Convolutive Non-Negative Matrix Factorization for CQT Transform using Itakura-Saito Divergence

Fabio Louvatti do Carmo; Evandro Ottoni Teatini Salles

Abstract— **This paper proposes a modification of the Non-negative Matrix Factorization (NMF) in a single-channel audio source separation problem. NMF is widely used in such problem because of its easy implementation and parts-based separation properties. However, the original NMF uses Short Time Fourier Transform (STFT) as a spectral representation of the data, which has matrix representation, and it does not support data at non-regular grid, such as Constant-Q Transform (CQT). CQT has a strong appeal in audio processing because it approximates the human auditory system in a reasonably way. Besides to usage of CQT as spectral representation, this paper presents a convolutive NMF approach using Itakura-Saito divergence (ISD) to work with irregularly-sampled data, here defined as NRCNMF-IS. The scale invariance property of ISD is interesting for audio applications. The NRCNMF-IS was tested and compared with its matricial version. Utilizing performance metrics, the statistical results show that the use of CQT as spectral representation yields better results than the STFT representation.**

Keywords— **Non-negative Matrix Factorization, Monaural separation, Blind source separation, Constant-Q Transform, Itakura-Saito.**

I. Introduction

The separation task of multiples sounds from monaural audio is a hard problem and it has been widely approached on literature. A commonly used method to such decomposition is the Non-negative Matrix Factorization (NMF). The NMF has gained popularity in [1] due to its sparse nature in the decomposition process (part-based) and its simplicity in implementation. NMF has been used in audio applications as monaural speech separation [2], identification of auditory objects with time-varying spectrum [3] and automatic music transcription [4].

The NMF formulation in [5] defines that a matrix $\mathbf{X}_{N\times M} \in \mathbb{R}^+$ is approximated by the product of two matrices $\mathbf{W}_{N\times R} \in \mathbb{R}^+$ and $\mathbf{H}_{R\times M} \in \mathbb{R}^+$, such that a cost function is minimized. Such a cost function is built from a divergence measure between $\mathbf{X}$ and $\mathbf{WH}$. The most commonly used measures are Frobenius norm, Kullback-Leibler (KL) divergence and Itakura-Saito (IS) divergence.

The vast majority of the methods that use NMF, represents the audio signals through the spectrogram builted from the Short Time Fourier Transform (STFT). Such a representation is linearly spaced in frequency, so it does not efficiently map the frequencies corresponding to the musical notes. In contrast, the Constant-Q Transform (CQT), originally introduced in [6], describes a geometrical resolution in frequency. The CQT is important in speech and music processing, because it is based on human perception of sound. In [7] it is presented a very reliable invertible CQT approach based on non-stationary Gabor frames (CQ-NSGT). However, the CQ-NSGT yields the samples in a non-regular grid, and matrix representation is no longer possible, making it impossible to use with classic NMF.

A reformulation of NMF has been introduced in [8] in order to use it with irregularly-spaced samples, here called Non-Regular NMF (NRNMF). Thereby, it is possible to use NMF with CQ-NSGT for audio applications. The reformulation of the traditional model basically consists of vectoring the matrix $\mathbf{X}_{N\times M} \in \mathbb{R}^+$ as $\mathbf{x}_{K\times 1} \in \mathbb{R}^+$ and creating two other vectors $\mathbf{t}_{K\times 1}$ and $\mathbf{f}_{K\times 1}$ that contain the time and frequency positions of each sample, respectively. The approximation $\hat{\mathbf{x}}_{K\times 1} \in \mathbb{R}^+$ is given by

$$\hat{\mathbf{x}} = \sum_{r=0}^{R-1} \mathbf{v}_r \odot \mathbf{g}_r, \tag{1}$$

where $\mathbf{v}_{K\times R} \in \mathbb{R}^+$ and $\mathbf{g}_{K\times R} \in \mathbb{R}^+$ are the matrices to be found, $\odot$ means element-wise product and R is the number of parts to be factored. It is worth noting that now the values in $\mathbf{t}$ and $\mathbf{f}$ can be real, and not more integers as in the matrix case.

