EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization

Yuki Mitsufuji[1*] and Axel Roebel[2]

Abstract

This paper proposes a new method to enhance the performance of non-negative tensor factorization (NTF), one of the most prevalent source separation techniques nowadays. The enhancement is mainly achieved by introducing weights on bin-wise NTF cost functions, which differentiates NTF target components from other components so that the target should be approximated more precisely than others. Assuming sources are distributed sparsely in a 2-D sound field, the target components approximating a target source are exclusively selected by a user, or from accompanying images by means of providing a spatial cue to an NTF framework. The spatial cue is given in a similar format to the well-known binaural feature, inter-channel level difference (IID). This helps incorporate the spatial cue into the system, since the similar features of this format can be easily calculated from every spectrogram bin. The weighting functions are designed taking into account the distance between the spatial cue and the calculated features. Namely, the largest values are assigned to the spectrogram bins where the features present the highest similarity to the spatial cue, and the value decreases in proportion to the distance between them. The method is evaluated in terms of separation quality, comparing the proposed algorithm to the conventional NTF technique, PARAFAC-NTF, as well as other source separation techniques. The evaluation results measured by the metric signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) demonstrate the effectiveness of the new method, improved primarily by the weighting function and the initialization based on IID, while demonstrating a decrease in computational costs, a significant problem with NTF.

1 Introduction

In the last few decades, non-negative matrix factorization (NMF) has become one of the most prevalent techniques to tackle the underdetermined source separation problem where the number of sources is greater than or equal to the number of observations. NMF 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. Cost function for NMF estimation has long been investigated by several researchers [1-3]. In particular, the Itakura-Saito divergence is known to be an appropriate cost function for approximation of audio spectra due to its scale-invariant nature. To complete NMF-based source separation, clustering of the decomposed basis elements follows the factorization to properly classify them to corresponding sources. A large number of related techniques have been developed so far [4-6].

Taking advantage of prior information for the purpose of enhancing the performance of NMF has been widely investigated. Smaragdis et al. have attempted to make use of a user-guided humming for the extraction of melodies in a mixture [7]. Diknen et al. investigated the Bayesian NMF model assigning different prior distributions for tonal and percussive signals [8]. Ewert et al. presented an extended approach that uses additional score information to guide the NMF process [9]. Although a number

*Correspondence: Yuhki.Mitsufuji@jp.sony.com
[1]Audio Technology Development Department, Sony Corporation, Tokyo 141-8610, Japan
Full list of author information is available at the end of the article

of researches related to the prior knowledge of time-frequency features have been introduced, it is not yet common to incorporate a spatial feature (hereafter, spatial cue) in NMF due to its algorithm framework. Since NMF which separates a single-channel signal does not produce any spatial information during the separation process, it has been difficult to associate a spatial cue until the emergence of multichannel NMF.

Non-negative tensor factorization (NTF), known as one of the multichannel NMF techniques, 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 being investigated for use in NTF [10-12]. Extension to the third dimension provides another matrix that describes the energy distribution of each basis component on every channel, which can also be regarded as spatial information. This technique enables the NMF approach to be adapted to easily accept a spatial cue [13].

This paper proposes a promising method to enhance NTF performance, taking advantage of a spatial cue given by users or from accompanying images. The enhancement is mainly achieved by introducing weights on bin-wise NTF cost functions, which differentiates a target component from other components. 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 weights to bins where the target is likely to exist and less weights to the others. Virtanen et al. proposed perceptually weighted NMF that provides perceptually motivated weights for each critical band in each frame in accordance with the loudness perception of the human auditory system [14]. Nevertheless, to our knowledge, no research regarding NMF that incorporates the weighting function for spatially focusing on target component estimation has been proposed. The evaluation results show that this method is advantageous in terms of separation quality over conventional PARAFAC-NTF and other source separation techniques such as the Degenerate Unmixing Estimation Technique (DUET) [15-17].

It should be noted that apart from NTF, there exist other approaches to address the source separation problem in multichannel audio. Especially, the algorithms based on local Gaussian models [18] using the spatial covariance matrix (SCM) for encoding spatial positions of source signals [19] have been shown to outperform the simpler NTF approach that will be used in the following. A rank-1 convolutive model assuming target sources existed in a non-reverberant environment has been proposed in [20]. Full-rank unconstrained model with NMF was further introduced by Arberet et al. to account for reverberant conditions [21,22]. A modular framework for these algorithms has been presented in [23,24]. These models present a significant improvement in terms of separation quality as compared to the NTF model. A problem with these models that use the expectation-maximization (EM) algorithm for optimization is the significant increase in computational complexity when compared to the NTF model. Sawada et al. have addressed this problem by introducing multiplicative update rules in place of the EM algorithm, but still their optimization requires significantly more computation time compared to single-channel NMF [25]. While we believe that the proposed weighting scheme can improve performance with other existing multichannel factorization algorithms, we selected the NTF algorithm for investigation into the effectiveness of our weighting scheme to limit the computational costs.

