2009-01-01

Musical Sound Source Separation using Extended Tensor Decompositions

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

Funder: Science Foundation Ireland

Musical Sound Source Separation using Extended Tensor Decompositions

Derry Fitzgerald
Dublin Institute of Technology, derry.fitzgerald@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.

Musical Sound Source Separation using extended tensor decompositions

Derry FitzGerald[†]

†Audio Research Group, Dublin Institute of Technology
Kevin St., Dublin 2, Ireland
Email: derry.fitzgerald@dit.ie

Abstract—Recently, tensor decompositions have found use in sound source separation. In particular, non-negative tensor decompositions have received a lot of attention due to their ability to decompose audio spectrograms into meaningful "parts" such as individual notes. Extensions to the basic non-negative tensor factorisation framework allow the incorporation of additional constraints, such as shift-invariance in both frequency and time. This enables the factorisations to capture more complex structures than individual notes, such as individual sources playing different pitches and time-evolving instrument timbres. Further music specific constraints such as harmonicity and source-filter modeling have been shown to improve separation performance for musical signals. Other recent advances also allow the incorporation of Bayesian priors into these models, thereby further improving the separations obtained.

1. Introduction

Recently, non-negative matrix factorisation (NMF) and non-negative tensor factorisation (NTF) techniques have been the focus of much attention as a means of carrying out musical sound source separation from single and multichannel mixtures [1, 2]. Standard NMF and NTF decompositions return basis functions which correspond to meaningful parts such as notes played by a pitched instrument. However, clustering these basis functions to their respective sources proved difficult. Later, shift invariant approaches, which could model a given pitched instrument with a single basis function which was then shifted in frequency to approximate different notes played by the instrument, were developed to overcome this limitation [3, 4].

While sucessful in overcoming the problem of clustering the basis functions, these techniques were not without their drawbacks. In particular, shift-invariance in frequency necessitated the use of a time-frequency transform with log-frequency resolution such as the Constant Q Transform [5]. As the mapping from a log-frequency domain to a linear frequency domain is an approximate mapping, this caused problems with the resynthesis of the separated sources. This problem could be ameliorated somewhat by incorporating the mapping into the signal model and optimising for reconstruction error in the linear frequency domain [6].

Further, using a single basis function to model an instrument is only a valid approximation over a limited pitch range, and in practice, the timbre of pitched instruments changes with frequency. In an attempt to overcome this, Virtanen and Klapuri proposed incorporating a source-filter model [7], which allowed the timbre to change by using a fixed filter to approximate the resonant structure of the instrument.

Finally, none of the above techniques constrained the basis functions of pitched instruments to be harmonic. An inital attempt to include such a constraint was made by Raczynski et al [8], who zeroed the basis functions in regions where no harmonic activity was expected. An additive synthesis approach to incorporating harmonicity was later proposed in [9]. The addition of harmonicity was also found to constrain the optimisation problem sufficiently to allow the simultaneous separation of pitched and unpitched sources, and the resulting model is described in greater detail in Section 2.

The rest of the paper is organised as follows: Section 1.1 describes the notation conventions used throughout the paper, while section 2 describes a generalised tensor factorisation model capable of separating mixtures of both pitched and unpitched sources. Section 3 describes Bayesian extensions to this model, while section 4 describes application of the algorithm to a number of real-world situations. Finally, section 5 provides some conclusions.

1.1. Notation

For the rest of the paper, all tensors, regardless of the number of dimensions, are signified by the use of caligraphic letters such as $\mathcal{A}$. $\langle \mathcal{AB} \rangle_{\{a,b\}}$ 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 [10]. For ease of notation, as all tensors are now instrument or source specific, the subscripts are implicit in all tensors within summations. Elementwise multiplication is denoted by $\otimes$ and all division is taken as elementwise.

2. Source-Filter Sinusoidal Non-negative Tensor Factorisation

Given an r-channel mixture of pitched and unpitched musical instruments, magnitude spectrograms are obtained

for each channel, resulting in $\mathcal{X}$, an $r \times n \times m$ tensor with n the number of frequency bins and m the number of time frames. The tensor is then modelled as:

$$\mathcal{X} \approx \hat{\mathcal{X}} = \sum_{k=1}^{K} \mathcal{G} \circ \langle\langle\langle \mathcal{F}\mathcal{H}\rangle_{\{2,1\}}\mathcal{W}\rangle_{\{3,1\}}\langle\mathcal{S}\mathcal{P}\rangle_{\{2,1\}}\rangle_{\{2:3,1:2\}} + \sum_{l=1}^{L} \mathcal{M} \circ \langle\mathcal{B}\langle\mathcal{C}\mathcal{Q}\rangle_{\{1,1\}}\rangle_{\{2,1\}} \quad (1)$$

where pitched instruments are modelled by the first right-hand side term and unpitched or percussion instruments by the second term. K denotes the number of pitched instruments and L denotes the number of unpitched instruments. The individual elements of the model are described below.

