SOUND SOURCE SEPARATION BASED ON NON-NEGATIVE TENSOR FACTORIZATION INCORPORATING SPATIAL CUE AS PRIOR KNOWLEDGE

Yuki Mitsufuji

Sony Corporation, Tokyo, Japan

Axel Roebel[1]

IRCAM-CNRS-UPMC UMR 9912, 75004, Paris, France

ABSTRACT

This paper concerns a new method of source separation that uses a spatial cue given by a user or from accompanying images to extract a target sound. The algorithm is based on non-negative tensor factorization (NTF), which decomposes multichannel spectrograms into three matrices. The components of one of the three matrices represent spatial information and are associated with the spatial cue, thus indicating which bins of the spectrogram should be given preference. When a spatial cue is available, this method has a great advantage over conventional PARAFAC-NTF in terms of both computational costs and separation quality, as measured by evaluation metrics such as SDR, SIR and SAR.

***Index Terms*—** Audio source separation, Signal reconstruction, Sparse representation, Nonnegative Tensor Factorization

1. INTRODUCTION

Source separation is a key technology that has enabled major breakthroughs in various fields of audio signal processing, such as automatic speech recognition and music transcription. It can generally be enhanced by incorporating prior knowledge, which is mainly provided as a property of the target source. For example, the incorporation of a spatial cue has the potential to improve applications such as Sound Zoom by enabling the extraction of target sounds that might be located in a certain direction. Below, we discuss source separation with a spatial cue, beginning with a review of previous studies.

The techniques for solving the source separation problem can be classified into two groups: position-given source separation (PGSS) e.g., beam forming [1, 2] and blind source separation (BSS). In the 1990s, a new BSS technique called independent component analysis (ICA) was developed that automatically finds the directions of the sources in a mixture, thereby enabling the extraction of a target source [3, 4]. The main drawback is that the algorithm can extract at most N sources for an N-channel signal. Another approach, called DUET, which is capable of de-mixing from 2ch-stereo signals, requires the sources to be sparse with respect to the spectrogram bins [5, 6].

The early 2000s saw the development of a new BSS technique called non-negative matrix factorization (NMF) [7, 8], which does not rely on directional information. It is based on the idea that a mixture is a composite of a number of object basis elements, each of which represents an underlying characteristic of the sources. Estimation is carried out by simple matrix factorization, with all the elements being non-negative. NMF also eliminates another problem with ICA, namely, that the number of microphones should be greater than or equal to the number of target sources. However, NMF requires the clustering of basis elements after factorization to classify them so that the target sound can be resynthesized. A large number of related techniques have been developed so far [9, 10, 11].

Nevertheless, it is unlikely that NMF can incorporate a spatial cue due to its algorithm framework. Since NMF does not produce any spatial information during the separation process, it is difficult to associate a spatial cue. Sawada et al. proposed a new NMF method that incorporates a spatial matrix; but they did not consider the use of a spatial cue [12].

As a generalized NMF technique, non-negative tensor factorization (NTF) extends the NMF idea to tensors. An n-way tensor is a generalization of the mathematical concepts of scalar, vector, and matrix (e.g., a two-way tensor is a matrix). Specifically, a three-way tensor, which can be regarded as a collection of multichannel spectrograms, is now being investigated for use in NTF [13, 14, 15]. Extension to the third dimension provides another matrix that describes the energy distribution of each base component on every channel, which can also be regarded as spatial information. This technique enables the NMF approach to be adapted to PGSS, which, as a result, can easily accept a spatial cue. In addition, since a spatial cue indicates which bins of the tensor spectrogram are important, it is possible to improve the quality of an approximation to the specific bins of the tensor by giving more weight to bins where the target is likely to exist and less weight to the others. To our knowledge, no one has yet used NMF for PGSS. This paper describes a new NTF method that incorporates a spatial cue. The results show that this method is advantageous in terms of separation quality and computa-

[1] This research has been funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287674 (3DTVS).

