(n), \tag{1}$$

with C non-stationary components $x_l(n)$, $l = 1, 2, \ldots, C$. We will assume that the signal is sparse in the STFT domain. The STFT of the discrete-time signal is defined as

$$S_N(n,k) = \sum_{m=-N/2}^{N/2-1} x(n+m)w(m)e^{-j\frac{2\pi}{N}mk}, \tag{2}$$

at instant n and frequency k. The window function of length N is $w(m)$. The windowed signal $x(n,m) = x(n+m)w(m)$, which is K-sparse in the STFT domain, can be written as

$$x(n,m) = \sum_{i=1}^{K} A_i(n)e^{j2\pi mk_i/N} \tag{3}$$

The signal and its STFT can be represented in the matrix form

$$\mathbf{S}_N(n) = \mathbf{W}_N\mathbf{H}_N\mathbf{x}(n) \tag{4}$$

$$\mathbf{H}_N\mathbf{x}(n) = \mathbf{W}_N^{-1}\mathbf{S}_N(n), \tag{5}$$

where $\mathbf{x}(n)$ is the vector of the original signal within the window, $\mathbf{W}_N$ is the DFT matrix of size $N \times N$ with coefficients $W(m,k) = e^{(-j2\pi km/N)}$ and $\mathbf{H}_N$ is the matrix with the window values at its diagonal. With suitable overlapping, the analysis and reconstruction of the whole signal based on STFT is straightforward [1], [2], [14].

With the assumption that the signal is sparse in the STFT domain, we can reconstruct it with a reduced number of samples, according to the compressive sensing theory [7], [8]. The number of randomly positioned available samples for the reconstruction is $N_A \ll N$. For a given n the available

signal samples are at the positions $n+m \in \mathbb{N}_A$, where $\mathbb{N}_A = \{n+m_1, n+m_2, ..., n+m_{N_A}\}$.

The number of unavailable/missing samples is $N_M = N - N_A$. The available samples of the windowed signal are

$$\mathbf{y}_n = [x(n+m_1)w(m_1), ..., x(n+m_{N_A})w(m_{N_A})]^T. \quad (6)$$

Note that

$$\mathbf{y}_n = \mathbf{A}\mathbf{S}_N(n),$$

where $\mathbf{A}$ is the partial inverse DFT matrix which corresponds to the positions of the available samples.

The goal of compressive sensing is reconstructing the missing samples of the original sparse signal from the available samples. A general compressive sensing formulation is

$$\min \|\mathbf{S}_N(n)\|_0 \text{ subject to } \mathbf{y}_n = \mathbf{A}\mathbf{S}_N(n).$$

In this paper we will assume that the initial STFT is calculated using the available samples only

$$S_{N0}(n,k) = \sum_{i=1}^{N_A} x(n+m_i)w(m_i)e^{-j\frac{2\pi}{N}m_i k} \quad (7)$$

$$\mathbf{S}_{N0}(n) = N\mathbf{A}^H\mathbf{y}_n. \quad (8)$$

The mean and the variance of this STFT, i.e. calculated using the available signal samples only, at a given instant n, are [12]

$$E\{S_{N0}(n,k)\} = \sum_{i=1}^{K} N_A A_i(n)\delta(k-k_i) \quad (9)$$

$$\text{var}\{S_{N0}(n,k)\} = N_A \frac{N_M}{N-1}\sum_{i=1}^{K}|A_i(n)|^2\left(1-\delta(k-k_i)\right), \quad (10)$$

where $\delta(k) = 1$ only for $k = 0$ and $\delta(k) = 0$, elsewhere.