The models in [5] and [8] are related to the classic NMF. However, such traditional approach is weak for audio tasks, because it not considers the temporal information of signal. In [3] and [9] was presented convolutive NMF models for matrix representation, where the relationship between multiples observations in close interval time is described by a spectral bases sequence $\mathbf{W}_t$, which corresponds to the coefficients in $\mathbf{H}$ varying over time. In [10], it was presented a convolutive version of [8] using CQ-NSGT as spectral representation for blind source separation (BSS) application.

The algorithms presented in [8], [9] and [10] use the KL divergence as cost function. It is known that the KL divergence is a special case of β-divergence when $\beta = 1$. In [11], it was shown that the factorization performed with $\beta > 0$ relies more heavily on the largest data values and less precision is to be expected in the estimation of the low-power components. Audio spectra typically exhibit exponential power decrease along frequency and commonly include low-power transient components. In this respect, KL divergence is not very appropriate for audio applications. In addition to the usage of CQ-NSGT as spectral representation of the data, this paper also proposes a modification in [10] using IS divergence as cost function (NRCNMF-IS), whose expression is given by

$$d_{IS}(x|\hat{x}) = \frac{x}{\hat{x}} - \ln\frac{x}{\hat{x}} - 1. \tag{2}$$

The IS divergence is a particular case of β-divergence when $\beta = 0$ and it has interesting properties, such as the scale

Fabio Louvatti do Carmo; Evandro Ottoni Teatini Salles. Federal University of Espirito Santo (UFES), Vitoria-ES, Brazil, E-mails: fabio.carmo@aluno.ufes.br, evandro@ele.ufes.br.

invariance, meaning that low-power components of $\hat{\mathbf{x}}$ bear the same relative importance as high-power ones. Therefore, factorization with IS divergence are relevant to decomposition of audio spectra and also has good perceptual properties of the reconstructed signals.

This paper is organized as follows. The section II presents the NMF algorithms for matrix-structured data and the proposed algorithm for irregularly sampled data utilizing Itakura-Saito as cost function. Section III shows the experimental results, comparing the two algorithms presented in section II, and also presents the methodology and the parameters used. And section IV draws the conclusions.

II. NMF FOR SAMPLES ON A NON-REGULAR GRID

A. Classic and Convoltutive NMF

Using IS divergence as cost function, the classic NMF results in the following iterative process

$$\begin{aligned} \mathbf{H} &\leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T((\mathbf{WH})^{-2}\mathbf{X})}{\mathbf{W}^T(\mathbf{WH})^{-1}} \\ \mathbf{W} &\leftarrow \mathbf{W} \odot \frac{((\mathbf{WH})^{-2}\mathbf{X})\mathbf{H}^T}{(\mathbf{WH})^{-1}\mathbf{H}^T}, \end{aligned} \tag{3}$$

where $\odot$ and $\frac{[.]}{[.]}$ are element-wise multiplication and division, respectively. At each iteration, the matrices $\mathbf{W}$ and $\mathbf{H}$ are normalized in order to obtain an unbiased estimate. The normalization is performed as follows

$$\begin{aligned} \mathbf{w}_N^r &\leftarrow \frac{\mathbf{w}_N^r}{\sum_{n=0}^{N-1} \mathbf{w}_N^r}, \quad \forall r = 0, ..., R-1 \\ \mathbf{h}_M^r &\leftarrow \mathbf{h}_M^r . \sum_{n=0}^{N-1} \mathbf{w}_N^r, \quad \forall r = 0, ..., R-1, \end{aligned} \tag{4}$$

where $\mathbf{w}_N^r$ corresponds to each column of $\mathbf{W}$ and $\mathbf{h}_M^r$ to each row of $\mathbf{H}$. The columns of $\mathbf{W}$ represent the spectral patterns found by the algorithm, and the rows of $\mathbf{H}$ indicate the position and intensity that these patterns are repeated along the spectrum. This approach is weak for audio applications, since such patterns have no temporal dependency between them. Such temporal dependence is added to the generative model as follows

