NMF WITH SPECTRAL AND TEMPORAL CONTINUITY CRITERIA FOR MONAURAL SOUND SOURCE SEPARATION

Julian M. Becker, Christian Sohn and Christian Rohlfing

Institut für Nachrichtentechnik
RWTH Aachen University
D-52056 Aachen, Germany

ABSTRACT

Nonnegative Matrix Factorization (NMF) is a well suited and widely used method for monaural sound source separation. It has been shown, that an additional cost term supporting temporal continuity can improve the separation quality [1]. We extend this model by adding a cost term, that penalizes large variations in the spectral dimension. We propose two different cost terms for this purpose and also propose a new cost term for temporal continuity. We evaluate these cost terms on different mixtures of samples of pitched instruments, drum sounds and other acoustical signals. Our results show, that penalizing large spectral variations can improve separation quality. The results also show, that our alternative temporal continuity cost term leads to better separation results than the temporal continuity cost term proposed in [1].

Index Terms— audio source separation, nonnegative matrix factorization

1. INTRODUCTION

Nonnegative matrix factorization (NMF) is a frequently used method in audio source separation, e.g. [2, 3]. It was introduced by Paatero [4], but only became popular after Lee and Seung published efficient algorithms for its computation [5]. NMF is able to factorize audio signals into a specified number of components which correspond to individual sound events. These events can be assigned to the original sources by clustering, resulting in estimated separated sources.

As NMF was not originally developed for source separation, there are various options to extend it, to better adapt it to the task of audio source separation. Several extensions have been proposed, some using convolutive bases instead of multiplicative ones [6, 7], others extending the matrix factorization model to a tensor factorization model, to separate multichannel recordings [8]. Yet others introduce additonal constraints such as sparsity or temporal continuity [1, 9]. An overview over different versions of NMF can be found in [10].

In this paper we investigate an extension that adds a temporal continuity constraint to NMF. Based on this idea, we propose constraints supporting spectral continuity and also propose a new temporal continuity constraint. Our results show, that a spectral continuity constraint can be beneficial for audio source separation and that our alternative temporal continuity constraint results in better separation than the temporal continuity criterion proposed in [1]. We also show, that combining spectral and temporal constraints can be beneficial.

The paper is structured as follows: In Section 2, we provide basic information about NMF, the application of NMF in source separation and the temporal continuity criterion proposed in [1]. In Section 3, we propose constraints for spectral continuity and introduce our alternative temporal continuity term. In Section 4 we present our experimental results. Finally, in Section 5, we give our conclusions.

2. FUNDAMENTALS

2.1. Nonnegative Matrix Factorization

NMF approximates a nonnegative matrix $\mathbf{X}$ of size $K \times N$ by a product of two nonnegative matrices $\mathbf{B}$ and $\mathbf{G}$

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{B}\mathbf{G}, \tag{1}$$

with $\mathbf{B}$ of size $K \times I$ and $\mathbf{G}$ of size $I \times N$. I is a user defined parameter, which is usually chosen smaller than K and N.

The matrices $\mathbf{B}$ and $\mathbf{G}$ are iteratively calculated by minimizing an adequat distance function $c(\mathbf{B}, \mathbf{G})$ between $\mathbf{X}$ and $\tilde{\mathbf{X}}$. Commonly used distance functions are the Euclidean distance, the Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) distance. Lee and Seung [5] introduced efficient multiplicative update rules for the square of the Euclidean distance as well as for the KL divergence, resulting in convergence to a local minimum of the distance function. These update rules can be calculated using the gradient of $c(\mathbf{B}, \mathbf{G})$ with respect to $\mathbf{B}$,

$$\nabla_{\mathbf{B}} c(\mathbf{B}, \mathbf{G}) = \nabla^{+}_{\mathbf{B}} c(\mathbf{B}, \mathbf{G}) - \nabla^{-}_{\mathbf{B}} c(\mathbf{B}, \mathbf{G}) \tag{2}$$

where $\nabla^{+}_{\mathbf{B}} c(\mathbf{B}, \mathbf{G})$ and $\nabla^{-}_{\mathbf{B}} c(\mathbf{B}, \mathbf{G})$ are elementwise nonnegative terms of the gradient. Aquivalently, $\nabla_{\mathbf{G}} c(\mathbf{B}, \mathbf{G})$ is the gradient with respect to $\mathbf{G}$. The update rules are

$$\mathbf{B} \leftarrow \mathbf{B} \otimes \frac{\nabla^{-}_{\mathbf{B}} c(\mathbf{B}, \mathbf{G})}{\nabla^{+}_{\mathbf{B}} c(\mathbf{B}, \mathbf{G})} \tag{3}$$