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

tional costs over conventional PARAFAC-NTF [16]. NTF is also capable of separating two sounds coming from the same direction, which is a great improvement over previous source separation methods. However, this paper focuses only on the incorporation of a spatial cue. Adaptation to a moving target is also outside the scope of this study.

This paper is organized as follows: Section 2 briefly explains NTF. Section 3 describes a possible incorporation of spatial cue into NTF. Section 4 shows evaluation results on quality and computational costs. Finally, Section 5 presents some concluding remarks.

2. NON-NEGATIVE TENSOR FACTORIZATION

A multichannel audio signal that has been transformed into a set of spectrograms (one for each channel) will be regarded as a three-way tensor V and approximated by $\widehat{\mathrm{V}}$. $\widehat{\mathrm{V}}$ is created as a superposition of P feature tensors, each produced by means of an outer product of three vectors q_p, w_p, and h_p, respectively representing the channel, frequency, and time factors of the feature tensor. To adapt the NTF representation $\widehat{\mathrm{V}}$ to the target tensor spectrogram V the following optimisation problem is solved:

$$\min_{\mathrm{Q,W,H}} \sum_{j,k,l} g_{jkl} d_\beta(v_{jkl}|\hat{v}_{jkl}) + \alpha(\mathrm{H}) \text{ s.t. } \mathrm{Q,W,H} \geq 0 \ , \quad (1)$$

with

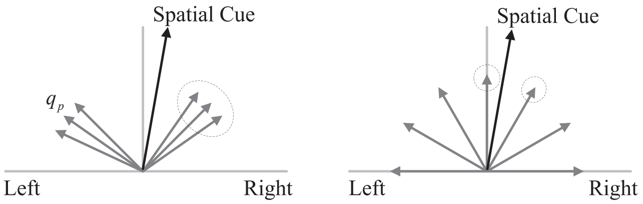
$$\hat{v}_{jkl} = \sum_p q_{jp} w_{kp} h_{lp} \ .$$

Here the matrices Q, W, and H are assembled from the vectors q_p, w_p, and h_p, having elements q_{jp}, w_{kp}, and h_{lp}. The elements of the tensor $\widehat{\mathrm{V}}$ are denoted as v_{jkl}. $\alpha(\mathrm{H})$ represents additional constraints on matrix H, which are taken into account during minimisation of the cost function. The $\beta-$divergence, d_β, is suitable for NTF, allowing the separation quality to be changed, subject to the parameter β [17]. When β equals to 2, 1, or 0, the NTFs are called EUC-NTF, KL-NTF, or IS-NTF, respectively. g_{jkl} denotes one of the bins of the weighting tensor, G, in bin-wise $\beta-$divergence. It allows controlling the impact of the error observed in the different elements of V. For standard PARAFAC-NTF, $g_{jkl} = 1$ for all the bins.

The update rules for training the three matrices are derived from the derivatives of the cost function:

$$\mathrm{Q} \leftarrow \mathrm{Q} \cdot \left(\frac{\langle \mathrm{G} \cdot \mathrm{V} \cdot \widehat{\mathrm{V}}^{\cdot(\beta-2)}, \mathrm{W} \circ \mathrm{H} \rangle_{\{2,3\},\{1,2\}}}{\langle \mathrm{G} \cdot \widehat{\mathrm{V}}^{\cdot(\beta-1)}, \mathrm{W} \circ \mathrm{H} \rangle_{\{2,3\},\{1,2\}}} \right)^{\cdot\gamma} , \quad (2)$$

$$\mathrm{W} \leftarrow \mathrm{W} \cdot \left(\frac{\langle \mathrm{G} \cdot \mathrm{V} \cdot \widehat{\mathrm{V}}^{\cdot(\beta-2)}, \mathrm{Q} \circ \mathrm{H} \rangle_{\{1,3\},\{1,2\}}}{\langle \mathrm{G} \cdot \widehat{\mathrm{V}}^{\cdot(\beta-1)}, \mathrm{Q} \circ \mathrm{H} \rangle_{\{1,3\},\{1,2\}}} \right)^{\cdot\gamma} , \quad (3)$$