$$\hat{\mathbf{X}} = \sum_{t=0}^{T-1} \mathbf{W}_t . \overset{t\rightarrow}{\mathbf{H}}, \tag{5}$$

where T is the length of each spectral sequence and $\overset{t\rightarrow}{[.]}$ denotes a column shift operator of t steps to the right. With this new generative model, the new multiplicative rules can be seen as a set of T NMF operations in (3) that are combined to the end. Thereby, the algorithm needs to update $T+1$ matrices at each iteration

$$\begin{aligned} \mathbf{H} &\leftarrow \mathbf{H} \odot \frac{\mathbf{W}_t^T \overset{\leftarrow t}{\left[\frac{\mathbf{X}}{\hat{\mathbf{X}}^2}\right]}}{\mathbf{W}_t^T \overset{\leftarrow t}{\left[\frac{\mathbf{1}}{\hat{\mathbf{X}}}\right]}} \\ \mathbf{W}_t &\leftarrow \mathbf{W}_t \odot \frac{\left[\frac{\mathbf{X}}{\hat{\mathbf{X}}^2}\right] \overset{t\rightarrow}{\mathbf{H}}^T}{\left[\frac{\mathbf{1}}{\hat{\mathbf{X}}}\right] \overset{t\rightarrow}{\mathbf{H}}^T} \\ & \forall t \in [0, ..., T-1]. \end{aligned} \tag{6}$$

At each iteration, all $\mathbf{W}_t$ and $\mathbf{H}$ are updated, where $\mathbf{H}$ is calculated as the average of their updates for each $\mathbf{W}_t$. Note that for the case $T = 1$, the algorithm in (6) becomes (3).

B. Convolutive NMF for samples on non-regular grid

The algorithms cited in the previous subsection usually use STFT for matricial representation of the spectrogram in $\mathbf{X}$. However, CQ-NSGT does not allow a matrix structure of the samples, and to use CQ-NSGT as a spectral representation along with NMF, it is necessary to vectorize the classic algorithm. The vectorization of (6) yields

$$\begin{aligned} \hat{\mathbf{x}} &= \sum_{t=0}^{T-1} \sum_{r=0}^{R-1} \mathbf{V}_{r,t} \odot \overset{t\rightarrow}{\mathbf{G}}_r \\ \mathbf{P}_{t,1} &= \frac{\mathbf{V}_t \odot \overset{t\rightarrow}{\mathbf{G}}}{\hat{\mathbf{x}}^2 \cdot \mathbf{1}_{1\times R}} \\ \mathbf{P}_{t,2} &= \frac{\mathbf{V}_t}{\hat{\mathbf{x}} \cdot \mathbf{1}_{1\times R}} \\ \mathbf{P}_{t,3} &= \frac{\overset{t\rightarrow}{\mathbf{G}}}{\hat{\mathbf{x}} \cdot \mathbf{1}_{1\times R}} \\ \mathbf{G} &= \frac{\mathbf{D}_t \cdot (\overset{\leftarrow t}{\mathbf{P}}_{t,1} \odot (\overset{\leftarrow t}{\mathbf{x}} \cdot \mathbf{1}_{1\times R}))}{\mathbf{D}_t \cdot \overset{\leftarrow t}{\mathbf{P}}_{t,2}} \\ \mathbf{V}_t &= \frac{\mathbf{D}_f \cdot (\overset{\leftarrow t}{\mathbf{P}}_{t,1} \odot (\overset{\leftarrow t}{\mathbf{x}} \cdot \mathbf{1}_{1\times R}))}{\mathbf{D}_f \cdot \overset{\leftarrow t}{\mathbf{P}}_{t,3}}, \end{aligned} \tag{7}$$