This paper is organized as follows: Section 2 briefly explains NTF-based source separation. Section 3 describes a possible incorporation of spatial cue into NTF. Section 4 discusses the proposed method. Section 5 shows evaluation results on quality and computational costs. Finally, Section 6 presents some concluding remarks.

2 NTF-based source separation

2.1 Non-negative tensor factorization

A multichannel audio signal that has been transformed into a set of spectrograms (one for each of the J channels) can 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 optimization problem is solved:

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

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} where j indicates the channel index, k the bin index of the spectrogram, and l the time index of the spectrogram. $\alpha(\mathrm{H})$ represents additional constraints on matrix H, which are taken into account during minimization of the cost function. The β-divergence, d_β, is suitable for NTF, allowing the separation quality to be changed, subject to the parameter β [26]. When β is equal 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 β-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:

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

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

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

where $\nabla_{\mathrm{H}}\alpha(\mathrm{H}) = \nabla_{\mathrm{H}}^{+}\alpha(\mathrm{H}) - \nabla_{\mathrm{H}}^{-}\alpha(\mathrm{H})$, both $\odot$ and / denote element-wise calculations, A $\circ$ B denotes $J \times K \times P$ tensor with elements $a_{jp}b_{kp}$ when A and B are $J \times P$ and $K \times P$ [27], and $\langle \mathrm{A}, \mathrm{B} \rangle_{\{C\},\{D\}}$ denotes a contracted product [15]. 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 [3].

2.2 Wiener filtering

As soon as the approximation of spectrogram $\widehat{\mathrm{V}}$ composed of multiple basis elements Q, W, and H has been completed by NTF, Wiener filtering is followed to extract the target signal such that

$$y_{jkl} = \frac{\sum_{p \in P_{tar}} q_{jp} w_{kp} h_{lp}}{\hat{v}_{jkl}} x_{jkl}, \tag{5}$$

where P_{tar} denotes the collection of bases considered as the target group. x_{jkl} and y_{jkl} denote the short-time Fourier transform (STFT) of input audio signal and the separated target signal, respectively. It should be noted that the more sophisticated method called Multichannel Wiener Filter employing spatial covariance matrices is known to give a better performance in more complex mixing scenario [28,29].

3 Incorporation of spatial cues

We devised two ways of incorporating a spatial cue. Figure 1 shows a 2-D 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. It can be represented more intuitively by the following equation,

$$\overleftarrow{Q}_p = 2 \tan^{-1} \left(\frac{q_{1,p}}{q_{0,p}} \right)^{\delta}, \tag{6}$$

where $\delta = 1.0$ when magnitude spectrums are used as an input of NTF and $\delta = 0.5$ for power spectrums. $\overleftarrow{Q}$ denotes the angles of the arrows in radians clockwise from the horizontal axis in Figure 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 $\overleftarrow{Q}$, specifically called $\overleftarrow{Q}_{\mathbf{sc}}$. However, it should be noted that these arrows serve only for visualization purposes and are different from the azimuth angles in the real world.

Idea 1 (Figure 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 (Figure 1, left). Interpolation between the two groups (three arrows on the left and three arrows on the 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

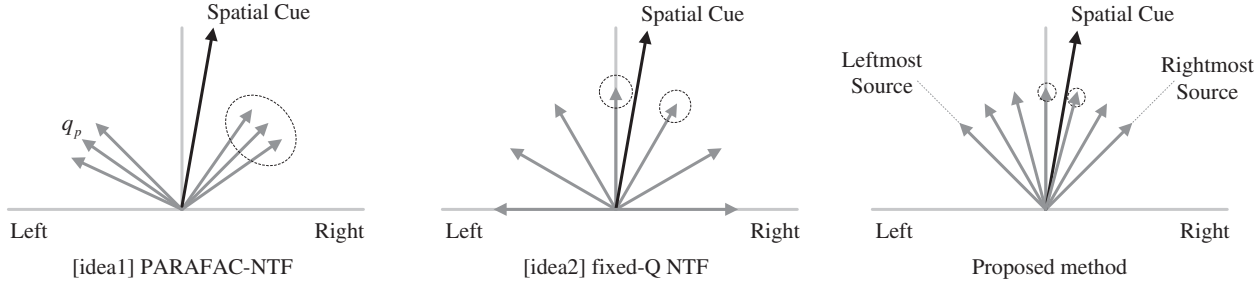

Figure 1 2-D representation of a channel matrix Q. Three different solutions incorporating NTF.

the basis element candidates are sufficiently close to each other. We call idea 1 PARAFAC-NTF (p-NTF).

