2009-01-01

On the use of the Beta Divergence for Musical Source Separation

Derry Fitzgerald
Technological University Dublin, derry.fitzgerald@tudublin.ie

Matt Cranitch
Cork Institute of Technology, matt.cranitch@cit.ie

Eugene Coyle
Technological University Dublin, Eugene.Coyle@tudublin.ie

Funder: Enterprise Ireland

Audio Research Group

Articles

Dublin Institute of Technology *Year* 2009

On the use of the Beta Divergence for Musical Source Separation

Derry Fitzgerald* Matt Cranitch†

Eugene Coyle‡

*Dublin Institute of Technology, derry.fitzgerald@dit.ie

†Cork Institute of Technology, matt.cranitch@cit.ie

‡Dublin Institute of Technology, Eugene.Coyle@dit.ie

— Use Licence —

Attribution-NonCommercial-ShareAlike 1.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution.
  You must give the original author credit.
- Non-Commercial.
  You may not use this work for commercial purposes.
- Share Alike.
  If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

For any reuse or distribution, you must make clear to others the license terms of this work. Any of these conditions can be waived if you get permission from the author.

Your fair use and other rights are in no way affected by the above.

On the use of the Beta Divergence for Musical Source Separation

Derry FitzGerald[†] Matt Cranitch[*] and Eugene Coyle[†]

Audio Research Group
Dublin Institute of Technology
Kevin St, Dublin

Dept. of Electronic Engineering
Cork Institute of Technology
Rossa Avenue, Cork

E-mail: [†]derry.fitzgerald@dit.ie [*]matt.cranitch@cit.ie

Abstract — **Non-negative Tensor Factorisation based methods have found use in the context of musical sound source separation. These techniques require the use of a suitable cost function to determine the optimal factorisation, and most work has focused on the use of the generalised Kullback-Liebler divergence, and more recently the Itakura-Saito divergence. These divergences can be regarded as limiting cases of the parameterised Beta divergence. This paper looks at the use of the Beta Divergence in the context of musical source separation with a view to determining an optimal value of Beta for this problem. This is considered for both magnitude and power spectrograms. In an effort to avoid potential local minima in the Beta divergence, the use of a "tempered" Beta Divergence is also explored.**

Keywords — **Non-negative Tensor Factorisation, Sound Source Separation, Beta Divergence**

I Introduction

Much research has been carried out on the use of non-negative matrix factorisation (NMF) and non-negative tensor factorisation (NTF) for the purposes of musical sound source separation [1, 2]. The majority of this work has focused on the use of the generalised Kullback-Liebler divergence (KLD) as a cost function as it has been found to work reliably for sound source separation. This divergence is defined as:

$$D_{KL}\left(\mathcal{X}\|\hat{\mathcal{X}}\right) = \sum\left(\mathcal{X}log\frac{\mathcal{X}}{\hat{\mathcal{X}}} - \mathcal{X} + \hat{\mathcal{X}}\right) \quad (1)$$

where $\mathcal{X}$ is a tensor containing the original data, $\hat{\mathcal{X}}$ is an estimate of the original data, and summation takes place over all dimensions of the tensors.

More recently it has been suggested that the Itakura-Saito divergence (ISD) may be a more useful divergence for performing NMF and NTF [3]. This is due to the fact it is scale invariant and so low energy components have the same relative importance as high energy components. Further, the ISD can be considered a statistical model of superimposed gaussian components when dealing with power spectrograms. ISD is defined as:

$$D_{IS}\left(\mathcal{X}\|\hat{\mathcal{X}}\right) = \sum\left(\frac{\mathcal{X}}{\hat{\mathcal{X}}} - log\frac{\mathcal{X}}{\hat{\mathcal{X}}} - 1\right) \quad (2)$$

Both of the above divergences can be considered as limiting cases of a more general parameterisable divergence, known as the Beta Divergence. This divergence was first proposed for use with non-negative matrix factorisation (NMF) techniques by Kompass[4] and later by Cichocki et al [5].

$$D_B\left(\mathcal{X}\|\hat{\mathcal{X}},\beta\right) = \sum\left(\mathcal{X}\frac{\mathcal{X}^{\beta-1} - \hat{\mathcal{X}}^{\beta-1}}{\beta\left(\beta-1\right)} + \hat{\mathcal{X}}^{\beta-1}\frac{\hat{\mathcal{X}} - \mathcal{X}}{\beta}\right) \quad (3)$$

For $\beta = 2$ the squared Euclidean distance is obtained,as $\beta \to 1$ the divergence tends to KLD, and so for $\beta = 1$ we define the Beta divergence as KLD. Similarly, for $\beta \to 0$ ISD is obtained, and for $\beta = 0$ we define the divergence as ISD.