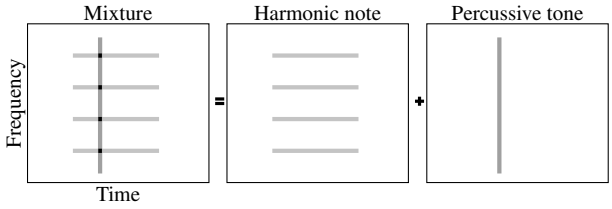

Fig. 1: Idealized illustration of the magnitude spectrogram of one harmonic note and one percussive tone.

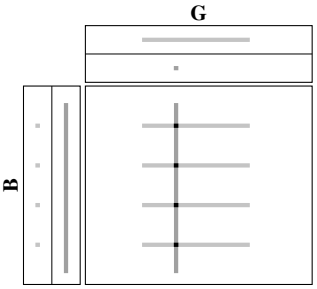

Fig. 2: Example of correct factorization with NMF

and

$$\mathbf{G} \leftarrow \mathbf{G} \otimes \frac{\nabla_{\mathbf{G}}^{-} c(\mathbf{B}, \mathbf{G})}{\nabla_{\mathbf{G}}^{+} c(\mathbf{B}, \mathbf{G})}, \tag{4}$$

where $\otimes$ denotes elementwise multiplication and the divisions are also elementwise. For the methods presented in thispaper, this generalized formulation of the update rules is sufficient. However, the exact update rules for KL-divergence and squared Euclidean distance can be found in [5] and for the IS-distance in [10].

2.2. Source separation using NMF

To use NMF for source separation, the time signal $\mathbf{x}$, consisting of M sources $\mathbf{s}_m$, first has to be transformed to time-frequency domain using short time Fourier transform (STFT). This results in a complex valued spectrogram $\underline{\mathbf{X}}$, which has a spectral and a temporal dimension. The NMF can be applied to the magnitude $\mathbf{X} = |\underline{\mathbf{X}}|$ of this spectrogram for audio source separation. Figures 1 and 2 illustrate this procedure. Figure 1 shows an idealized magnitude spectrogram of a mixture of one harmonic note (horizontal structure) and one percussive tone (vertical structure) for $M = 2$. Figure 2 shows the result of factorization of this mixture using NMF with $I = 2$. The matrices $\mathbf{B}$ (on the left) and $\mathbf{G}$ (on top) are the result of the NMF. The columns of $\mathbf{B}$ (these vectors will from now on be denoted $\mathbf{b}_i$) capture the spectral shape of the acoustical events and can therefore be interpreted as spectral bases. The vectors $\mathbf{g}_i$ (the rows of the matrix $\mathbf{G}$) can be interpreted as temporal activation vectors. The matrices $\tilde{\mathbf{C}}_i = \mathbf{b}_i \mathbf{g}_i$ can be interpreted as spectrograms of individual acoustical events.
$\tilde{\mathbf{X}} = \sum_i \tilde{\mathbf{C}}_i$ is only an approximation for $\mathbf{X}$. However, it is

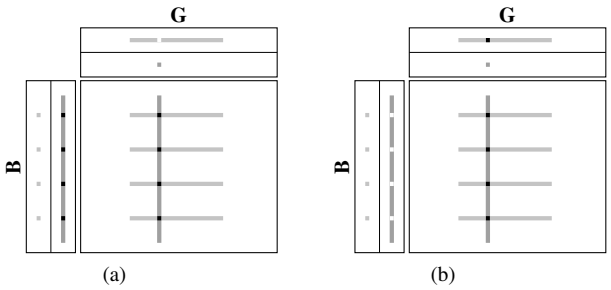

Fig. 3: Examples of incorrect factorizations with NMF

desirable that the factorized spectrograms exactly sum up to the original spectrogram $\mathbf{X}$. This is done in a filtering step, which is also used to restore phase, using the phase of the mixture:

$$\hat{\underline{\mathbf{C}}}_i = \underline{\mathbf{X}} \otimes \left(\frac{\tilde{\mathbf{C}}_i}{\sum_i \tilde{\mathbf{C}}_i} \right). \tag{5}$$