In idea 2 (Figure 1, center), 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 elements located at both sides of the spatial cue are selected as target elements. We call idea 2 fixed-Q NTF (f-NTF), and each direction is numbered with a direction index, d, ranging from 0 to $D-1$.

Figure 2 shows block diagrams of two different solutions. The big difference that can be observed in this figure is that in p-NTF, the spatial cue cannot be passed to the NTF process, whereas in f-NTF it is, which allows NTF to take advantage of spatial information. Two things make idea 2 worth focusing on: the computational efficiency, since the channel matrix, Q, does not have to be updated thanks to spatial information provided from a spatial cue; and the potential improvement in quality due to the prior knowledge provided by the spatial cue.

The next section presents more details and a variant of the f-NTF method that is called spatial cue (sc-NTF).

4 Proposed method

4.1 Choice of divergence

As mentioned in Section 2.1, three settings of the β-divergence are commonly used for NMF and NTF: Euclidean distance ($\beta = 2$), KL divergence ($\beta = 1$), and Itakura-Saito (IS) divergence ($\beta = 0$). Their differences were investigated by C. Févotte [2]. One important characteristic of the IS divergence that is not shared with the two other types of divergence is that the absolute scale of given audio does not affect the total cost of the divergence. That is, the unnoticeably small spectrogram bins can be approximated as well as the dominant bins. We assume that IS-NTF is thus more appropriate when a relatively small signal might come from a direction close to that of the spatial cue. However, this assumption is probably true only when there is little ambient noise [30]. Thus, we selected IS-NTF for our initial experiments and used noise-free input signals, such as commercial music. Another motivation for employing IS divergence comes from a statistical perspective. It has been shown that the ML estimation of a sum of complex Gaussian components representing for each spectral bin is equivalent to minimizing IS divergence between the ideal and estimated power spectrograms [31]. While there are clear advantages of the IS divergence, IS-NTF suffers from the fact that it is more often caught in local minima.

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 [32]. For example, the training could start with EUC-NTF (NTF based on Euclidean distance), 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.

4.2 Initialization of channel matrix

The initialization of channel matrix, Q, is based on the understanding of the matrix Q explained in Section 3.

$$\overleftarrow{B}_{kl} = 2\tan^{-1}\left(\frac{v_{1,kl}}{v_{0,kl}}\right)^{\delta}, \tag{7}$$

where $v_{0,kl}$ and $v_{1,kl}$ are the spectrogram bins for the left and right channels, respectively. This feature concerns the arrows in Figure 1 and their relationships to

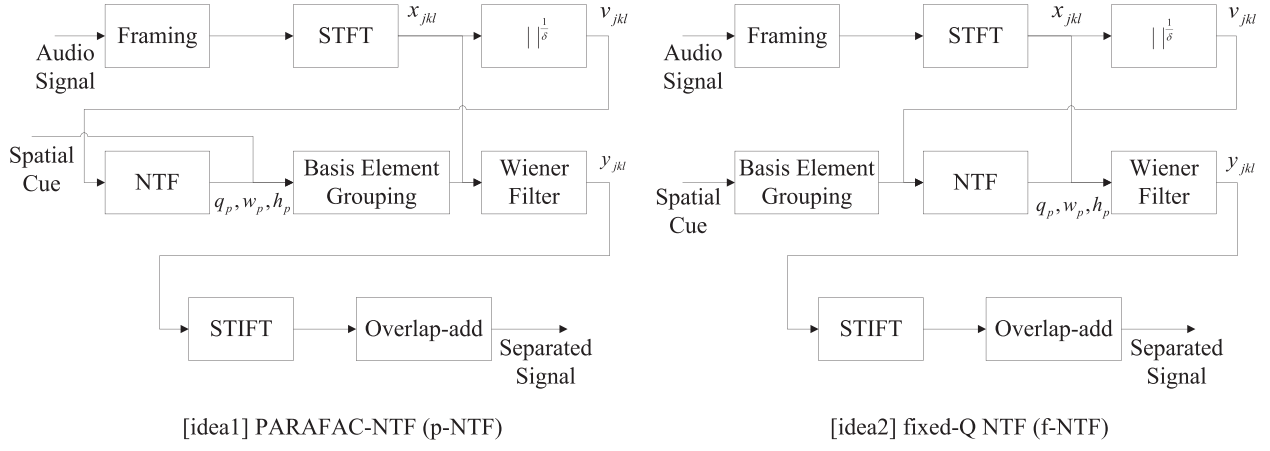

Figure 2 Block diagrams of two different solutions. Block diagrams for two different solutions using NTF.