The optimal choice of β depends on the statistics of the data being investigated, and experi-

ments have been performed on determining the optimal parameter of β for speech signals when using standard NMF and convolutive NMF[6]. However, the Beta Divergence has yet to be evaluated for musical signals, particularly in the context of the extended non-negative tensor factorisation models proposed for musical signals proposed in [2, 7].

II Musical Source Separation Model

The musical source separation algorithm used in this paper is a harmonicity enforcing additive synthesis based model, where each instrument or source is modelled by a set of harmonic weights [2]. These weights are invariant to pitch and so each note played by an instrument will use the same weights regardless of pitch. This is a simplification of the actual situation where the timbre of the instrument will vary with pitch and so a source-filter model was also incorporated to allow the timbre of instruments to change with pitch, resulting in improved separations. The evolution of the instrument timbres with time is modelled by incorporating shift-invariance in time. A linear mixing model is also assumed.

To further improve the separations, the parameters of each source, such as the number of harmonics and number of shifts in time can be set independently, offering considerable flexibility. The utility of this can be seen in that modelling a flute will typically require less harmonics than a piano or violin and so the number of harmonics can be adjusted accordingly if a flute is known to be present. Further, if knowledge of the pitch range of the instruments is known as is the case in score-assisted separation [8], then the pitch range of the individual instruments can be set according to this information. This has the effect of reducing the possibility of errors in the separations. This generalised model was first presented in [7], but no numerical results on the effectiveness of the generalised model were presented at that time.

In the following, $\langle \mathcal{AB} \rangle_{\{a,b\}}$ denotes contracted tensor multiplication of $\mathcal{A}$ and $\mathcal{B}$ along the dimensions a and b of $\mathcal{A}$ and $\mathcal{B}$ respectively. Outer product multiplication is denoted by $\circ$. Indexing of elements within a tensor is notated by $\mathcal{A}(i,j)$ as opposed to using subscripts. This notation follows the conventions used in the Tensor Toolbox for Matlab, which was used to implement the following algorithm [9]. For ease of notation, as all tensors in the model are instrument-specific, the subscript k is implicit in all tensors. Elementwise multiplication is denoted by $\otimes$ and all division is taken as elementwise.

For an r-channel mixture, individual channel magnitude spectrograms are obtained and combined in a single tensor $\mathcal{X}$, of size $r \times n \times m$ tensor where n is the number of frequency bins and m is the number of time frames. The tensor is then modelled as:

$$\mathcal{X} \approx \hat{\mathcal{X}} = \sum_{k=1}^{K} \mathcal{G} \circ \langle \langle \mathcal{RW} \rangle_{\{3,1\}} \langle \mathcal{SP} \rangle_{\{2,1\}} \rangle_{\{2:3,1:2\}} \tag{4}$$

with $\mathcal{R} = \langle \mathcal{FH} \rangle_{\{2,1\}}$ and K denotes the number of pitched instruments.

$\mathcal{G}$ is a tensor of size r, containing the gains of a given pitched instrument in each channel. $\mathcal{F}$ is of size $n \times n$, where the diagonal elements contain a filter which attempts to model the formant structure of an instrument, thus allowing the timbre of the instrument to alter with frequency. $\mathcal{H}$ is a tensor of size $n \times z_k \times h_k$ where z_k and h_k are respectively the number of allowable notes and the number of harmonics used to model the kth instrument, and where $\mathcal{H}(:,i,j)$ contains the frequency spectrum of a sinusoid with frequency equal to the jth harmonic of the ith note. $\mathcal{W}$ is a tensor of size $h_k \times p_k$ containing the harmonic weights for each of the p_k shifts in time that describe the kth instrument. $\mathcal{S}$ is a tensor of size $z_k \times m$ which contains the activations of the z_k notes associated with the kth source, and in effect contains a transcription of the notes played by the instrument. $\mathcal{P}$ is a translation tensor of size $m \times p_k \times m$, which translates the activations in $\mathcal{S}$ across time, thereby allowing the model to capture temporal evolution of the harmonic weights.