$\mathcal{G}$ 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 main diagonal contains 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 of size $n \times z_k \times h_k$ where z_k is the number of allowable notes and h_k the number of harmonics used to model the kth instrument. Here $\mathcal{H}(:, i, j)$ contains the frequency spectrum of a sinusoid with frequency equal to the jth harmonic of the ith note. $\mathcal{H}$ remains fixed during optimisation. $\mathcal{W}$ 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. This allows the model to capture evolution in the harmonic weights with time.

In the case of the unpitched instruments, $\mathcal{M}$ is a tensor of size r containing the gains of an unpitched instrument in each channel. $\mathcal{B}$ is of size $n \times q_l$ and contains a set of frequency basis functions which model the temporal evolution of the timbre of the unpitched instrument where q_l is the number of translations in time used to model the lth instrument. $\mathcal{C}$ is a tensor of size m which contains the activations of the lth instrument, and $\mathcal{Q}$ is a translation tensor of size $m \times q_l \times m$ used to shift the activations in $\mathcal{C}$ in time.

As opposed to the model presented in [9], here each instrument can have its parameters set individually, such as the number of harmonics or the number of allowable notes. For example, a flute can be modelled with fewer harmonics than a piano, and so the model parameters can be adjusted accordingly. This increased flexibility can be leveraged to improve the separations obtained from the algorithm.

Once a suitable metric for measuring reconstruction of the original data, such as the generalised Kullback-Liebler divergence, is chosen, iterative update equations can be derived for each of the model variables. These update equations take the form

$$\mathcal{R} = \mathcal{R} \otimes \frac{\nabla^{-}_{\mathcal{R},D(\mathcal{X}\|\hat{\mathcal{X}})}}{\nabla^{+}_{\mathcal{R},D(\mathcal{X}\|\hat{\mathcal{X}})}} \quad (2)$$

where $\mathcal{R}$ represents a given variable in the model to be updated, $D\left(\mathcal{X} \parallel \hat{\mathcal{X}}\right)$ denotes the reconstruction metric, and where $\nabla^{-}_{\mathcal{R},D(\mathcal{X}\|\hat{\mathcal{X}})}$ and $\nabla^{+}_{\mathcal{R},D(\mathcal{X}\|\hat{\mathcal{X}})}$ represent the negative and the positive parts respectively of the partial derivative of the reconstruction metric with respect to $\mathcal{R}$.

3. Bayesian Extensions

Recently Virtanen et al. have proposed a number of Bayesian extensions to NMF for the purposes of audio separation [11]. This work focused on the use of gamma priors to incorporate constraints such as prior knowledge of the frequency characteristics of the sources and temporal continuity on the note activations. These additional terms are added directly to the chosen reconstruction metric.

Priors over the unpitched frequency basis functions $\mathcal{B}$ were derived by Virtanen et al by assuming that each entry of each prior is independently drawn from a Gamma distribution, where the distribution was defined as:

$$G(y; a, b) = y^{a-1} b^{-a} e^{-y/b} / \Gamma(a) \quad (3)$$

where $\Gamma(a)$ is the gamma function. In the case of the frequency basis functions for unpitched instruments this can be expressed as

$$\begin{aligned} p(\mathcal{B}(v,r)) &= G(\mathcal{B}(v,r); \alpha_{v,r}, \beta^{-1}_{v,r}) \\ &= \mathcal{B}(v,r)^{\alpha_{v,r}-1} \beta^{\alpha_{v,r}}_{v,r} e^{-\mathcal{B}(v,r)\beta_{v,r}} / \Gamma(\alpha_{v,r}) \end{aligned} \quad (4)$$

where v denotes the vth frequency bin, r denotes the rth of q_l frames used to model the temporal evolution of the source. The hyperparameters $\alpha_{v,r}$ and $\beta_{v,r}$ can be chosen independently for each source, and a simple interpretation of $\beta^{-1}_{v,r}$ is as a set of weights which describe the typical frequency content of a given source. For example, $\beta^{-1}_{v,r}$ could be a typical frequency spectrogram of an unpitched instrument such as a hi-hat.