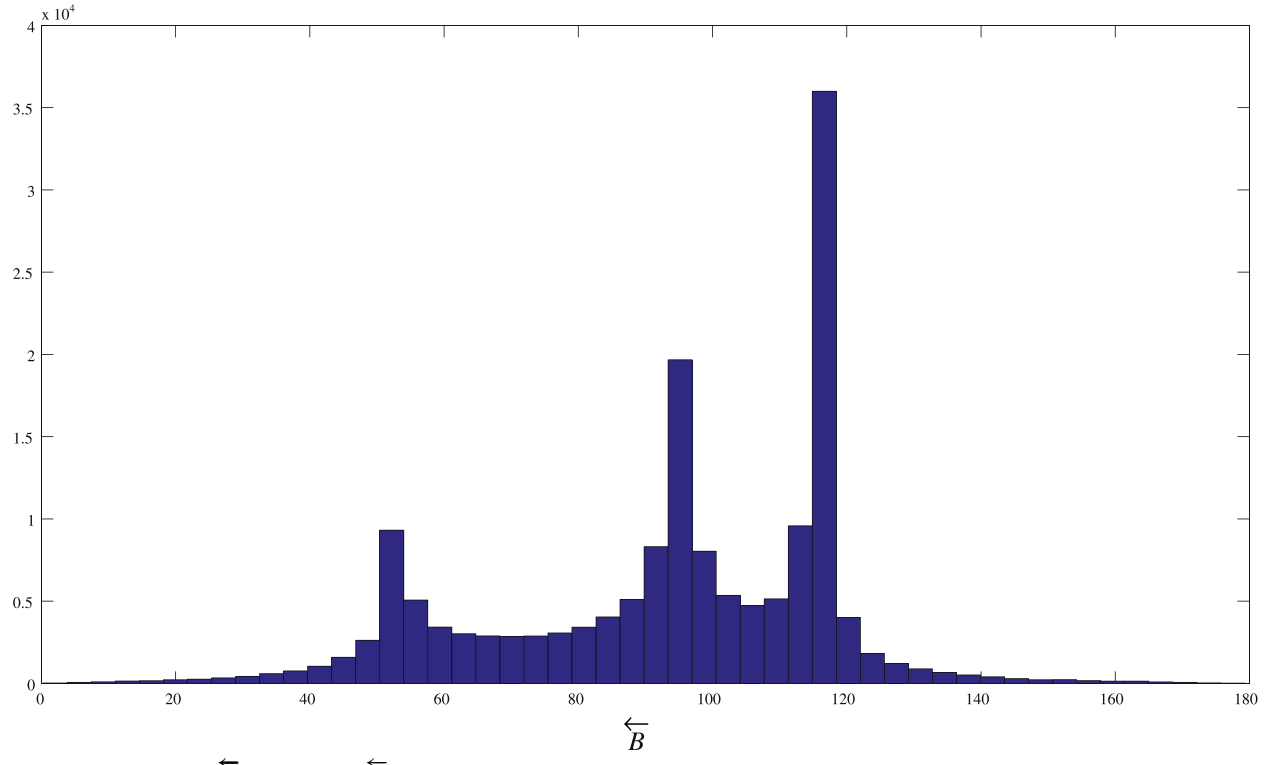

Figure 3 Histogram of $\overleftarrow{B}$. Histogram of $\overleftarrow{B}$ of the three-source mixture. Three peaks are observed.

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 $\overleftarrow{B}$ (Figure 3), which represents the dominant presence of the sources. The basis elements are preferentially allocated by initializing channel matrix, Q, 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. Figure 3 shows the histogram of $\overleftarrow{B}$ calculated from a mixture of three audio instruments placed in different positions. The length of arrows corresponds to a frequency of each bin. As we can see the arrows in three directions in the histogram, it is highly likely that sources exist at 50° to 60°, 90° to 100°, and 110° to 120°. However, the allocation of basis elements to the left of the leftmost peak and to the right of the rightmost peak is not required since the superposition of the sources never appears outside of these peaks. It is therefore preferable to allocate basis elements inside the range spanned by the left and rightmost peaks that exist in the measured histogram of $\overleftarrow{B}$, as can be seen in the right image of Figure 1.