Multiplicative update equations can then be derived for each of the free variables in the model in a manner similar to that described in [10]. Defining

$$\mathcal{D} = \frac{\mathcal{X}}{\hat{\mathcal{X}}^{2-\beta}} \tag{5}$$

and

$$\mathcal{O} = \hat{\mathcal{X}}^{\beta-1} \tag{6}$$

the update equations, which have not been previously presented in published work, for the model parameters are as follows:

$$\mathcal{G} = \mathcal{G} \otimes \frac{\langle \langle \mathcal{D} \langle \mathcal{RW} \rangle_{\{3,1\}} \rangle_{\{2,1\}} \langle \mathcal{SP} \rangle_{\{2,1\}} \rangle_{2:4,[3,1,2]\}}}{\langle \langle \mathcal{O} \langle \mathcal{RW} \rangle_{\{3,1\}} \rangle_{\{2,1\}} \langle \mathcal{SP} \rangle_{\{2,1\}} \rangle_{2:4,[3,1,2]\}}} \tag{7}$$

$$\mathcal{F} = \mathcal{F} \otimes \frac{\langle \langle \mathcal{GD} \rangle_{\{1,1\}} \langle \langle \mathcal{TW} \rangle_{\{3,1\}} \langle \mathcal{SP} \rangle_{\{2,1\}} \rangle_{\{2:3,1:2\}} \rangle_{\{2,2\}}}{\langle \langle \mathcal{GO} \rangle_{\{1,1\}} \langle \langle \mathcal{TW} \rangle_{\{3,1\}} \langle \mathcal{SP} \rangle_{\{2,1\}} \rangle_{\{2:3,1:2\}} \rangle_{\{2,2\}}} \tag{8}$$

$$\mathcal{W} = \mathcal{W} \otimes \frac{\langle \langle (\mathcal{G} \circ \mathcal{R})\, \mathcal{D} \rangle_{\{1:2,1:2\}} \langle \mathcal{SP} \rangle_{\{2,1\}} \rangle_{\{[1,3],[1,3]\}}}{\langle \langle (\mathcal{G} \circ \mathcal{R})\, \mathcal{O} \rangle_{\{1:2,1:2\}} \langle \mathcal{SP} \rangle_{\{2,1\}} \rangle_{\{[1,3],[1,3]\}}} \tag{9}$$

$$\mathcal{S} = \mathcal{S} \otimes \frac{\langle \langle \big(\mathcal{G} \circ \langle \mathcal{RW} \rangle_{\{3,1\}}\big)\, \mathcal{D} \rangle_{\{1:2,1:2\}} \mathcal{P} \rangle_{\{2:3,[2,1]\}}}{\langle \langle \big(\mathcal{G} \circ \langle \mathcal{RW} \rangle_{\{3,1\}}\big)\, \mathcal{O} \rangle_{\{1:2,1:2\}} \mathcal{P} \rangle_{\{2:3,[2,1]\}}} \tag{10}$$

	SDR	SIR	SDR
Standard	8.33	23.20	8.67
Generalised	8.98	24.50	9.28

Table 1: Comparison of performance (in dB) of standard model and generalised model using KLD as cost function

The model parameters are randomly initialised to positive values and the use of multiplicative updates then ensures non-negativity.

III Testing the Beta Divergence

A set of 40 test signals containing mixtures of pitched instruments was used to test the Beta Divergence as a cost function for the separation of musical signals. These test signals contained equal numbers of both stereo and mono mixtures, and full details of the test signals can be found in [2].

Three commonly used separation metrics, namely Signal to Distortion ratio (SDR), Signal to Interference ratio (SIR) and Signal to Artifacts ratio (SAR) were used to measure the separations obtained on these signals for various values of β. SDR attempts to provide an overall measure of the separation quality and takes into account interference from other sources as well as other artifacts due to separation and resynthesis. SIR provides a measure of the presence of other sources in the separated source and SAR provides a measure of artifacts present due to separation and resynthesis. A more detailed description of these metrics, and code to implement these measures can be found at [11],and [12] respectively.