In general, time-varying signals (such as audio signals) are not strictly sparse in the STFT domain. Because of their nonstationarity, the signals are either approximately sparse or not sparse. We say that a signal is K-sparse in a transformation domain if it has nonzero coefficients only at positions $k \in \mathbb{K} = \{k_1, k_2, \ldots, k_K\}$ and others are zero-valued. A signal is approximately sparse if the coefficients at $k \in \mathbb{K}$ are significantly larger than the coefficients at $k \notin \mathbb{K}$. A signal is said to be nonsparse if the coefficients at the positions $k \notin \mathbb{K}$ are of the same order as the coefficients at $k \in \mathbb{K}$. To use the theory of compressive sensing for any of these signals, the sparsity assumption has to be made. In this paper, we will examine the influence of the non-reconstructed coefficients on audio signals obtained by assuming that signals are K-sparse in the STFT domain.

The signal is reconstructed by estimating the positions of the nonzero components and calculating the unknown signal amplitudes $A_i(n)$ based on the known $x(n+m_i)$. The reconstruction is done in an iterative way [11]. In the first step, the position of the largest component is found as

$$k_1 = \arg\max\{\mathbf{S}_{N0}(n)\}.$$

Matrix $\mathbf{A}_1$ is formed from matrix $\mathbf{A}$ by omitting all rows except the row corresponding to the found position k_1. The first STFT estimate is

$$\mathbf{S}_{N1}(n) = (\mathbf{A}_1^H\mathbf{A}_1)^{-1}\mathbf{A}_1^H\mathbf{y}_n.$$

The signal is reconstructed and subtracted from the original signal at that position. The STFT estimate is calculated again with this new signal and its maximum position is at k_2. A new set of positions of available samples $\mathbb{K} = \{k_1, k_2\}$ is formed with the corresponding matrix $\mathbf{A}_2$. The new estimate $\mathbf{S}_{N2}(n)$ is calculated and the signal is reconstructed. The procedure is repeated K times. The procedure can be presented with a pseudo-code:

$\mathbf{K} = \emptyset, \quad \mathbf{y}_r = \mathbf{y}_n$
for $i = 1 : K$
  $\mathbf{S}_{N0}(n) = N\mathbf{A}^H\mathbf{y}_r$
  $k = \arg\{\max_k |\mathbf{S}_{N0}(n)|\}$
  $\mathbf{K} = \{\mathbf{K}, k\}$
  $\mathbf{A}_K = \mathbf{A}(K,:)$
  $\mathbf{S}_{NK}(n) = (\mathbf{A}_K^T\mathbf{A}_K)^{-1}\mathbf{A}_K^T\mathbf{y}_n$
  $S_{NK}(n,k) = S_{NK}(n,k), \quad k \in \mathbf{K}$
  $S_{NK}(n,k) = 0, \quad k \notin \mathbf{K}$
  $\mathbf{s}_r = \mathbf{W}_N^{-1}\mathbf{S}_{NK}(n)$
  $\mathbf{y}_r = \mathbf{y}_n - \mathbf{s}_r, \quad$ for $\quad n \in \mathbf{N_A}$
end
$\mathbf{S}_{NR}(n) = \mathbf{W}_N\mathbf{s}_r$

The reconstructed signal STFT is $\mathbf{S}_{NR}(n)$. How the nonsparsity influences the reconstruction process will be presented next.

III. Nonsparsity in Time-Frequency Analysis

The error which is produced by the reconstruction of nonsparse signal with a sparsity constraint is calculated. We assume that the compressive sensing conditions for the reconstruction are satisfied.

Let consider a signal $x(n)$ with time-varying components. Its STFT is denoted $S_N(n,k)$. In matrix form it is $\mathbf{S}_N(n)$. The number of samples within a window is N. The available signal samples are at N_A random positions, defined by $n+m \in \mathbb{N}_A$. We assume that the signal is reconstructed as it were K-sparse and that the reconstruction conditions are met for this sparsity. The reconstructed signal with K nonzero STFT coefficients at $k \in \mathbb{K}$ is denoted by $\mathbf{S}_{NR}(n)$. The error in the reconstructed coefficients with respect to the K corresponding STFT coefficients if the original signal were used is:

$$\|\mathbf{S}_{NK}(n) - \mathbf{S}_{NR}(n)\|_2^2 = K\frac{N_M}{N_A N}\|\mathbf{S}_N(n) - \mathbf{S}_{NK}(n)\|_2^2, \quad (11)$$

where $\mathbf{S}_{NK}$ is equal to the original signal STFT $\mathbf{S}_N$ at the reconstructed positions, $\mathbf{S}_{NK}(n) = \mathbf{S}_N(n)$ for $k \in \mathbb{K}$ and $\mathbf{S}_{NK}(n) = 0$ for $k \notin \mathbb{K}$. Note that $\|\mathbf{S}_N(n)\|_2^2 = E\{\sum_k |S_N(n,k)|^2\}$ and $\mathbf{S}_{NK}(n)$ is the K-sparse version of

$\mathbf{S}_N(n)$. The elements of vector $\mathbf{S}_{NK}(n)$ are $S_{NK}(n,k) = S_N(n,k)$ for $k \in \mathbb{K}$, and $S_{NK}(n,k) = 0$ for $k \notin \mathbb{K}$. The reconstructed $\mathbf{S}_{NR}(n)$ is formed in the same way. The coefficients at $k \in \mathbb{K}$ are the results from the reconstruction procedure. The remaining coefficients are set to zero.

Since the CS conditions are satisfied, we can detect K signal components with amplitudes $A_i, i = 1 \dots K$, using the algorithm explained in Section II and perform the reconstruction. The reconstructed signal $\mathbf{S}_{NR}(n)$ has K reconstructed components. The remaining $N-K$ signal components with amplitudes $(A_{K+1}(n), A_{K+2}(n), ..., A_N(n))$ are not reconstructed. They produce noise in the reconstructed components. The variance of the noise from one non-reconstructed signal component, Eq. (10), is

$$|A_i(n)|^2 N_A N_M/(N-1). \tag{12}$$

The scaling factor is N/N_A for the reconstructed components since the signal amplitudes in $\mathbf{S}_{N0}(n)$ are proportional to N_A and we know that the amplitudes are recovered to their original values as if all samples were available (proportional to N). That means that the scaling factor for the noise variance in the reconstructed components is $(N/N_A)^2$. Therefore, the variance of noise caused by a non-reconstructed component to a reconstructed component is

$$|A_i(n)|^2 \frac{N^2}{N_A^2} \frac{N_A N_M}{N-1} \cong |A_i(n)|^2 N \frac{N_M}{N_A}. \tag{13}$$

This analysis is valid for one component signal. For a K-component signal, the white noise energy in the reconstructed components will be K times larger than the variance in one reconstructed component. Total noise caused by the non-reconstructed components is

$$\|\mathbf{S}_{NR}(n) - \mathbf{S}_{NK}(n)\|_2^2 = KN \frac{N_M}{N_A} \sum_{i=K+1}^{N} |A_i(n)|^2. \tag{14}$$

Energy corresponding to the non-reconstructed components only, can be written as

$$\|\mathbf{S}_N(n) - \mathbf{S}_{NK}(n)\|_2^2 = \sum_{i=K+1}^{N} |NA_i(n)|^2. \tag{15}$$

From (14) and (15) follows

$$\|\mathbf{S}_{NR}(n) - \mathbf{S}_{NK}(n)\|_2^2 = K \frac{N_M}{N_A N} \|\mathbf{S}_N(n) - \mathbf{S}_{NK}(n)\|_2^2.$$

The cases when the original signal is exactly of sparsity K, i.e. $\mathbf{S}_N(n) = \mathbf{S}_{NK}(n)$, and when all samples are available, i.e. $N = N_A$, produce no error.