4.3 Initialization of frequency matrix and time matrix

Initialization of frequency matrix, W, and time matrix, H, is simply carried out by taking advantage of information of the histogram such that

$$w_{kp} = \frac{1}{N_{Grp(d)}J} \sum_{j} \sum_{l \in Grp(d)} v_{jkl} \quad k,l,p \in Grp(d), \tag{8}$$

$$h_{lp} = \frac{1}{N_{Grp(d)}J} \sum_{j} \sum_{k \in Grp(d)} v_{jkl} \quad k,l,p \in Grp(d), \tag{9}$$

where $N_{Grp(d)}$ denotes the frequency per bin of the histogram, and $Grp(d)$ denotes the collection of bases allocated to the direction with the index d. Normalization of the matrices follows for the purpose of concentrating the energy of input tensor V into H.

$$\mathrm{Q}_p = \mathrm{Q}_p/|\mathrm{Q}_p|_1,\ \mathrm{W}_p = \mathrm{W}_p/|\mathrm{W}_p|_1, \mathrm{H}_p = |\mathrm{Q}_p|_1|\mathrm{W}_p|_1\mathrm{H}_p, \tag{10}$$

$$e_d = \sum_{d} \sum_{p \in Grp(d)} |\mathrm{H}_p|_1, \tag{11}$$

where $|\cdot|_1$ denotes the L1-norm, and e_d denotes the directional energy associated with the direction index d.

4.4 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:

$$g_{jkl} = \exp\left(-\frac{\psi}{D}\left|\frac{\overleftarrow{Q}_{\mathbf{sc}} - \overleftarrow{Q}_p}{\Delta\overleftarrow{Q}}\right|\right) \quad k,l,p \in Grp(d), \tag{12}$$

where ψ determines the shape of the exponential function. Figure 4 changes the weighting parameter ψ and $\overleftarrow{Q}_p$ that creates different shapes while forcing $\overleftarrow{Q}_{\mathbf{sc}}$ to point toward 100°. The weighting values of different ψ when $\overleftarrow{Q}_p = 72$ are accentuated with markers. When ψ equals 0, all the weights for bin-wise cost functions become 1, which boils down the update rules of sc-NTF described in Section 2.1 to those used for PARAFAC-NTF.

4.5 Constraints

The energy for each direction should be estimated by adding all the basis elements in matrix H over time, equal to the procedure done by Equation 11. The estimated energy is fixed so that it can be used as a reference to constrain the energy distribution of the estimated tensor. This should reduce the likelihood of being trapped in local minima. Here, we again use the IS divergence to measure distance. The constraint on energy in a given direction is

$$\alpha(\hat{v}_{jkl}) = \mu\sum_{d=0}^{D-1} d_{IS}\left(\sum_{jkl\in Grp(d)} v_{jkl} \,\middle|\, \sum_{jkl\in Grp(d)} \hat{v}_{jkl}\right). \tag{13}$$

By taking into account the normalization procedure in Equation 10, the equation can be boiled down to

$$\alpha(\mathrm{H}) = \mu\sum_{d=0}^{D-1} d_{IS}\left(e_d \,\middle|\, \sum_{p\in Grp(d)} |\mathrm{H}_p|_1\right), \tag{14}$$

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

For IS-NTF, the following should hold for the derivative of the constraint:

$$\nabla\alpha(h_p) = \frac{\mu}{e_d} - \frac{\mu\sum_{p\in Grp(d)}|\mathrm{H}_p|_1}{(e_d)^2}. \tag{15}$$

5 Evaluation

5.1 Separation quality

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) [33]. 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 [34]). 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 2-D 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.

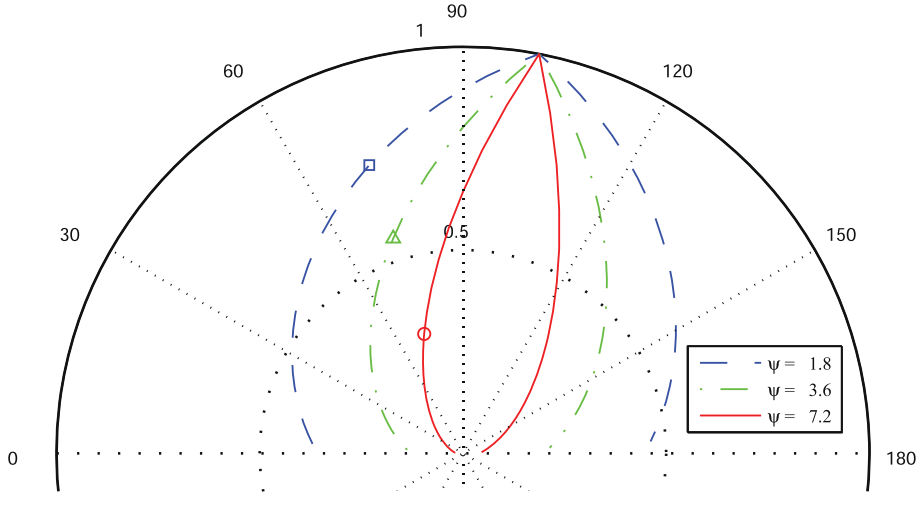

Figure 4 Spatial representation of the weighting function, G. The weighting function controlled its shape by the weighting parameter ψ. Three different shapes are shown when $\overleftarrow{Q}_{\mathrm{sc}} = 100$.