where the matrices $\mathbf{D}_t$ e $\mathbf{D}_f$ are calculated from the time and frequency coordinates of the samples in $\mathbf{x}$. These matrices indicate the relative positions between all the samples. When using CQ-NSGT as a spectral representation, the frequency bins are geometrically spaced, i.e. it is given a minimum frequency f_{min} and the subsequent frequencies are given by $f_i = f_{\text{min}} \cdot 2^{\frac{i}{B}}$, where B is the number of bins per octave. Therefore, the frequency coordinates in $\mathbf{f}$ are predefined values, and then the matrix $\mathbf{D}_f$ can be calculated as follows

$$\mathbf{D}_f(i,j) = \begin{cases} 1, & \mathbf{f}(i) = \mathbf{f}(j) \\ 0, & \mathbf{f}(i) \neq \mathbf{f}(j) \end{cases} \tag{8}$$

However, the time coordinates are not regular due to the Q-constant characteristic. It is known that Q is the quality of the filter and it is given by the ratio of the central frequency of the filter to its bandwidth. Thereby, the window sizes are different for each frequency bin. The Figure 1 shows an example of how the samples are positioned in time and frequency plane.

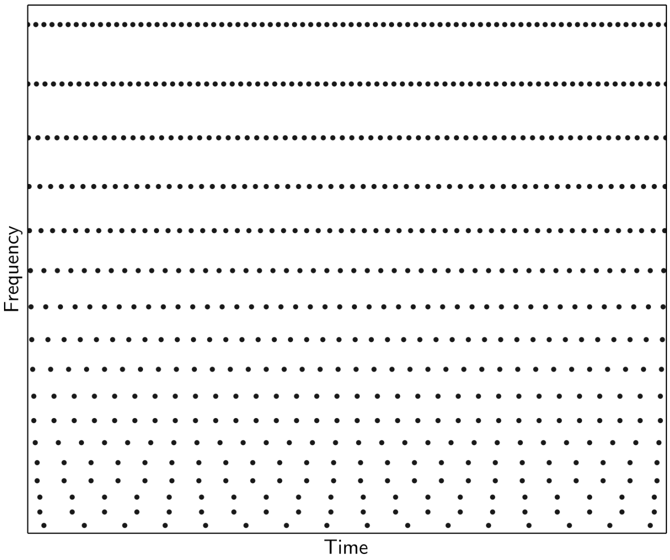

Fig. 1. Example of non-regular grid in the time-frequency plane for the CQT.

Thus, the vector $\mathbf{t}$ has a high chance of being formed by single elements, and if the matrix $\mathbf{D}_t$ is calculated as in (8), $\mathbf{D}_t$ will be a null matrix. To mitigate this problem, in [8] it is calculated the matrix $\mathbf{D}_t$ using a Gaussian kernel.

$$\mathbf{D}_t(i,j) = \exp\left(\frac{-|\mathbf{t}(i) - \mathbf{t}(j)|^2}{\sigma_t^2}\right). \tag{9}$$

However, in the calculation of $\mathbf{D}_t$ in (9), a single value of variance is used. If the chosen variance is too small, the algorithm will not learn any structure in the spectrogram, whereas the case of large variance, spectral blurring will occur. To overcome this problem, the calculation of $\mathbf{D}_t$ is modified and it uses a variance that varies with each bin of frequency. Then, the variance in (9) becomes

$$\sigma_{t,\text{bin}}^2 = \frac{-d_{\text{bin}}^2}{\ln \gamma}, \quad 0 < \gamma < 1, \tag{10}$$

where d_{bin} is the distance between neighboring samples for each frequency bin, and γ is the value of the Gaussian kernel evaluated in the neighboring sample. Figure 2 illustrates the samples and the Gaussian kernels with the variances adapted to each bin of frequency.

The shift operators $\overset{t\rightarrow}{[\cdot]}$ and $\overset{\leftarrow t}{[\cdot]}$ in (7) are performed differently from the matrix version in (6). Figure 1 shows an example of an irregular time-frequency grid, and it may be noted that time-shifting must be conducted so that more high frequency samples are shifted than the low ones, i.e. for a shift of one sample at low frequency, it is necessary to shift several at high frequency so that the spectrogram does not deform. Figure 3 exemplifies the shift operator in irregularly-spaced samples.