As no results comparing the performance of the generalised model described above with the model described in [2] have previously been published, these are provided in Table 1 to provide a baseline against which to compare the performance of the Beta Divergence, where Standard refers to the model used in [2] and Generalised the model presented in this paper. In both cases KLD was used as a cost function. It can be seen that the generalised model has resulted in improvements of around 0.6 dB for both SDR and SIR, and an improvement of 1.3 dB for SIR, showing that the increased flexibility of the generalised model does offer improved performance over models where the parameters are the same for all instruments.

The value of β was varied from 2 to 0 in steps of 0.1, giving a total of 21 different values of β. It can be seen that this covers the squared Euclidean distance as well as both KLD and ISD. For each of the 40 test signals, the average separation performance was obtained from the individual scores for each source in a given test signal, thereby providing a measure of overall separation for each signal. These were then averaged across all the test signals to provide a measure of the effectiveness of the separations obtained for each value of β.

Further, these tests were ran using both magnitude and power spectrograms as inputs to the model. All previous work on this model had focused on the use of the magnitude spectrogram only, as this had been found to work better with KLD than power spectrograms. However, in light of the use of the Beta Divergence in this paper it was felt that it was necessary to revisit the use of power spectrograms in case some value of β in conjunction with power spectrograms gave better performance than using magnitude spectrograms with the Beta Divergence.

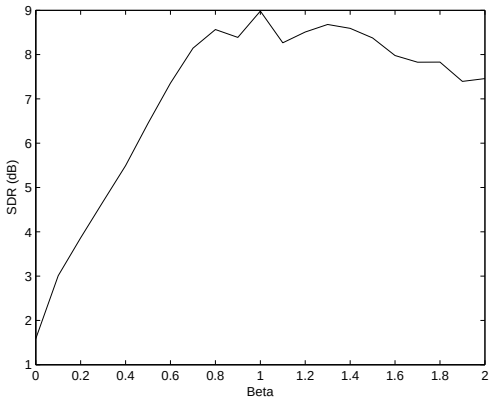

Fig. 1: Average SDR obtained for various values of β (Magnitude Spectrograms)

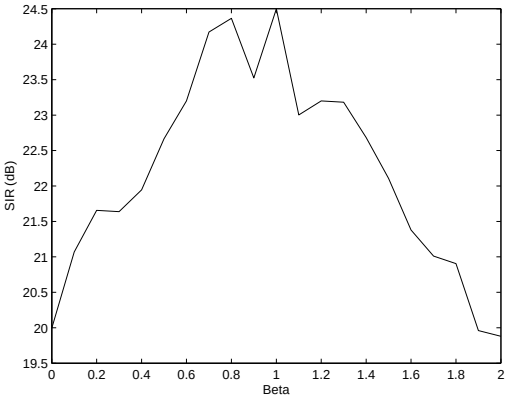

Fig. 2: Average SIR obtained for various values of β (Magnitude Spectrograms)

Figure 1 shows the average SDR obtained for magnitude spectrograms from the test signals for the various values of β. It can be seen that the optimal value of β occurs when $\beta = 1$, with a tailoff in performance as the value of β moves away from 1 in either direction, with a more noticeable decrease in performance as β tends to zero. A similar trend

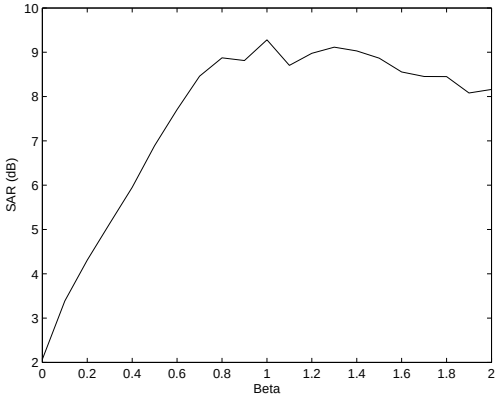

Fig. 3: Average SAR obtained for various values of β (Magnitude Spectrograms)

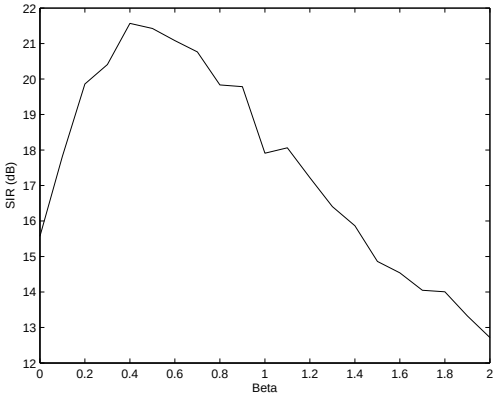

Fig. 5: Average SIR obtained for various values of β (Power Spectrograms)

can be observed for SIR, shown in figure 2, and for SAR, shown in figure 3. For all metrics, it can be seen that using KLD ($\beta = 1$) outperforms all other values of β. This justifies the widespread use of KLD for musical source separation when using magnitude spectrograms to date.