[idea1] PARAFAC-NTF

[idea2] fixed-Q NTF

Fig. 1. Two different solutions incorporating NTF (see section 3).

$$\mathrm{H} \leftarrow \mathrm{H} \cdot \left(\frac{\langle \mathrm{G} \cdot \mathrm{V} \cdot \widehat{\mathrm{V}}^{\cdot(\beta-2)}, \mathrm{Q} \circ \mathrm{W} \rangle_{\{1,2\},\{1,2\}} + \nabla^-_{\mathrm{H}}\alpha(\mathrm{H})}{\langle \mathrm{G} \cdot \widehat{\mathrm{V}}^{\cdot(\beta-1)}, \mathrm{Q} \circ \mathrm{W} \rangle_{\{1,2\},\{1,2\}} + \nabla^+_{\mathrm{H}}\alpha(\mathrm{H})} \right)^{\cdot\gamma} , \quad (4)$$

where $\nabla_{\mathrm{H}}\alpha(\mathrm{H}) = \nabla^+_{\mathrm{H}}\alpha(\mathrm{H}) - \nabla^-_{\mathrm{H}}\alpha(\mathrm{H})$, both $\cdot$ and / denote element-wise calculations, $\mathrm{A} \circ \mathrm{B}$ denotes $J \times K \times P$ tensor with elements $a_{jp}b_{kp}$ when A and B are $J \times P$ and $K \times P$ [18], and $\langle \mathrm{A}, \mathrm{B} \rangle_{\{C\},\{D\}}$ denotes a contracted product [16]. Setting parameter γ to the proper value guarantees that the cost function decreases monotonically when $g_{jkl} = 1$ for all the bins and the constraints are zero [19].

3. INCORPORATION OF SPATIAL CUES

We devised two ways of incorporating a spatial cue. Fig.1 shows a 2D representation of a channel matrix, Q, for 2ch-stereo signals. The small arrows represent the basis elements of the channel matrix. Their positions depend on the values for each channel: for example, the basis element $q_p = [0.5, 0.5]^T$ means that the source is coming from the center, and the basis element $q_p = [0.9, 0.1]^T$ means that the source is closer to the left channel.

$$\overleftarrow{Q}_p = 2 \tan^{-1} \left(\frac{q_{1,p}}{q_{0,p}} \right), \quad (5)$$

where $\overleftarrow{Q}$ denotes the angles of the arrows in radians clockwise from the horizontal axis in Fig.1. The big arrow indicates a spatial cue that is specified independently from outside. It is totally independent of the positions of the small arrows at this moment, and it is given in the same format as the elements of Q, (e.g., $sc = [0.4, 0.6]_T$).

Idea 1 (Fig.1, left) applies standard PARAFAC-NTF to audio signals. Factorization produces the channel matrix, Q, the elements of which will be linked to the spatial cue at the end of the process. This may, however, pose a problem when the spatial cue is far from the basis element candidates (Fig.1, left). Interpolation between the two groups (three arrows on left and three arrows on right) in another space might not be helpful in creating sound in the direction of the spatial cue.

However, Idea 1 yields good performance when the spatial cue and the basis element candidates are sufficiently close to each other.

In Idea 2 (Fig.1, right), the basis elements of the channel matrix, Q, are evenly spaced before the start of the NTF process. The directions remain fixed throughout the process while matrix W and matrix H are trained by means of NTF update rules. The basis element candidates are selected to be the ones closest to the spatial cue. Assuming that the basis elements cover the whole sound field, this technique yields a more robust approximation for any spatial cue. Two things make Idea 2 worth focusing on: the computational efficiency, since the channel matrix, Q, does not have to be updated; and the potential improvement in quality due to the prior knowledge provided by the spatial cue. We call Idea 1 PARAFAC-NTF (p-NTF) and Idea 2 fixed-Q NTF (f-NTF). The next section explains a method based on f-NTF.