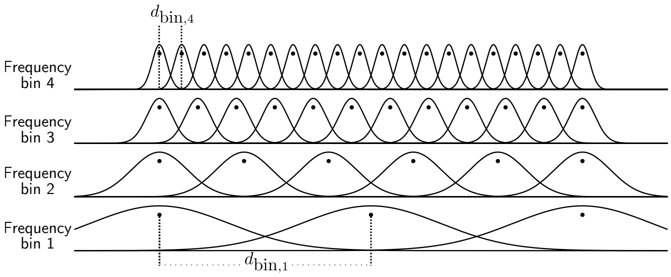

Fig. 2. Example of Gaussian kernel with variance adapting to each frequency bin, calculated with $\gamma = 0.01$.

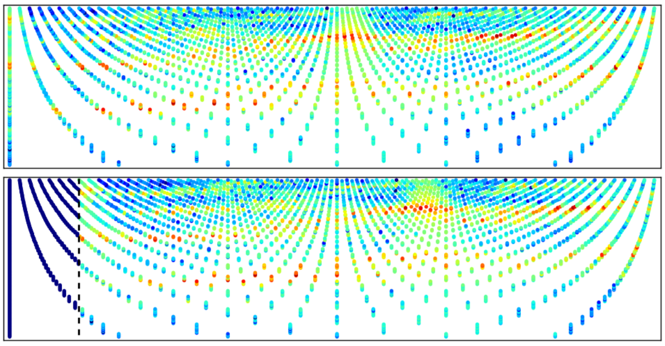

Fig. 3. Example of shift of 0.2 seconds in a CQ-NSGT spectrogram.

III. Experimental Results

A. Musical Application

The NRCNMF-IS is tested and compared to its matricial version, Convolutive Non-negative Matrix Factorization, based on IS-divergence too (CNMF-IS). The NRCNMF-IS uses CQ-NSGT as spectral representation, while CNMF-IS uses the STFT. The task is blind source separation, where the sources are estimated in an unsupervised mode. All the algorithm were developed using the Python programming language.

The audio to be factored is a sequence of notes played on a piano, as the Figure 4 shows. The piano sound was generated from a MIDI file and it was synthesized utilizing the Yamaha C5 Grand Piano SoundFont, with 16kHz sampling rate. The expected result is to separate each note of the piece played.

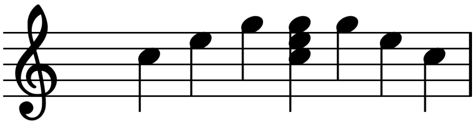

Fig. 4. Sheet music of the audio played by the piano.

B. Performance Evaluation

In this work we use as performance evaluation the measures presented in [12]. The factorization quality is measured in terms of Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR) and Source-to-Artifact Ratio (SAR), which calculate the amount of distortion, interference, and artifacts, respectively. These three measures are standard metrics for blind source separation and they are calculated with a toolbox called BSS Eval[1]. It is important to note that higher values of SDR, SAR and SIR mean better estimates of the sources.

C. Results and Discussion

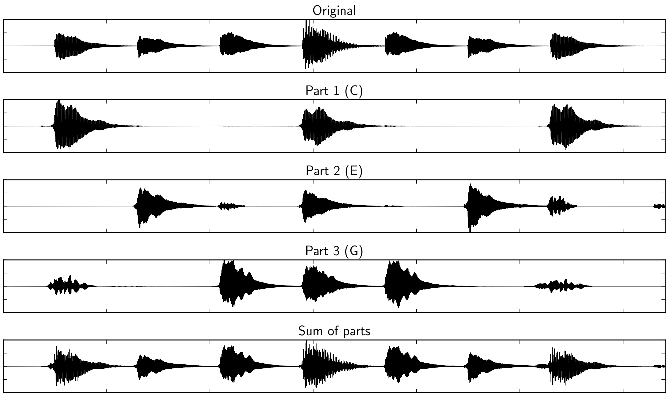

Fig. 5. Time signals resulting from an execution of CNMR-IS.