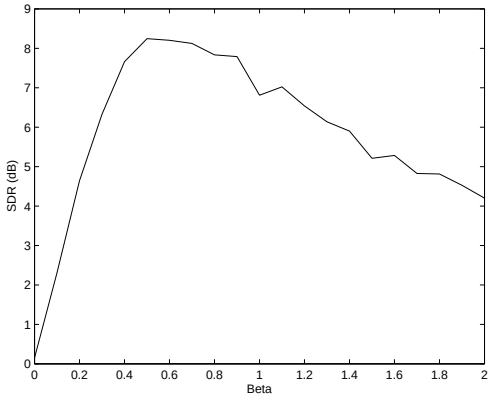

Fig. 4: Average SDR obtained for various values of β (Power Spectrograms)

Figure 4 shows the average SDR obtained using power spectrograms from the test signals. In comparison to magnitude spectrograms where the optimal value was $\beta = 1$, here the optimal value is $\beta = 0.5$. Again there is a notable drop in performance as β tends to zero, with a lesser falloff as β goes towards two. It can also be seen that the maximum SDR value is around 0.75 dB lower than the maximum SDR when using magnitude spectrograms.

In the case of SIR for power spectrograms (shown in figure 5), the optimal value occurs at $\beta = 0.4$, with the maximun SIR being around 3 dB lower than that of the highest value of the magnitude spectrograms. For SAR, (see figure 6) the maximum value occurs at $\beta = 0.6$, with a difference of around 0.5 dB between the largest power spectrogram value and the largest magnitude spectrogram value.

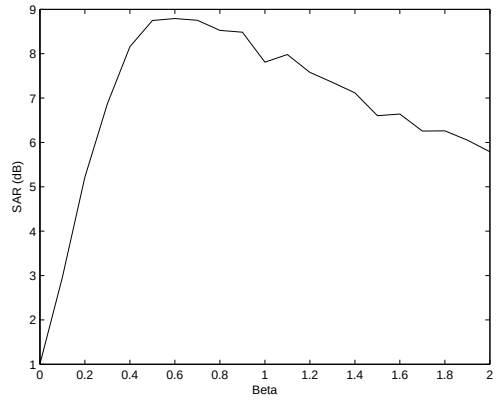

Fig. 6: Average SAR obtained for various values of β (Power Spectrograms)

The above suggests that a value of around $\beta = 0.5$ is optimal for the sound separation model when using power spectrograms. It also shows that with respect to the metrics used KLD with magnitude spectrograms outperforms all other values of β. However, it should be noted that these metrics do not always correspond with human perception of separation. To this end, informal listening tests on the separation quality were performed.

The results of the informal listening tests were broadly in line with the results obtained using the metrics. In general, using a magnitude spectrogram results in better separation performance across a much larger range of β values, whereas with power spectrograms separation was generally quite poor in the range $2 \leq \beta \leq 1.5$. In the region of ± 0.2 of the optimal values given by the metrics, little or no difference could be heard in the quality of the separations regardless of whether magnitude or power spectrograms were used. Finally, as β

got closer to zero, the separation performance degraded. In particular, it was observed that notes played by one source were often stolen by another source, resulting in very noticeable artifacts in the separations.

Overall, both the metrics and informal listening tests show that the performance of magnitude spectrograms with $\beta \approx 1$ and power spectrograms with $\beta \approx 0.5$ are better than separations obtained with other β values. Further, it has been observed that separation performance is effectively equivalent in these cases. This justifies the use of KLD with magnitude spectrograms in previous research using this model.

IV The Tempered Beta Divergence

It has recently been noted that the Beta divergence is convex with respect to the optimisation of any individual parameter in NMF and NTF-based models if $1 \leq \beta \leq 2$, but is non-convex in the range $0 \leq \beta < 1$ [13]. It was suggested that this made the Beta divergence prone to local minima in this range. In particular, with respect to the ISD, it was shown that using a tempered version of the ISD resulted in improved performance in a standard NMF framework. The tempered ISD uses β as a temperature parameter, which is varied over the course of the iterations of the algorithm. Initially β is set in the convex region, typically at $\beta = 2$, and after a set number of iterations at this value, the value of β was then gradually reduced to 0 over a number of iterations before remaining at 0 until convergence.