IV. Results

The presented theory is illustrated on three audio signals. Audio signals are usually approximately sparse or nonsparse. The first considered audio signal is the benchmark signal from MATLAB. The second signal is a recorded speech, while the third one is also a benchmark signal from MATLAB with faster varying time-frequency representation.

A. Example 1

Let us consider the audio signal 'train'. This signal is included in the MATLAB software. The original STFT of the signal is shown in Fig. 1(top). The STFT is calculated with a Hann(ing) window with a half of its length overlapping. This window and overlapping allowed very simple and direct signal reconstruction. It is assumed that the sparsity is $K = 55$ and 50% of samples are missing. The STFT of the signal with the remaining 50% of available samples is shown in Fig. 1(middle). The reconstructed STFT is presented in Fig. 1(bottom). The total error in dB caused by the reconstruction for various sparsity levels K is shown in Fig. 2. The theoretical results are presented with the blue solid line and the estimated error is shown with the red stars. Agreement between theoretically obtained error energy and estimated one is very high.

B. Example 2

Now we will assume a recorded version of the words "You and I". This signal was recorded on a MacBook Air laptop using MATLAB with a sampling frequency of 44.1 kHz, 16-bit A/D conversion and single-channel mode. As in the previous example, it is assumed that there is only a half of the samples available. The assumed sparsity is $K = 75$. The original signal, the signal with available samples and the reconstructed one are shown in Fig. 3. The total error in dB caused by the reconstruction for various sparsity levels K is shown in Fig. 4. As in the previous example, agreement between theory and estimation is very high.

C. Example 3

Let us consider another built-in MATLAB audio signal 'mtlb'. The signal 'mtlb' is the spoken word "MATLAB". The original STFT of the signal is shown in Fig. 5(top). It can be seen that the time-frequency representation is varying fast in this case. We again used a Hann(ing) window with 50% of its length overlapping. It is assumed that the sparsity is $K = 150$ and 20% samples are missing. The STFT with missing samples is shown in Fig. 5(middle). The reconstructed STFT is presented in Fig. 5(bottom). The total error in dB for various sparsity levels K is shown in Fig. 6. The signal in the reconstruction is considered as sparse, although its components cover almost the whole frequency range.

V. Conclusions

The influence of nonsparsity to the reconstruction of audio signals that are approximately sparse in the time-frequency domain is analysed in this paper. The relation for the reconstruction error is derived. The reconstruction results are examined on three different examples, which include some recorded data. The derived formula agrees with the numerical calculations of the reconstruction error.