This Wiener-like filtering is a frequently used postprocessing step of the results of the NMF [3], [11].
In a more complex mixture, I would have to be chosen corresponding to the number of the acoustical events in the mixture. Usually I is higher than the number of sources M, therefore the resulting spectrograms $\hat{\underline{\mathbf{C}}}_i$ have to be clustered to the melodies of the original sources. This is done in a clustering step, resulting in the estimated spectrograms $\hat{\underline{\mathbf{S}}}_m$ for the original sources.
These spectrograms are then transformed into time domain by inverse short-time Fourier transform (ISTFT). This step results in the estimations for the original sources in time domain $\hat{\mathbf{s}}_m$.

2.3. NMF with temporal continuity

The results of the NMF differ, depending on the initial elements of $\mathbf{B}$ and $\mathbf{G}$, because the NMF only leads to a local minimum of the distance function. This can lead to problems: For example, the spectrogram in Figure 2 could also be factorized incorrectly, as shown in Figure 3. In these examples, the overlapping points of the two sources are assigned incorrectly, leading to gaps and peaks either in the temporal activation vector of the harmonic source, or the spectral basis vector of the percussive source.
To prevent this unwanted result, Virtanen proposed to add an additional temporal continuity term to the cost function of the NMF [1]. With this addition, the cost term to be minimized transforms to

$$c(\mathbf{B}, \mathbf{G}) = c_r(\mathbf{B}, \mathbf{G}) + \alpha_t c_t(\mathbf{G}), \tag{6}$$

where $c_r(\mathbf{B}, \mathbf{G})$ ist a reconstruction error term (e.g. KL-divergence), $c_t(\mathbf{G})$ is the temporal continuity term and α_t is a weight to adjust the influence of the temporal continuity

term. For $\alpha_t = 0$ this model equals the standard NMF. The update rule for $\mathbf{G}$ transforms to

$$\mathbf{G} \leftarrow \mathbf{G} \otimes \frac{\nabla^-_{\mathbf{G}} c_r(\mathbf{B},\mathbf{G}) + \alpha_t \nabla^-_{\mathbf{G}} c_t(\mathbf{G})}{\nabla^+_{\mathbf{G}} c_r(\mathbf{B},\mathbf{G}) + \alpha_t \nabla^+_{\mathbf{G}} c_t(\mathbf{G})}, \tag{7}$$

while the update term for $\mathbf{B}$ stays the same as in Equation (3). The temporal continuity term proposed by Virtanen is a temporal squared difference (TSD) cost term, which can be calculated as

$$c_t(\mathbf{G}) = \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{n=2}^{N} (g_{n,i} - g_{n-1,i})^2, \tag{8}$$

with $\sigma_i = \sqrt{(1/N)\sum_{n=1}^{N} g_{n,i}^2}$ being the standard deviation of each component i. $g_{n,i}$ denotes one element of the matrix $\mathbf{G}$ at indizes n and i. The negative and positive gradient terms of this cost function are

$$[\nabla^-_{\mathbf{G}} c_t(\mathbf{G})]_{i,n} = 2N \frac{g_{i,n-1} + g_{i,n+1}}{\sum_{l=1}^{N} g_{i,l}^2} + \frac{2N g_{i,n} \sum_{l=2}^{N} (g_{i,l} - g_{i,l-1})^2}{\left(\sum_{l=1}^{N} g_{i,l}^2\right)^2} \tag{9}$$

and

$$[\nabla^+_{\mathbf{G}} c_t(\mathbf{G})]_{i,n} = \frac{4N g_{i,n}}{\sum_{l=1}^{N} g_{i,l}^2} \tag{10}$$