In light of the improved performance of ISD using this approach it was decided to test a tempered version of the Beta divergence to see if improved performance could be obtained in this manner. To this end, a starting value of $\beta = 2$ was used and final values in the range $0 \leq \beta \leq 1.5$ were investigated. Both power and magnitude spectrograms were again used in testing.

However, it was found that in all cases the use of the Tempered Beta divergence gave considerably poorer performance than using fixed values of β throughout optimisation. This is contrary to the results obtained in [13] when using tempering with ISD in a standard NMF model, and so the use of the tempered Beta Divergence will not be considered when using the separation model contained in this paper in the future.

V Conclusions

The use of the Beta Divergence for musical sound source separation in the context of extended tensor factorisation models has been explored. Both power spectrograms and magnitude spectrograms were used in testing, and it was found that for magnitude spectrograms using the KLD was optimal with respect to the metrics used. For power spectrograms, the Beta Divergence with $\beta \approx 0.5$ was found to perform best, with KLD and magnitude spectrograms outperforming the best value of the power spectrograms. Informal listening tests suggest that both of these cases give similar separation performance, thus justifying the use of KLD in previous research.

Also tested was a tempered version of the Beta Divergence. This was done in an effort to overcome potential local minima in the divergence, but the tempered Beta Divergence was found to give poorer performance in all cases.

The use of both power and magnitude spectrograms, as well as a fractional divergence has also prompted the question as to whether there is an optimal power to raise the magnitude spectrograms by for the purposes of sound source separation. This will be an area for future research.

References

[1] P. Smaragdis, J.C. Brown, "Non-negative Matrix Factorization for Polyphonic Music Transcription", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 177-180, October 2003

[2] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended Nonnegative Tensor Factorisation Models for Musical Sound Source Separation", Computational Intelligence and Neuroscience, vol. 2008, Article ID 872425, 15 pages, 2008. doi:10.1155/2008/872425

[3] C. Fevotte, N. Bertin and J.-L. Durrieu. "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," Neural Computation, vol. 21, no 3, Mar. 2009

[4] R. Kompass. "A generalized divergence measure for non-negative matrix factorization". In Neuroinformatics workshop, Torun, Poland, Sept. 2005.

[5] A. Cichocki, R. Zdunek, and S.-I. Amari, "Csiszar's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms", In: Independent Component Analysis and Blind Signal Separation, ICA 2006, Charleston, SC, USA, March 5-8, 2006, Lecture Notes in Computer Science, Vol. 3889, Springer, pp. 32 - 39, 2006

[6] P. D. O'Grady. "Sparse Separation of Under-Determined Speech Mixtures". PhD Thesis, National University of Ireland Maynooth, 2007

[7] D. FitzGerald, M. Cranitch, and E. Coyle,"Musical Source Separation using

Generalised Non-negative Tensor Factorisation Models",Workshop on Music and Machine Learning, International Conference on Machine Learning, Helsinki, 2008

[8] J. Woodruff,B. Pardo, and R. Dannenberg,"Remixing Stereo Music with Score-informed Source Separation",Proceedings of the 7th International Conference on Music Information Retrieval,2006.

[9] B. W. Bader and T. G. Kolda, MATLAB Tensor Toolbox Version 2.2, http://csmr.ca.sandia.gov/ tgkolda/TensorToolbox/, January 2007.

[10] D. Lee, and H. Seung, "Learning the parts of objects by non-negative matrix factorisation", Nature, vol 401, no 6755, pp. 788-791, 1999.

[11] E. Vincent, R. Gibonval and C. Fevotte. "Performance Measurement in Blind Audio Source Separation", IEEE Transactions on Speech and Audio Processing, vol. 14 no. 4 pp 1462-1469, July, 2006.

[12] BSS_Eval toolbox available at http://bass-db.gforge.inria.fr/bss_eval/

[13] N. Bertin, C. Fevotte, and R. Badeau, "A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription",International Conference on Acoustics, Speech, and Signal Processing ICASSP09, Taipei, Taiwan, April 19-24, 2009.