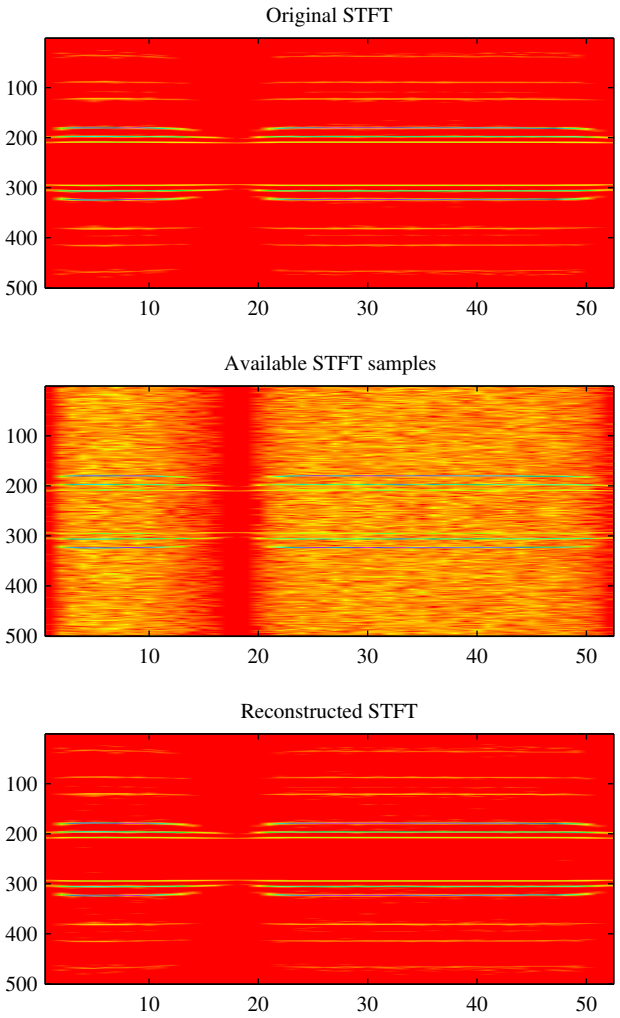

Fig. 1: STFT reconstruction of the audio signal 'train': Original STFT (top); STFT of the signal with available samples (middle); Reconstructed STFT (bottom)

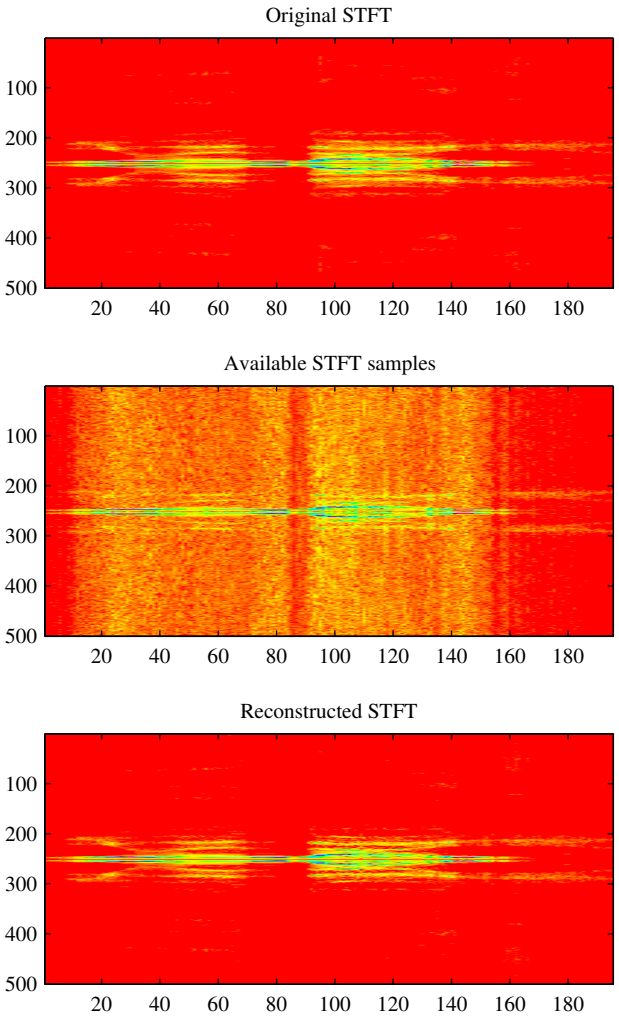

Fig. 3: STFT reconstruction of the recorded audio signal "You and I": Original STFT (top); STFT of the signal with available samples (middle); Reconstructed STFT (bottom)

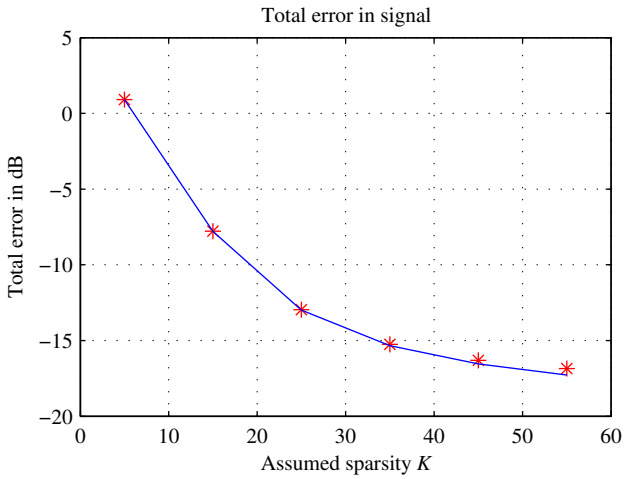

Fig. 2: Total error energy after the reconstruction with various sparsity levels of the audio signal 'train'. Blue line represents theoretically obtained error and red stars represents estimation.