Similar priors can easily be incorporated for the harmonics weights $\mathcal{W}$ and the formant filter $\mathcal{F}$, thereby allowing additional knowledge of the sources to be incorporated in an intuitive manner. In effect, the update equations are modified to include a set of weights which point the updates towards the source known to be present. The modified update equations including the gamma prior take the form

$$\mathcal{R} = \mathcal{R} \otimes \frac{(\alpha - 1)/\mathcal{R} + \nabla^{-}_{\mathcal{R},D(\mathcal{X}\|\hat{\mathcal{X}})}}{\beta + \nabla^{+}_{\mathcal{R},D(\mathcal{X}\|\hat{\mathcal{X}})}} \quad (5)$$

where α and β are tensors containing the hyperparmeters for the gamma distribution. In practice, all elements in α are set to 1.

Virtanen et al. encourage temporal continuity on the basis function activations through the use of a gamma chain. Taking the activations of the pitched sources as an example,

this chain is constructed through the use of an auxiliary tensor $\mathcal{Z}$ of size $z_k \times m + 1$, which is defined as follows:

$$\begin{aligned} \mathcal{Z}(i,1) &\sim G(\mathcal{Z}(i,1); a+1, (ab)^{-1}) \\ \mathcal{S}(i,\tau) \mid \mathcal{Z}(i,\tau) &\sim G(\mathcal{S}(i,\tau); a, (\mathcal{Z}(i,\tau)a)^{-1}) \\ \mathcal{Z}(i,\tau) \mid \mathcal{S}(i,\tau) &\sim G(\mathcal{Z}(i,\tau+1); a+1, (\mathcal{S}(i,\tau)a)^{-1}) \end{aligned} \quad (6)$$

where τ indicates the time index in frames and lies between 1 and m, and i indexes over $1 : z_k$. In this context a acts as a coupling parameter between frames, and larger values of a result in more strongly coupled adjacent frames.

It should be noted that this approach can easily be adapted to deal with other forms of continuity. For example, it would be expected that a formant filter $\mathcal{F}$ would vary smoothly with frequency, and so continuity in frequency would be of benefit in this situation. Similarily, continuity can be imposed on the harmonic weights $\mathcal{W}$, either between the strengths of the partial, which would correspond to assuming spectral smoothness, a principal found to be useful in music transcription [12] or by imposing continuity on the temporal evolution of the partials.

The update equations incorporating the gamma chain take the general form

$$\mathcal{R} = \mathcal{R} \otimes \frac{2a/\mathcal{R} + \nabla^{-}_{\mathcal{R},D(\mathcal{X}\|\hat{\mathcal{X}})}}{a\mathcal{T} + \nabla^{+}_{\mathcal{R},D(\mathcal{X}\|\hat{\mathcal{X}})}} \quad (7)$$

where $\mathcal{T}$ is a tensor which depends on the auxilliary tensor $\mathcal{Z}$ for the gamma chain. In the case of $\mathcal{S}$, updates for $\mathcal{Z}$ are given by:

$$\mathcal{T} = \mathcal{Z}(i, 1:m) + \mathcal{Z}(i, 2:m+1) \quad (8)$$

and

$$\mathcal{Z}(i,\tau) = \begin{cases} 1/(\mathcal{S}(i,1)+b), & \tau = 1 \\ 2/(\mathcal{S}(i,\tau)+\mathcal{S}(i,\tau-1)), & 1 < \tau < m+1 \\ 1/\mathcal{S}(i,\tau) & \tau = m+1 \end{cases} \quad (9)$$

Similar auxiliary tensors can be defined for both $\mathcal{F}$ and $\mathcal{W}$, though it should be noted that for $\mathcal{F}$ the gamma chain will be defined only for the main diagonal.

4. Applications

This section presents a number of real-world applications of the separation algorithm, with particular reference to the separation of pitched instruments from percussion and noise. Figure 1 shows an excerpt from "Rosanna" by Toto, along with the separated pitched and drum instruments respectively. It can be seen that the separation of the drum sounds is quite clean, with little or no evidence of pitched instruments, while the prominent transients of the drum sounds are not evident in the pitched separation. On listening, little or no evidence of the pitched sound can be heard in the separated drums, while there is still some evidence of the drums in the pitched signal, though at a much reduced volume. This appears to be because the pitched separation also appears to capture relatively long tailed reverberations of the drum events, in part due to the continuity constraint imposed on the pitched instruments.