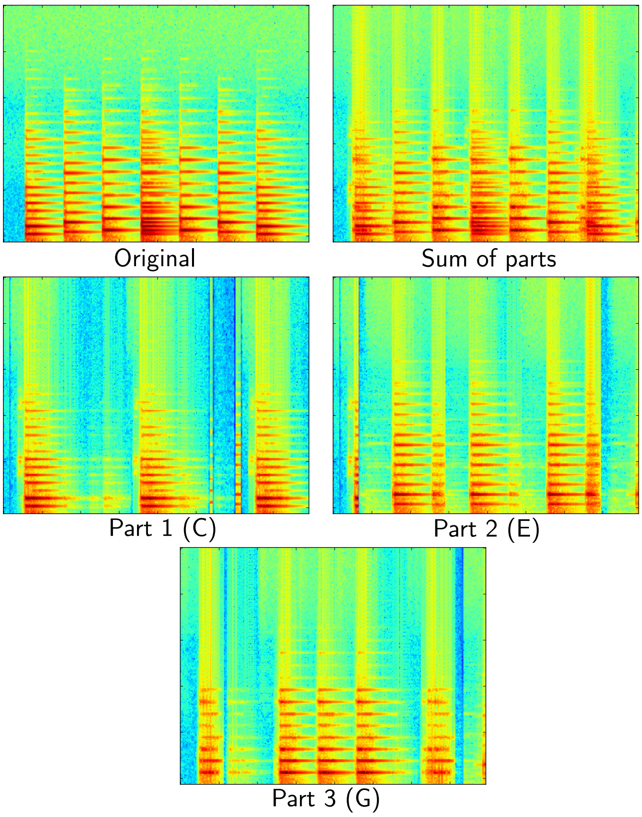

Fig. 6. Spectrograms resulting from an execution of CNMR-IS.

The STFT was calculated using frame size of 512 samples and hop size 128. The time-shifting in CNMF-IS was tested with different lengths in order to find the most suitable. The CNMF-IS algorithm was initialized with random values in all runs. After tests, the best value of time shifting was 50, which is approximately the duration of each note played. The Table I presents the metrics evaluated in an execution of the CNMF-IS. Figures 5 and 6 show the time signals and spectrograms of the execution evaluated in Table I.

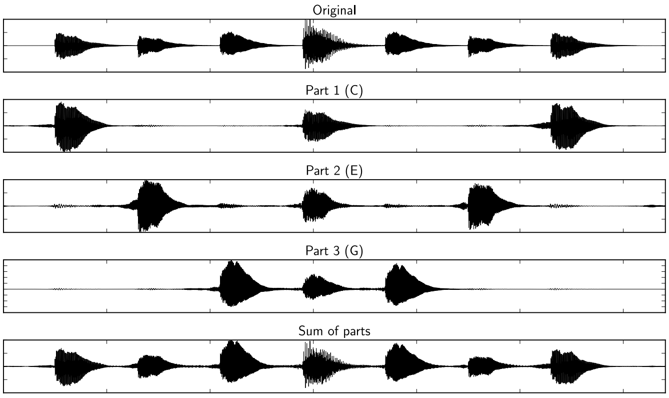

Fig. 7. Time signals resulting from an execution of NRCNMR-IS.