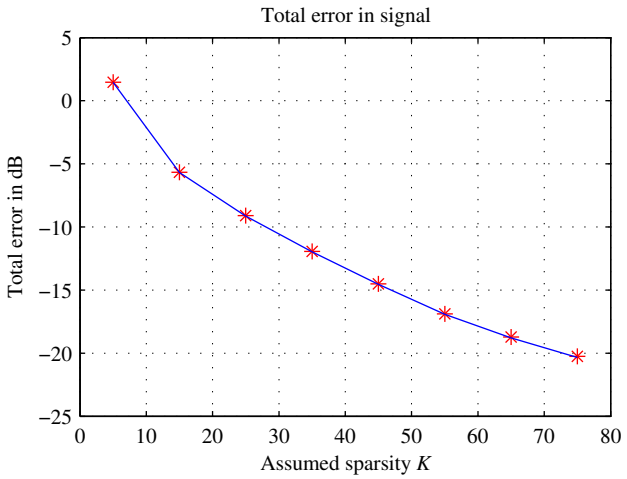

Fig. 4: Total error energy after the reconstruction with various sparsity levels of the recorded audio signal "You and I". Blue solid line represents theory, red stars represents estimation

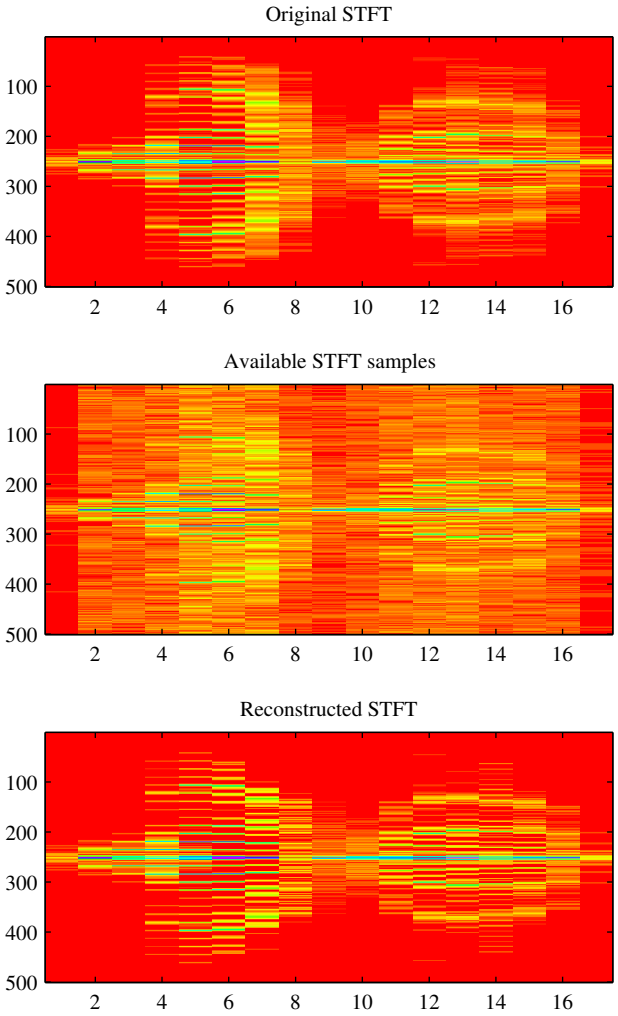

Fig. 5: STFT reconstruction of the audio signal 'mtlb': Original STFT (top); STFT of the signal with available samples (middle); Reconstructed STFT (bottom)

