Score-Informed Sparseness for Source Separation

Christian Rohlfing[1], Martin Spiertz[1], Volker Gnann[1]

[1] *Institut für Nachrichtentechnik, RWTH Aachen University, Aachen, Germany, Email: spiertz@ient.rwth-aachen.de*

Introduction

Audio source separation is a useful preprocessing step for remixing or transcription of music. It can be shown, that the separation quality increases, if the separation algorithm gets additional side information, e.g. the score of the current mixture [5]. In many cases the score of a musical piece is not available and has to be extracted by a professional musician or an automatic music transcription algorithm. To avoid both necessities, we will propose a source separation algorithm, which utilizes only the temporal activity (TA) of each instrument in the mixture. Compared to the whole score, this TA can be evaluated with much less experience in music transcription. To improve separation quality, the TA controls the sparsity of a non-negative tensor factorization. We will show, that for certain mixtures, this TA is a sufficient information for source separation. If TA is not sufficient, it can be utilized as a preprocessing step for further blind source separation algorithms.

Preliminaries

The spectrograms $\boldsymbol{X}_c \in \mathbb{R}_+^{K,T}$, obtained by the short-time Fourier transform of each channel signal $x_c(t)$ of the stereo mixture signal, are combined into the tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}_+^{K,T,2}$ with $c \in \{1,2\}$. K is the number of frequency bins and T the number of time-slots. $x_c(t)$ is the mixture of M mono sources.

Source separation with NTF

As described in [2], the Non-Negative Tensor Factorization (NTF) factorizes a tensor $\boldsymbol{\mathcal{V}}$ of size K-by-T-by-C in three non-negative matrices:

$$\boldsymbol{\mathcal{V}}(k,t,c) \approx \widetilde{\boldsymbol{\mathcal{V}}}(k,t,c) = \sum_{i=1}^{I} \boldsymbol{B}(k,i)\boldsymbol{G}(t,i)\boldsymbol{A}(c,i) \qquad (1)$$

where $\boldsymbol{B} \in \mathbb{R}_+^{K,I}$, $\boldsymbol{G} \in \mathbb{R}_+^{T,I}$ and $\boldsymbol{A} \in \mathbb{R}_+^{C,I}$. The index i denotes the i-th and I the total number of factorized components. I has to be set by the user.
The core of the factorization is a set of update rules which try to minimize a certain cost function.
Motivated by [3], we choose the Itakura-Saito (IS) divergence as the cost function to be minimized by the NTF:

$$d_{\mathrm{IS}}(\boldsymbol{\mathcal{V}} \mid \widetilde{\boldsymbol{\mathcal{V}}}) = \left|\left| \boldsymbol{\mathcal{V}} \oslash \widetilde{\boldsymbol{\mathcal{V}}} - \log(\boldsymbol{\mathcal{V}} \oslash \widetilde{\boldsymbol{\mathcal{V}}}) - 1 \right|\right|_1 \qquad (2)$$

and set $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{X}} \otimes \boldsymbol{\mathcal{X}}$ to the power spectrogram tensor. $\otimes$ and $\oslash$ denote elementwise multiplication and division.
In this paper, we use the respective multiplicative update rules extended to the NTF [1]. As the IS divergence is a limit case of the β-divergence ($\beta = 0$), we describe now the general multiplicative update rules of the β-divergence, proposed in [4] to ensure non-negativity of elements:

$$\mathbf{B}(k,i) \leftarrow \mathbf{B}(k,i)\frac{\sum_{t,c} \xi_1(k,t,c)\mathbf{G}(t,i)\mathbf{A}(c,i)}{\sum_{t,c} \xi_2(k,t,c)\mathbf{G}(t,i)\mathbf{A}(c,i)} , \qquad (3)$$

$$\mathbf{G}(t,i) \leftarrow \mathbf{G}(t,i)\frac{\sum_{k,c} \xi_1(k,t,c)\mathbf{B}(k,i)\mathbf{A}(c,i)}{\sum_{k,c} \xi_2(k,t,c)\mathbf{B}(k,i)\mathbf{A}(c,i)} , \qquad (4)$$

$$\mathbf{A}(c,i) \leftarrow \mathbf{A}(c,i)\frac{\sum_{k,t} \xi_1(k,t,c)\mathbf{G}(t,i)\mathbf{B}(k,i)}{\sum_{k,t} \xi_2(k,t,c)\mathbf{G}(t,i)\mathbf{B}(k,i)} . \qquad (5)$$