3. PROPOSED CONTINUITY TERMS

The NMF with temporal continuity favors factorizations with continuous temporal activation vectors. This way temporal gaps as in Figure 3(a) or temporal peaks as in Figure 3(b) are circumvented. These problems usually only occur for harmonic components. While for harmonic components temporal continuity and spectral discontinuity can be assumed, spectral components usually show a spectral continuity and temporal discontinuity. Thus, gaps and peaks also occur in the spectral vectors $\mathbf{b}_i$ for percussive components (see Fig. 3). Therefore, we propose to use an additional cost term favoring continuous spectral basis vectors.
With this addition, the cost of Equation (6) transforms to

$$c(\mathbf{B},\mathbf{G}) = c_r(\mathbf{B},\mathbf{G}) + \alpha_t c_t(\mathbf{G}) + \alpha_s c_s(\mathbf{B}), \tag{11}$$

where $c_s(\mathbf{B})$ is a spectral continuity term and α_s is a weight to adjust the influence of the spectral continuity term. The update rule for $\mathbf{B}$ transforms to

$$\mathbf{B} \leftarrow \mathbf{B} \otimes \frac{\nabla^-_{\mathbf{B}} c_r(\mathbf{B},\mathbf{G}) + \alpha_s \nabla^-_{\mathbf{B}} c_s(\mathbf{B})}{\nabla^+_{\mathbf{B}} c_r(\mathbf{B},\mathbf{G}) + \alpha_s \nabla^+_{\mathbf{B}} c_s(\mathbf{B})}, \tag{12}$$

while the update term for $\mathbf{G}$ stays the same as in Equation (7).

3.1. Spectral continuity terms

We propose two differently motivated terms for spectral continuity.

3.1.1. Spectral squared difference

The spectral squared difference (SSD) term is motivated by Virtanens TSD (Eq. (8)). Aquivalently to this term, the proposed SSD term is

$$c_s(\mathbf{B}) = \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{k=2}^{K} (b_{i,k} - b_{i,k-1})^2, \tag{13}$$

with $\sigma_i = \sqrt{(1/K)\sum_{k=1}^{K} b_{n,i}^2}$ being the standard deviation of each component i. The negative and positive gradient terms of this cost function, $[\nabla^-_{\mathbf{B}} c_s(\mathbf{B})]_{k,i}$ and $[\nabla^-_{\mathbf{B}} c_s(\mathbf{B})]_{k,i}$ can be calculated equivalently to Equations (9) and (10).

3.1.2. Spectral flatness

The spectral flatness (SF) term is motivated by the MPEG-7 [12] spectral flatness descriptor, which estimates a measure for spectral flatness as the ratio between the geometric and the arithmetic mean of spectral power coefficients. Using the spectral flatness as cost term $c_s(\mathbf{B})$ would mean that flatness is penalized, thus leading to a sparseness criteria (see [9]). As we want to favor flat basis vectors, we take the inverse of the spectral flatness descriptor as cost function. The proposed SF term is

$$c_s(\mathbf{B}) = \sum_{i=1}^{I} \frac{\frac{1}{K}\sum_{k=1}^{K} b_{k,i}}{\sqrt[K]{\Pi_{k=1}^{K} b_{k,i}}}. \tag{14}$$

The negative and positive gradient terms of this cost function are

$$[\nabla^-_{\mathbf{B}} c_s(\mathbf{B})]_{k,i} = \frac{\sum_{l=1}^{K} b_{l,i}}{K^2 b_{k,i} \sqrt[K]{\Pi_{l=1}^{K} b_{l,i}}} \tag{15}$$

and

$$[\nabla^+_{\mathbf{B}} c_s(\mathbf{B})]_{k,i} = \frac{1}{K \sqrt[K]{\Pi_{l=1}^{K} b_{l,i}}}. \tag{16}$$

3.2. Alternative temporal continuity term

Motivated by the inverse behavior of percussive and harmonic signals in time and frequency, we also propose to use a temporal flatness (TF) term as alternative to Virtanens TSD. The proposed TF term is

$$c_t(\mathbf{G}) = \sum_{i=1}^{I} \frac{\frac{1}{N}\sum_{n=1}^{N} g_{i,n}}{\sqrt[N]{\Pi_{n=1}^{N} g_{i,n}}}. \tag{17}$$