4. PROPOSED METHOD

4.1. Initialization

We selected IS-NTF (NTF using Itakura-Saito divergence) for our initial experiments and used noise-free input signals, such as commercial music. In spite of the better quality it provides, the main concern with IS-NTF is instability arising from the initialization of matrices [20]. IS-NTF is more prone to local minima than other types of NTF because the formula has convex and concave parts. The simplest solution to this problem might be to perform a number of training runs and then select the best results from among them. Another approach to mitigating this effect is tempering NTF by changing the type of divergence during the iterations [21]. For example, the training could start with EUC-NTF, which is relatively robust with regard to local minima, and finish up with IS-NTF, which produces better results. This would require that developers carefully control β, and more iterations than usual would probably be needed. We came up with a two-part solution to this problem. One part involves taking advantage of the prior knowledge provided by the spatial cue, which is discussed in the next section. The other part is based on the feature similar to the idea of $\overleftarrow{Q}$. This similar feature is easily obtained from the equation in the case of 2ch-stereo signals.

$$\overleftarrow{B}_{kl} = 2\tan^{-1}\left(\frac{v_{1kl}}{v_{0kl}}\right), \tag{6}$$

where v_{0kl} and v_{1kl} are the spectrogram bins for the left and right channels, respectively. This feature concerns the arrows in Fig.1 and their relationships to each spectrogram bin. It is possible to determine the locations of sources with respect to the bins by searching for the peaks in the histogram of $\overleftarrow{B}$, which represents the dominant presence of the sources. The basis elements are preferentially allocated, based on the histogram of $\overleftarrow{B}$: More elements are allocated to directions where sources are likely to exist, although some are allocated to cover all directions. The number of directions, D, and the range of arrows should be selected depending on the application.

4.2. Weighting function

Since the spatial cue indicates which direction should be given preference, and since the histogram of $\overleftarrow{B}$ indicates which source is dominant for a given direction, it is possible to approximate the spectrogram bin associated with the spatial cue more precisely than other bins. This is easy to accomplish by using the proper weighting tensor, G, in the cost function:

$$g_{jkl} = \exp\bigl(-\psi|(d_{sc} - d)|\bigr) \quad c, k, l \in \mathit{Grp}(d), \tag{7}$$

where d_{sc} is the direction index associated with the spatial cue, d ranges from 0 to $D-1$, *Grp* is the group of spectrogram bins obtained from the histogram of $\overleftarrow{B}$, and ψ determines the shape of the exponential function.

4.3. Constraints

This step takes advantage of the normalization procedure that is often done through NTF for the purpose of concentrating energy to single matrix. After matrices Q and W have been normalized by their own energies, which have then been multiplied by H, the energy for each direction can be estimated by adding all the basis elements in matrix H over time. The estimated energy is fixed so that it can be used as a reference to constrain the training in the initial stages of the iteration. This should reduce the likelihood of being trapped in local minima. Here, we again use the Itakura-Saito divergence to measure distance. The constraint on energy in a given direction is

$$\alpha(\mathrm{H}) = \mu \sum_{d=0}^{D-1} d_{IS}\left(\sum_{p\in P_d} |\mathrm{H}_p^{ini}|_1 \middle| \sum_{p\in P_d} |\mathrm{H}_p|_1\right), \tag{8}$$

s.t. $\mathrm{Q}_p = \mathrm{Q}_p/|\mathrm{Q}_p|_1,\ \mathrm{W}_p = \mathrm{W}_p/|\mathrm{W}_p|_1,\ \mathrm{H}_p = |\mathrm{Q}_p|_1|\mathrm{W}_p|_1\mathrm{H}_p,$

where $|\cdot|_1$ denotes the L1-norm, H^{ini} denotes the matrix H that was estimated in the initialization step, and P_d denotes the set of bases allocated to the direction with the index d. For IS-NTF, the following should hold for the derivative of the constraint:

$$\nabla\alpha(h_p) = \frac{\mu}{\left(\sum_{p\in P_d} |\mathrm{H}_p|_1\right)^{-1}} - \frac{\mu \sum_{p\in P_d} |\mathrm{H}_p^{ini}|_1}{\left(\sum_{p\in P_d} |\mathrm{H}_p|_1\right)^{-2}}. \tag{9}$$

We call this type of NTF a spatial-cue NTF (sc-NTF).