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. This difference can be seen in Figure 2. Our experiments on two grouping algorithms, k-means and k-nearest neighbor, to the ground truth showed that the latter yielded better results. The number of centroids is determined according to the number of basis components.

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 (1,024-point FFT, half overlap, the number of the basis elements $P = 90$, and the number of directions $D = 18$). 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 2011 paper [27]. 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 $\psi = 3.6$ for the weighting tensor G and $\mu = 300$ for the constraint. More details with results using different settings of ψ can be found in Figure 5.

Figure 6 shows test results for harmonic sound (nodrums), and Figure 7 shows results for percussive sound (wdrums). 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 the accuracy of approximation of sources far from the spatial cue. This can be deduced from the final value of the cost function. Table 1 shows the IS divergence per bin for four methods including sc-NTF without the weighting function G. sc-NTF using the weighting function produces worse results than p-NTF and sc-NTF without the weighting function in terms of approximation, due to the lesser weights of the broad range and the more weights of the relatively narrow target direction. 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, which means that the proposed method helps to avoid being trapped in local minima.

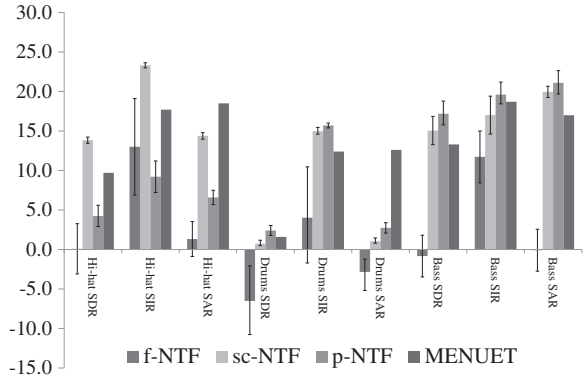

Figure 6 Test results for harmonic sound. Dataset nodrums. Mixture of three different instruments: bass, lead guitar, and second lead guitar.

5.2 Computational cost

As mentioned earlier, the omission of the calculation of the channel matrix Q should be a great advantage for sc-NTF. A simple experiment was conducted by measuring the runtime on an Intel Core i7 (2.80 GHz) processor for three NTF methods implemented in MATLAB code. It should be noted that the code was not particularly optimized by incorporating external libraries written in, for instance, C language. Instead, the built-in functions automatically provided by MATLAB, such as division and log functions, were used. The conditions of the experiment were the same as those for the quality evaluation

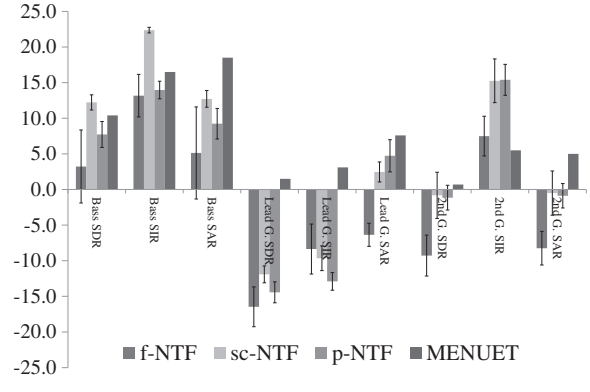

Figure 5 Test results of different shapes of the weighting function. Overall results of different settings of ψ is shown, from 0.9 to 7.2.

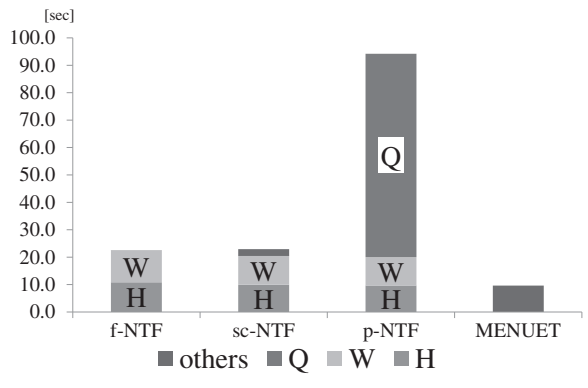

Figure 7 Test results for percussive sound. Dataset wdrums. Mixture of three different instruments: hi-hat, drums, and bass.