with $\xi_1(k,t,c) = \boldsymbol{\mathcal{V}}(k,t,c) \cdot \widetilde{\boldsymbol{\mathcal{V}}}^{\beta-2}(k,t,c)$ and $\xi_2(k,t,c) = \widetilde{\boldsymbol{\mathcal{V}}}^{\beta-1}(k,t,c)$.
$\boldsymbol{B}$, $\boldsymbol{G}$ and $\boldsymbol{A}$ are initialized with the absolute values of a normal-distribution. Afterwards sparsity is introduced to $\boldsymbol{G}$ as described in the following section.
The application of the NTF to $\boldsymbol{\mathcal{V}}$ (with $C = 2$) forms the three matrices $\boldsymbol{B}$, $\boldsymbol{G}$ and $\boldsymbol{A}$ as follows:
$\boldsymbol{B}$ consists of a set of frequency basis functions who are assumed to be valid for both stereo channels of the mixture signal, $\boldsymbol{G}$ contains the corresponding time envelopes and $\boldsymbol{A}$ the gain factors of each separated component i in each stereo channel c [2].
It can be shown, that the NTF is able to separate single notes or musical events from a spectrogram tensor [2], if the number of factorization-channels I is chosen large enough. For each event, the NTF assigns a column in $\boldsymbol{B}$, $\boldsymbol{G}$ and $\boldsymbol{A}$ corresponding to the event's frequency information or pitch, its temporal behaviour and its gain in each of the two stereo channels.

Score-informed sparseness

As a constraint to the separation process, we introduce sparseness to $\boldsymbol{G}$ as an initialization step prior to the factorization according to the temporal activity (TA) of each source signal. Therefore, a certain number of NTF-components $\boldsymbol{J}(m)$ ($\boldsymbol{J}(0) = 0$ with $I = \sum_{i=1}^{M} \boldsymbol{J}(m)$) is assigned to each source m.
For the purpose of this paper, it is sufficient to assume that the number of active instruments is non-decreasing. Each instrument has a starting time $\boldsymbol{T}(m)$. The extension to a more complex model with stop times of sources can be derived analogous.
Entries in columns of $\boldsymbol{G}$ are set to zero corresponding to the temporal activity of source m:

$$\boldsymbol{G}(t,i) = 0 \qquad \text{for } 1 \leq t < \boldsymbol{T}(m) , \qquad (6)$$

$$j(m) \leq i \leq j(m+1) - 1 . \qquad (7)$$

with $j(m) = 1 + \sum_{l=0}^{m-1} \boldsymbol{J}(l)$ being the first component assigned to source m. Due to the multiplicative update

rules, these entries remain zero during the factorization process.
This approach forces the NTF to learn frequency basis functions of source m and store it in one of the corresponding columns of $\boldsymbol{B}$.

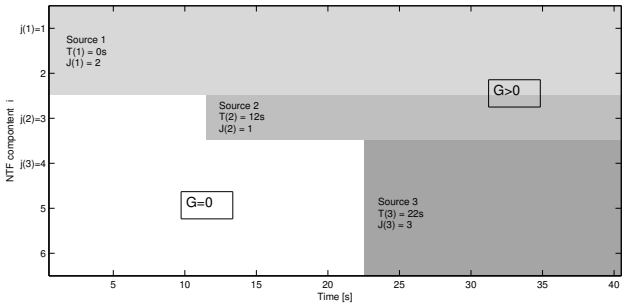

Figure 1: Sparse Matrix $\boldsymbol{G}$, $M = 3$ active sources.

Signal synthesis

The approximative tensors $\widetilde{\boldsymbol{\mathcal{S}}}_m$ of each source m are obtained by the assignment filter $\boldsymbol{\mathcal{H}}_m(k,t,c)$:

$$\boldsymbol{\mathcal{H}}_m(k,t,c) = \frac{\sum_{i=j(m)}^{j(m+1)-1} \boldsymbol{B}(k,i)\boldsymbol{G}(t,i)\boldsymbol{A}(c,i)}{\sum_{i=1}^{I} \boldsymbol{B}(k,i)\boldsymbol{G}(t,i)\boldsymbol{A}(c,i)} \qquad (8)$$

$\widetilde{\boldsymbol{\mathcal{S}}}_m$ is set to $\widetilde{\boldsymbol{\mathcal{S}}}_m(k,t,c) = \boldsymbol{\mathcal{H}}_m(k,t,c)\boldsymbol{\mathcal{X}}(k,t,c)$. The inverse short-time fourier transform of each channel spectrogram $\widetilde{\boldsymbol{S}}_{m,c}$ provides the estimated stereo signal $\tilde{s}_{m,c}(t)$.