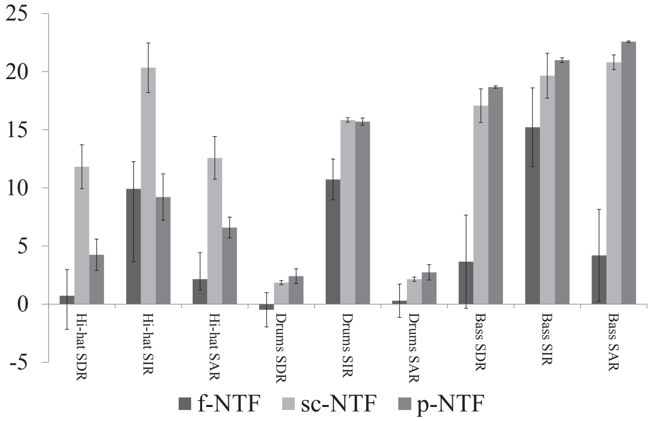


Fig. 2. Dataset wdrums.

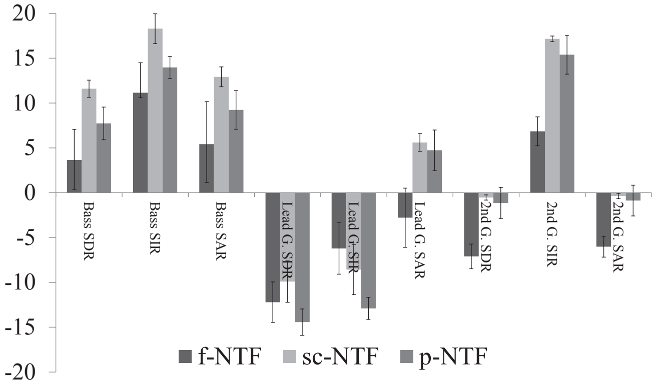


Fig. 3. Dataset nodrums.

5. EVALUATION

BSS Eval of MATLAB was used to evaluate the above method. It gives three standard metrics for source separation: signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) [22]. SDR is a global measure of the quality of source separation that encompasses the two other metrics; SIR indicates how well the target source is separated from interference; and SAR indicates how well the target source retains sound quality after separation. sc-NTF was compared with p-NTF and f-NTF. 2ch-stereo signals were obtained from the "Signal Separation Evaluation Campaign" Web site (SiSEC 2008 [23]). More specifically, we used development data from the underdetermined speech and music mixture task. These 2ch-stereo sources contain a number of instruments placed independently in a 2D field. The ground truth registered in a similar format as $\overleftarrow{Q}$, was also obtained from SiSEC 2008. It gives the location of each instrument.

For f-NTF, after separation is conducted, the basis elements pointing in a direction close to the ground truth are selected to be separated out. In contrast, p-NTF requires grouping after training. Our experiments on two grouping algorithms k-means and k-nearest neighbor to the ground truth showed that the latter yielded better results.

	f-NTF	sc-NTF	p-NTF
computational costs [s]	20.869	23.427	94.399

Table 1. Runtime test for the three different NTF.

On the other hand, for sc-NTF the bases are selected beforehand due to the link established between the allocated basis elements and the spatial cue. Resynthesis is followed by Wiener filtering to create the output signals used for evaluation (1024-point FFT, half overlap, the number of the basis elements $P = 90$). Tests were run 10 times to obtain an average, indicated by a bar, and 95% confidence, indicated by a line on top of the bar. This test was almost exactly the same as the one described by Févotte et al. in their 2010 paper [18]. The only difference was in the number of bases: They used $P = 9$ and we used $P = 90$. γ in the cost function was set to 1. We obtained better results when $\psi = 0.2$ for weighting tensor G and $\mu = 300$ for the constraint.