Table 1 Comparison of convergence

	f-NTF	sc-NTF (w/o G)	sc-NTF	p-NTF
IS divergence per bin	0.19300	0.16759	0.18300	0.16857

Itakura-Saito divergence per bin after 200 iterations.

described in Section 5.1 (the number of channels $J = 2$, the number of frequency bins $K = 513$, and the number of frames $L = 314$). Figure 8 shows that p-NTF takes almost four times as much computational power as the other NTF methods. Although sc-NTF requires initialization involving division, inverse tangents, and calculation of the histogram, the runtime is only slightly longer than that of f-NTF. Since the size of each matrix is different due to the required resolution, the computational power needed to update each matrix is not the same. Each matrix needs at least two contracted products and single division: $2 \times K \times L + J \times P$ for the channel matrix, $2 \times J \times L + K \times P$ for the frequency matrix, and $2 \times J \times K + L \times P$ for the time matrix. Thus, the results should depend on the numbers of occurrences of J, K, L, and P. In most cases, $J << K$ or L, which means that updating will probably take longer for the channel matrix, Q, than for the other two matrices. Another important point is that the burden of the built-in functions depends on the features of the LSI used to implement the NTF: Division generally needs a couple of instructions, but multiplication and addition usually need only one.

5.3 Comparison with DUET

An evaluation that compares sc-NTF with DUET has been carried out. Since a number of evolved versions of DUET have been proposed, we employed one of the most straightforward extensions of DUET called Multiple Sensor DUET (MENUET) [17]. Although the framework of MENUET requires the number of sources for the clustering at the end of the process, it is usually unknown in real world, particularly in applications such as Sound Zoom. Instead of providing the number of sources, which can be regarded as providing another piece of prior knowledge, we incorporated a spatial cue in the DUET system so that the closest centroid to the spatial cue can be extracted as a target centroid. The number of centroids is set to 18 to fully cover all the azimuths. Figures 6 and 7 show the comparison results between two different source separation techniques. In particular, the results of the two guitars give us a deep insight such that MENUET performs better than sc-NTF when separating such signals that have similar frequency characteristics, on the condition that the two sources are not coming from the same directions. The worse results of sc-NTF may be attributed to the nature of NTF that greedily exploits not only spatial information, but also time-frequency information for the approximation of the cost function. On the other hand, sc-NTF performs better in the case of separating the two percussions in spite of the close positions of the two instruments. This is due to the capability of NMF to capture repetitive structures of signals. Computational costs of MENUET can be seen in Figure 8. The number of iterations for clustering in MENUET is 30.

Figure 8 Comparison of computational cost. Runtime test for the three different NTFs. The runtime of sc-NTF includes the calculation of initialization, weighting function, and constraints.

6 Conclusion

We developed a new method of enhancing NTF performance by introducing weights on the NTF cost function, which is achieved by incorporating a spatial cue into the system. Two ways of incorporating a spatial cue into an 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. sc-NTF also showed competitive results against the variant of the DUET algorithm. Adaptation to a moving target and extension to the other state-of-the-art multichannel NMF will be subjects of future work.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The contributions of the second author, Axel Roebel, have been partly funded by the 3DTVS project in the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287674.

Author details
[1]Audio Technology Development Department, Sony Corporation, Tokyo 141-8610, Japan. [2]IRCAM-CNRS-UPMC UMR 9912, Paris 75004, France.