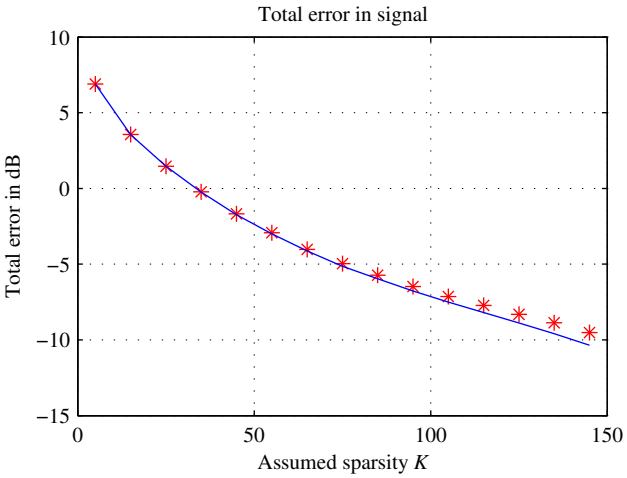

Fig. 6: Total error energy after the reconstruction with various sparsity levels of the audio signal 'mtlb'. Blue line represents theoretically obtained error and red stars represents estimation.

ACKNOWLEDGMENTS

This work is supported by the Montenegrin Ministry of Science, project grant funded by the World Bank loan: CS-ICT "New ICT Compressive sensing based trends applied to: multimedia, biomedicine and communications".

Corresponding author I. Stanković is with the Faculty of Electrical Engineering, University of Montenegro, on leave at GIPSA Lab, University of Grenoble Alpes.

REFERENCES

[1] B. Boashash, *Time Frequency Signal Analysis and Processing: A Comprehensive Reference*, 2^{nd} edition, Elsevier, 2016.

[2] L. Cohen, *Time-Frequency Analysis - Theory and Applications*, Prentice-Hall, 1995.

[3] V. C. Chen, H. Ling, *Time-frequency transforms for radar imaging and signal analysis*, Artech House, Boston, USA, 2002.

[4] B. Boashash, V. Sucic, "High performance time-frequency distributions for practical applications," *Wavelets and Signal Processing*, Birkhuser Boston, pp. 135–175, 2003.

[5] C. Richard, "Time-frequency-based detection using discrete-time discrete-frequency Wigner distributions," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2170–2176, September 2002.

[6] LJ. Stanković, M. Daković, T. Thayaparan, *Time-Frequency Signal Analysis with Applications*, Artech House, Boston, 2013.

[7] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol.52, no.4, pp. 1289–1306, April 2006.

[8] E. J. Candès, J. Romberg, T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol.52, no.2, pp. 489–509, February 2006.

[9] E. J. Candès, M. B. Wakin, "An Introduction to Compressive Sampling," *IEEE Signal Processing Magazine*, vol.21, no.2, pp. 21–30, March 2008.

[10] P. Flandrin, P. Borgnat, "Time-Frequency Energy Distributions Meet Compressed Sensing," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 2974–2982, June 2010.

[11] D. Needell, J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 301–321, May 2009.

[12] LJ. Stanković, S. Stanković, M. G. Amin, "Missing Samples Analysis in Signals for Applications to L-Estimation and Compressive Sensing," *Signal Processing*, vol.94, pp. 401–408, January 2014.

[13] LJ. Stanković, I. Stanković, M. Daković, "Nonsparsity Influence on the ISAR Recovery from a Reduced Set of Data," *IEEETransactions on Aerospace and Electronic Systems*, vol. 52, no. 6, December 2016.

[14] LJ. Stanković, "On the STFT Inversion Redundancy," *IEEE Transactions on Circuits and Systems II*, vol.63, no.3, pp. 284–288, March 2016.