Fig.2 shows test results for percussive sound (wdrums), and Fig.3 shows results for harmonic sound (nodrums). Most of the results show that sc-NTF outperforms both f-NTF and p-NTF. It is important to note that these results were obtained by sacrificing accuracy of the approximation of sources far from the spatial cue. This can be deduced from the final value of the cost function with respect to direction, but cannot be discussed further due to lack of space. The 95% confidence for both p-NTF and sc-NTF indicates that local minima were avoided. There is a large variance only in the results for f-NTF. For the sake of the omission of updating channel matrix Q, the runtime test shows great advantages for sc-NTF seen in Table.1, which was calculated with Matlab code.

6. CONCLUSION

We developed a new method of extracting a target sound from a mixture that employs a spatial cue. Two ways of incorporating a spatial cue into a NTF framework were devised. The one that employs a fixed channel matrix, Q, was further developed to improve the separation quality. The association of the spatial cue with the histogram of $\overleftarrow{B}$ clarifies which spectrogram bins should be given preference to obtain a better approximation. An evaluation of separation quality, which was carried out as in previous studies, demonstrated the effectiveness of the weighting tensor, G, and the energy constraints. In addition, the omission of the calculation of Q was a great advantage in the runtime test. In short, our algorithm combines the computational cost of f-NTF with the separation quality of p-NTF. Adaptation to a moving target and source separation from the same direction will be subjects of future work.

7. REFERENCES

[1] L. J. Griffiths, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Antennas and Propagation Society*, vol. 30, no. 1, pp. 22–34, 1982.

[2] H. Cox, R. M. Zeskind, and M. M. Owen, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, no. 10, pp. 22–34, 1987.

[3] A. Hyvarinen, "Survey on independent component analysis," in *Neural Computing Surveys 2*, 1999, pp. 94–128.

[4] K. Torkkola, "Blind separation for audio signals - are we there yet?," in *Workshop on Independent Component Analysis and Blind Signal Separation*, Aussois, France, 1999.

[5] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. of the Intern. Conf. on Digital Audio Effects (DAFx-03)*, London, UK, 2003, pp. 209–213.

[6] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals : Demixing n sources from 2 mixture," in *ICASSP*. 2000, vol. 5, pp. 2985–2988, IEEE.

[7] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," 2003.

[8] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *the International Computer Music Conference (ICMC)*, 2003.

[9] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, September 1-4, 2009.

[10] Jaiswal R., Fitzgerald D., Barry D., E. Coyle, and S. Rickard, "Clustering NMF basis functions using shifted NMF for monaural sound source separation," in *ICASSP*, 2011, pp. 245–248.

[11] J. M. Becker, M. Spiertz, and V. Gnann, "A propability-based combination method for unsupervised clustering with application to blind source separation," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2012)*, TelAviv, Israel, September 1-4, 2012, pp. 99–106.

[12] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of itakura-saito nonnegative matrix factorization," in *ICASSP*. 2012, pp. 261–264, IEEE.

[13] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *ICML*. 2005, vol. 119, pp. 792–799, ACM.

[14] A. Ozerov, C. Févotte, R. Blouet, and J. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *ICASSP*, 2011, pp. 257–260.

[15] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, 2008.

[16] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 2009.

[17] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," *Proc. of the Irish Signals and Systems Conf. (ISCC)*, 2008.

[18] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *CMMR*. 2010, vol. 6684, pp. 102–115, Springer.

[19] M. Nakano, H. Kameoka, J. Le Roux, Yu. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2010, pp. 283 –288.

[20] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *ICASSP*. 2012, pp. 129–132, IEEE.

[21] N. Bertin, C. Févotte, and R. Badeau, "A tempering approach for itakura-saito non-negative matrix factorization. with application to music transcription," in *ICASSP*. 2009, pp. 1545–1548, IEEE.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] "In signal separation evaluation campaign (sisec 2008):http://www.sisec.wiki.irisa.fr," .