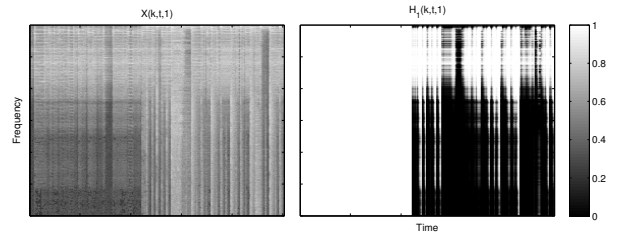

Figure 2: Spectrogram $\boldsymbol{\mathcal{X}}(k,t,1)$ of the mixture and assignment filter $\boldsymbol{\mathcal{H}}_1(k,t,1)$ of the guitar signal of the intro of "Money for Nothing" by Dire Straits. In the beginning of the intro, the guitar plays by itself and is later joined by drums and bass. This is indicated by the structure of $\boldsymbol{\mathcal{H}}_1$.

Temporal activity estimation

To approach a blind source separation scenario, a temporal activity estimation method is proposed:
For each column of the mixture spectrogram, an approximation of the neighboring columns is calculated by the NTF-algorithm (here with $\beta = 1$, $I = 1$). Afterwards, the cost-function $d_{\mathrm{div}}(\boldsymbol{\mathcal{V}} \mid \widetilde{\boldsymbol{\mathcal{V}}}) = \left|\left|\boldsymbol{\mathcal{V}} \otimes \log(\boldsymbol{\mathcal{V}} \oslash \widetilde{\boldsymbol{\mathcal{V}}}) - \boldsymbol{\mathcal{V}} + \widetilde{\boldsymbol{\mathcal{V}}}\right|\right|_1$ [3] is evaluated. As shown in figure 3, this signal yields a good estimation of changes in the complexity of the mixture signal over time and shows strong correlation with its semantics. These values are quantized with a small number of quantization levels. Derivation of the quantized signal gives information about the starting times $\boldsymbol{T}(m)$.

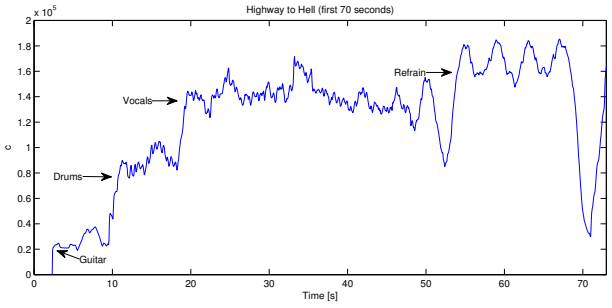

Figure 3: The first 70 seconds of the song "Highway to Hell" performed by AC/DC. At $\boldsymbol{T}(1) = 0$s the guitar begins to play, at $\boldsymbol{T}(2) = 9.5$s the drums and at $\boldsymbol{T}(3) = 18$s the vocals which are backed up by a choir in the refrain at $\boldsymbol{T}(4) = 53$s. Even the four repetitions of the chorus "highway to hell" are visible as four peaks in the refrain

Conclusions

We proposed an algorithm for source separation which utilizes the temporal activity of the sources as a side information. This algorithm shows very good separation results in practise but has the disadvantage, that it cannot separate sources who always appear at the same time. In this case, the algorithm can be used as a preprosessing step for further blind source separation.
The enhancement of the TA estimation algorithm to increase its robustness and the extension to estimate $\boldsymbol{J}(m)$ are topic of current research.

References

[1] Andrzej Cichocki, Anh Huy Phan, and Rafal Zdunek. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation; electronic version.* Wiley, Chichester, 2009.

[2] Derry FitzGerald, Matt Cranitch, and Eugene Coyle. Non-negative tensor factorisation for sound source separation. In *Proceedings of Irish Signals and Systems Conference*, 2005.

[3] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis. *Neural Computation*, 21(3):793–830, Mar. 2009.

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

[5] John Woodruff, Bryan Pardo, and Roger B. Dannenberg. Remixing stereo music with score-informed source separation. In *ISMIR*, pages 314–319, 2006.