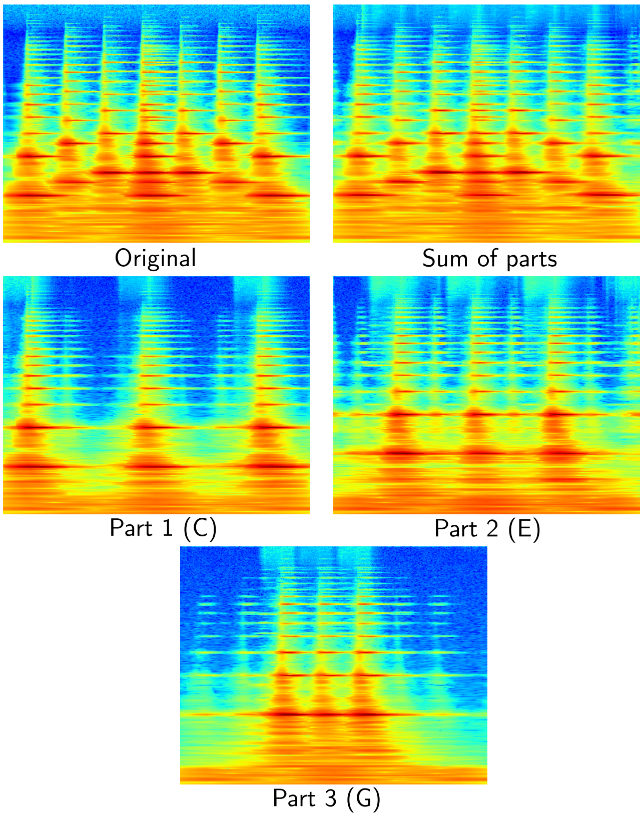

Fig. 8. Interpoled spectrograms resulting from an execution of NRCNMR-IS.

The CQ-NSGT was performed using minimum frequency $f_{\text{min}} = 110\text{Hz}$ (A1), maximum frequency $f_{\text{max}} = 7902.132\text{Hz}$ (B7) and bins per octave bpo $= 48$ (quarter-tones equivalent). Similarly done with CNMF-IS, multiples time-shifting was tested and the best value was 0.3 sec, which is also approximately the duration of each note played. The NRCNMF-IS algorithm was also initialized with random values in all runs. In addition, NRCNMF-IS has one more parameter and it was also tested. It is the parameter γ in (10), and the chosen value was 10^{-3}. The choice of γ is less critical

[1] http://bass-db.gforge.inria.fr/bss_eval/

than the time-shifting value. The time-shifting value is more sensitive because it defines the length of the spectral bases sequence $\mathbf{V}_t$, which will be repeated according to $\mathbf{G}$ along the estimated spectrogram. The Table II presents the metrics evaluated in an execution of the NRCNMF-IS. Figures 7 and 8 show the time signals and spectrograms of the execution evaluated in Table II. As previously mentioned, the CQ-NSGT spectrogram is sampled in an irregular grid as in the Figure 1, however the spectrograms presented in the Figures 7 and 8 were interpolated for better visualization.

The experiments of CNMF-IS and NRCNMF-IS were repeated 53 times and Figure 9 shows the statistics of all executions of both algorithms in terms of mean and standard deviation. The audio results of this article and others results are in the website `https://sites.google.com/site/nrcnmfis/`

Tables I and II show one execution time results of CNMF-IS and NRCNMF-IS respectivelly. In these two specific examples, we can see that CNMF-IS obtained the best results for Source 1 (C), while NRCNMF-IS is better in the others. However, it is important to note that despite this, the result of NRCNMF-IS is smoother in relation to human perception, being a more comfortable sound (It is possible to listen them from the above mentioned website). Although the values of the SAR metric are close in both cases, the artifacts in CNMF-IS are noisier than those present in NRCNMF-IS.

TABLE I

RESULTS OF AN EXECUTION OF THE CNMF-IS ALGORITHM.

	SDR	SIR	SAR
Source 1 (C)	**13.703**	**24.800**	**14.069**
Source 2 (E)	8.072	16.275	8.885
Source 3 (G)	8.582	15.518	9.685

TABLE II

RESULTS OF AN EXECUTION OF THE NRCNMF-IS ALGORITHM.