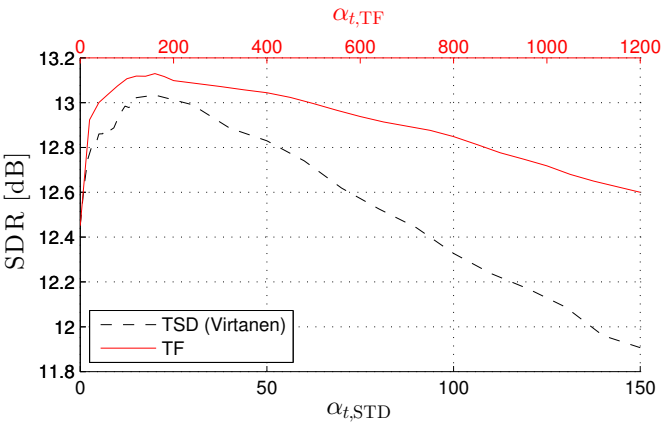

Fig. 4: Results for different values of α_t for Virtanens TSD and the proposed TF cost term, without using spectral cost terms.

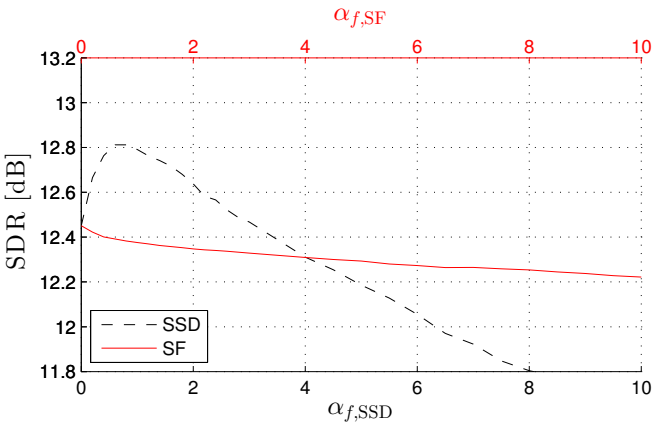

Fig. 5: Results for different values of α_s for the SSD and the spectal flatness cost term, without using temporal cost terms.

The negative and positive gradient terms of this cost function, $[\nabla^-_{\mathbf{G}} c_t(\mathbf{G})]_{i,n}$ and $[\nabla^-_{\mathbf{G}} c_t(\mathbf{G})]_{i,n}$ can be calculated equivalently to Equations (15) and (16).

4. EVALUATION & RESULTS

We performed source separation as described in Section 2.2 to evaluate the influence of the different additional terms. As we wanted to evaluate the separation quality of the NMF without being affected by errors introduced by a clustering algorithm, we use a non-blind clustering where the original signals are used as references for clusters, as described in [1]. As measure for separation quality, we use the Signal-to-distortion ratio (SDR) as proposed in [13].

4.1. Testset & Setup

The testset consists of 60 audio signals, including harmonic and percussive signals, speech, vocals and noise, each being sampled with 44.1 kHz. These signals were mixed in every possible two-source combination, resulting in 1770 mixtures. The testset is identical to the one used in [3].
For the short time fourier transform, we used a window size of $s_w = 2^{12}$ and a hop size of $s_h = 2^{11}$ samples. The parameter I of the NMF was set to 15 for every mixture. This value was chosen, because it had shown to be a suitable average number of components for this testset. We tested several reconstruction error terms for the NMF, namely IS-distance, KL-divergence and squared Euclidean distance. The KL-divergence produced the best separation results, therefore we use KL-divergence for evaluation.

4.2. Temporal cost terms

In a first experiment, we compared the influence of Virtanens TSD term and the TF term proposed in Equation (17). For this experiment, we set $\alpha_s = 0$, so the spectral cost terms would not influence the results. Figure 4 shows the average SDR over all 1770 mixtures for different values of α_t. The x-axes are scaled differently for better comparability. The result for $\alpha_t = 0$ equals the standard NMF.
The average SDR for the standard NMF was 12.45 dB. Both temporal cost terms resulted in higher SDR for different values of α_t. Virtanens TSD reached a maximum SDR of 13.03 dB for $\alpha_t = 20$. Our TF cost term reached a maximum SDR of 13.13 dB for $\alpha_t = 160$.
For higher values of α_t, the separation quality decreases for both temporal cost functions. However, for the TF term, the quality decreases slower than for the TSD. Temporal flatness seems to be a more robust method in terms of the correct choice of α_t.