Received: 27 June 2013 Accepted: 10 March 2014
Published: 28 March 2014

References

1. DD Lee, HS Seung, Algorithms for non-negative matrix factorization, in *NIPS* (Denver, Colorado, USA, 27 November–2 December 2000)
2. C Févotte, Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition, in *Machine Audition: Principles, Algorithms and Systems* (IGI Global Press, Hershey, 2010)
3. M Nakano, H Kameoka, J Le Roux, Y Kitano, N Ono, S Sagayama, Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence, in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (Kittila, Finland, 29 August–1 September 2010), pp. 283–288
4. M Spiertz, V Gnann, Source-filter based clustering for monaural blind source separation, in *International Conference on Digital Audio Effects (DAFx-09)* (Como, Italy, 1–4 September 2009)
5. R Jaiswal, D Fitzgerald, D Barry, E Coyle, S Rickard, Clustering NMF basis functions using Shifted NMF for monaural sound source separation, in *ICASSP* (Prague Congress Center, Prague, Czech Republic, 22–27 May 2011), pp. 245–248
6. JM Becker, M Spiertz, V Gnann, A probability-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, 12–15 March 2012), pp. 99–106
7. P Smaragdis, GJ Mysore, Separation by "humming": user-guided sound extraction from monophonic mixtures, in *WASPAA* (New Paltz, NY, USA, 18–21 October 2009)
8. O Dikmen, AT Cemgil, Unsupervised single-channel source separation using Bayesian NMF, in *WASPAA* (New Paltz, NY, USA, 18–21 October 2009)
9. S Ewert, M Müller, Using score-informed constraints for NMF-based source separation, in *ICASSP* (Kyoto, Japan, 25–30 March 2012)
10. A Shashua, T Hazan, Non-negative tensor factorization with applications to statistics and computer vision, in *ICML, Volume 119* (ACM, New York, 2005), pp. 792–799
11. D FitzGerald, M Cranitch, E Coyle, Extended nonnegative tensor factorisation models for musical sound source separation. Comput. Intell. Neurosci. **2008**(Article ID 872425), 15 (2008). 10.1155/2008/872425
12. A Ozerov, C Févotte, R Blouet, J Durrieu, Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation, in *ICASSP* (Prague Congress Center, Prague, Czech Republic, 22–27 May 2011), pp. 257–260
13. Y Mitsufuji, A Roebel, Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge, in *ICASSP* (Vancouver, BC, Canada, 26–31 May 2013)
14. TO Virtanen, Monaural sound source separation by perceptually weighted non-negative matrix factorization. Technical Report, Tampere University of Technology, 2007
15. A Cichocki, R Zdunek, AH Phan, S Amari, *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation* (Wiley, Hoboken, 2009)
16. Ö Yilmaz, S Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
17. S Araki, R Mukai, H Sawada, S Makino, Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. Signal Process. **87**(8), 1833–1847 (2007)
18. E Vincent, S Arberet, R Gribonval, Underdetermined instantaneous audio source separation via local Gaussian modeling, in *ICA* (Paraty, Brazil, 15–18 March 2009), pp. 775–782
19. NQK Duong, E Vincent, R Gribonval, Spatial covariance models for under-determined reverberant audio source separation, in *WASPAA* (New Paltz, NY, USA, 18–21 October 2009), pp. 129–132
20. A Ozerov, C Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. on Audio Speech Lang. Process. **18**(3), 550–563 (2010)
21. NQ Duong, E Vincent, R Gribonval, Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Trans. on Audio Speech Lang. Process. **18**(7), 1830–1840 (2010)
22. S Arberet, A Ozerov, NQ Duong, E Vincent, R Gribonval, F Bimbot, P Vandergheynst, Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation, in *2010 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)* (IEEE, Renaissance Hotel, Kuala Lumpur, Malaysia, 10–13 May 2010), pp. 1–4
23. A Ozerov, E Vincent, F Bimbot, A general modular framework for audio source separation, in *Latent Variable Analysis and Signal Separation* (Springer, New York, 2010), pp. 33–40
24. A Ozerov, E Vincent, F Bimbot, A general flexible framework for the handling of prior information in audio source separation. IEEE Trans. on Audio, Speech Lang. Process. **20**(4), 1118–1133 (2012)
25. H Sawada, H Kameoka, S Araki, N Ueda, Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Trans. on Audio Speech Lang. Process. **21**(5), 971–982 (2013)
26. FitzD Gerald, M Cranitch, E Coyle, On the use of the Beta divergence for musical source separation, in *Proc. Irish Signals Syst. Conf. (ISCC)* (Galway, Ireland, 18–19 June 2008)
27. C Févotte, 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, Volume 6684* (Springer, New York, 2010), pp. 102–115
28. S Doclo, M Moonen, GSVD-based optimal filtering for single and multimicrophone speech enhancement. IEEE Trans. Signal Process. **50**(9), 2230–2244 (2002)
29. TJ Klasen, M Moonen, TV den Bogaert, J Wouters, Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction, in *ICASSP* (Pennsylvania Convention Center/Marriott Hotel, Philadelphia, PA, USA, 18–23 March 2005)
30. F Weninger, B Schuller, M Wöllmer, G Rigoll, Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory, in *ICASSP* (Prague Congress Center, Prague, Czech Republic, 22–27 May 2011), pp. 5840–5843
31. C Févotte, N Bertin, JL Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
32. N Bertin, C Févotte, R Badeau, A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription, in *ICASSP* (IEEE, Taipei, Taiwan, 19–24 April 2009), pp. 1545–1548
33. E Vincent, R Gribonval, C Févotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
34. In Signal Separation Evaluation Campaign (SiSEC 2008) (2008). http://sisec2008.wiki.irisa.fr/tiki-index.php.

doi:10.1186/1687-6180-2014-40
Cite this article as: Mitsufuji and Roebel: **On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization.** *EURASIP Journal on Advances in Signal Processing* 2014 **2014**:40.