	SDR	SIR	SAR
Source 1 (C)	12.037	22.987	12.422
Source 2 (E)	**9.409**	**19.698**	**9.883**
Source 3 (G)	**9.546**	**19.558**	**10.050**

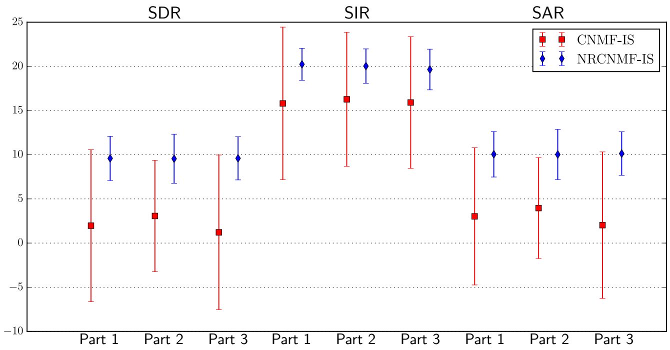

Fig. 9. Performance evaluation of CNMF-IS and NRCNMF-IS using BSS Eval. Statistics in terms of mean and standard deviation.

Figure 9 shows the mean and standard deviation of SDR, SIR and SAR metrics, after 53 experiment repetitions, for CNMF-IS and NRCNMF-IS. The NRCNMF-IS results were better in terms of SDR and SAR. With high SAR values, the audios obtained by NRCNMF-IS contain few artifacts, which explains why the sound is more pleasant. As for the SIR, which measures the interference between the sources, NRCNMF-IS obtained little variability in the results, whereas CNMF-IS sometimes achieved good results and sometimes bad ones. In all metrics, NRCNMF-IS results are less dispersed, i.e. they do not change much in different executions. The NRCNMF-IS results are more consistent since they have small values of standard deviation.

IV. Conclusions

In this paper we presented a Convolutive Non-negative Matrix Factorization approach for irregularly sampled data, using the Itakura-Saito divergence as cost function and Constant-Q Transform as spectral representation of the audio. To the didactic audio example used in this article it was observed that although the CQ-NSGT has an irregularly spaced samples in the time-frequency plane, it is possible to use such grid with the NMF algorithm and to obtain better results than the version using STFT. Future studies might explore a supervised approach in order to develop applications for automatic transcription of scores, for example. And, once NMF suffers from both a high computational cost and an intensive memory usage, it is interesting to develop an iterative online approach to mitigate both problems.

References

[1] D. D. Lee and H. S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[2] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” *Trans. Audio, Speech and Lang. Proc.*, vol. 15, pp. 1–12, Jan. 2007.

[3] P. Smaragdis, *Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs*, ch. Independent Component Analysis and Blind Signal Separation: Fifth International Conference, ICA 2004, Granada, Spain, September 22-24, 2004. Proceedings, pp. 494–499. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[4] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pp. 177–180, Oct 2003.

[5] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *In NIPS*, pp. 556–562, MIT Press, 2001.

[6] J. Brown, “Calculation of a constant q spectral transform,” *Journal of the Acoustical Society of America*, vol. 89, pp. 425–434, January 1991.

[7] T. G. M. Dörfler, N. Holighaus and G. Velasco, “Constructing an invertible constant-q transform with nonstationary gabor frames,” in *In Proceedings of the 14th International Conference on Digital Audio Effects (DAFx 11), Paris, France, 2011*, 2011.

[8] P. Smaragdis and M. Kim, “Non-negative matrix factorization for irregularly-spaced transforms.,” in *WASPAA*, pp. 1–4, IEEE, 2013.

[9] P. D. O’Grady and B. A. Pearlmutter, “Convolutive non-negative matrix factorisation with a sparseness constraint,” in *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pp. 427–432, Sept 2006.

[10] F. L. Carmo and E. O. T. Salles, “Uma abordagem convolutiva para a fatoração de matriz não-negativa em dados amostrados irregularmente,” *XXI CONGRESSO BRASILEIRO DE AUTOMÁTICA - CBA2016, 2016. v. 1. p. 1-6.*, 2016.

[11] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the beta-divergence,” *CoRR*, vol. abs/1010.1763, 2010.

[12] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, July 2006.