4.3. Spectral cost terms

In a second experiment, we compared the influence of the two proposed spectral cost terms. For this experiment, we set $\alpha_t = 0$, so the temporal cost terms would not influence the results. Figure 5 shows the average SDR over all 1770 mixtures for different values of α_s. The result for $\alpha_s = 0$ equals the standard NMF.
The average SDR for the standard NMF was 12.45 dB. Only the SSD cost term resulted in higher SDR for different values of α_s. It reached a maximum SDR of 12.81 dB for $\alpha_s = 0.8$. The SF cost term resulted in worse separation for every $\alpha_s > 0$.

4.4. Combination of temporal and spectral cost terms

In a third experiment, we tried to combine spectral and temporal cost terms to evaluate, if this combination can lead to even better separation results. Motivated by the results of the previous experiments, we decided to only use the TF and the SSD term.

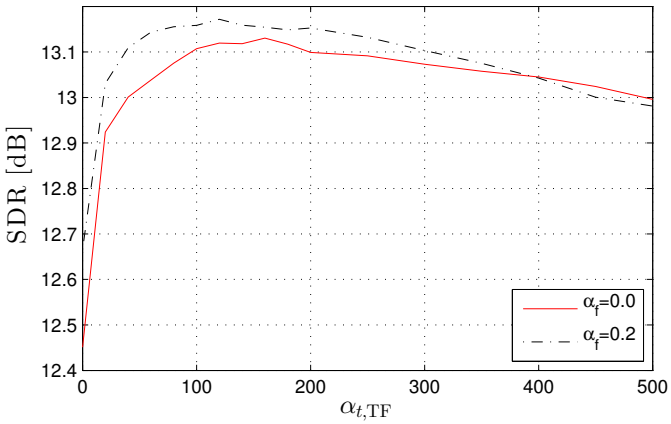

Fig. 6: Results for different values of α_s for the SSD cost term with different values of α_t for the TF cost term.

We performed source separation with different combinations of α_t and α_s. The results showed, that adding the SSD term with low α_s, while using TF improves the separation results slightly. Figure 6 shows the average SDR over different values of α_t, using TF, for $\alpha_s = 0.2$, which was the value leading to the best results. We added $\alpha_s = 0$, which equals the result in Figure 4, for comparison. It can be observed that adding an SSD term with $\alpha_s = 0.2$ improves the results. For higher values of α_t there is no more improvement.

5. CONCLUSION

In this paper, we proposed two additional cost terms for NMF, supporting continuous spectral basis vectors. Our results showed, that one of the proposed cost terms leads to better results than the standard NMF. We also proposed a temporal flatness cost term as alternative to Virtanens TSD cost term. Our results showed, that this TF cost term resulted in better separation. We also showed that combinations of temporal and spectral cost terms can improve separation results further.
Future research will aim for better methods to combine different cost terms by better adapting α_t and α_s to the different components and to find ways of generally adapting the values of α_t and α_s to different mixtures.

REFERENCES

[1] T. Virtanen, "Monaural Sound Source Separation by Nonnegative matrix Factorization With Temporal Continuity and Sparseness Criteria," in *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE, 2007, vol. 3, pp. 1066–1074.

[2] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 5, pp. V–V.

[3] M. Spiertz and V. Gnann, "Beta Divergence for Clustering in Monaural Blind Source Separation," in *128th AES Convention*. AES, 2010.

[4] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[5] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2000, pp. 556–562.

[6] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*. 2004, vol. 3195 of *Lecture Notes in Computer Science*, pp. 494–499, Springer Berlin Heidelberg.

[7] M. N. Schmidt and M. Mørup, "Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation," in *Proceedings of the 6th International Symposium on Independent Component Analysis and Blind Signal Separation*, 2006.

[8] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," 2005.

[9] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A comparative study on sparsity penalties for NMF-based speech separation: Beyond LP-norms," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 858–862.

[10] 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.

[11] R. Jaiswal, D. Fitzgerald, D. Barry, E. Coyle, and S. Rickard, "Clustering NMF Basis Functions Using Shifted NMF for Monaural Sound Source Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*. IEEE, 2011.

[12] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 audio and beyond: Audio content indexing and retrieval*, John Wiley & Sons, 2006.

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," in *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE, 2006.