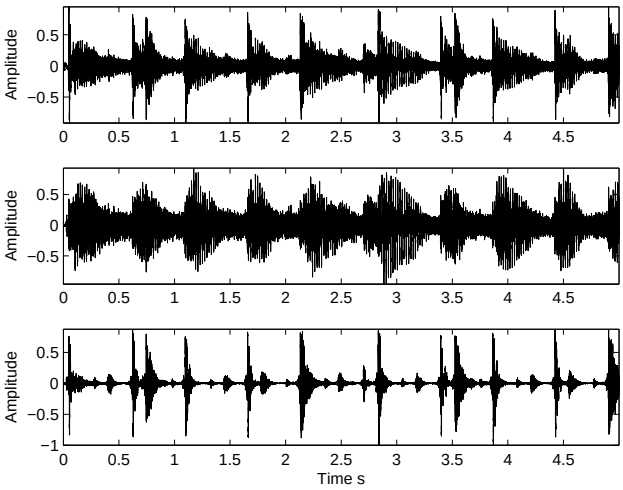

Figure 1: *(a) Original excerpt from "Rosanna" by Toto, (b) Separated pitched instruments, (c) Separated drums*

This can be seen more clearly in Figure 2, where a drum intro from the same song is passed through the same algorithm. The drum instruments can be seen to have been clearly captured, while the "pitched" separation consists mainly of more continuous noise. On listening to the "pitched" separation, it sounds like a drum signal passed through a reverb, and by adjusting the amplitudes of the "pitched" and unpitched parts, the drum ontro can be made to sound like it is a much bigger room.

Finally, Figure 3 shows an example of using the algorithm to denoise a degraded 78rpm recording. It can be seen that the noticeable transients have been completely removed from the recording, as well as much constant noise that is not visable on the waveform. On listening, the sound quality has been improved considerably, with the noise only audible in quieter passages of the recording.

5. Conclusions

A brief overview of NTF-based approaches to musical source separation has been given. An extended NTF-based approach was then presented which incorporated both source-filter modelling and harmonicity constraints. Following this, a number of Bayesian extensions to this model, using Gamma priors in particular, were discussed. These extensions can result in improved separation performance, particularly with regards to continuity constraints. A number of separation examples using these extensions were then shown, demonstrating the utility of these techniques in real world settings. However, a problem with

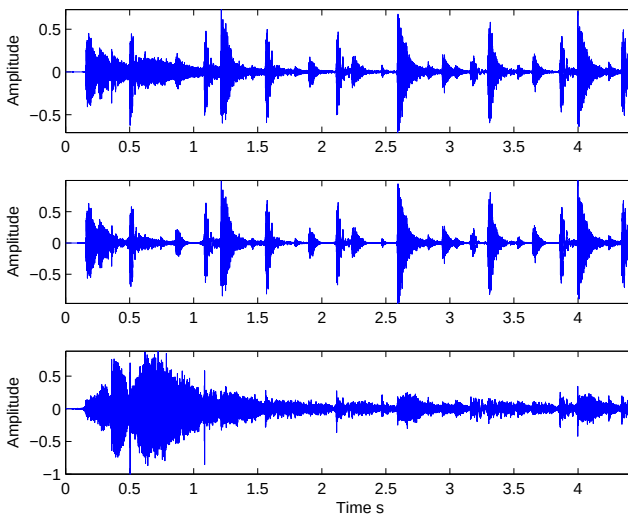

Figure 2: *(a) Drum intro from "Rosanna" by Toto, (b) Separated drum instruments, (c) Separated "reverb"*

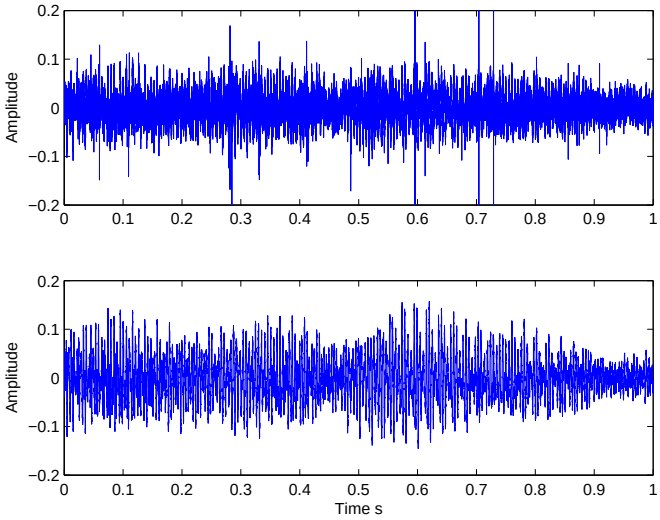

Figure 3: *(a) Original noisy signal, (b) Denoised signal*

these techniques is that they are computationally intensive. Future work will concentrate on improving the speed of these algorithms.

Acknowledgments

The author was supported in this research under Science Foundation Ireland's Stokes Lectureship scheme.

References

[1] M.N. Schmidt and M. Mørup, "Non-negative matrix factor 2-D deconvolution for blind single channel source separation", Int'l. Conf. on Independent Component Analysis and Blind Signal Separation, Lecture Notes in Computer Science, vol 3889 pp. 700-707, 2006

[2] T. Virtanen, "Sound Source Separation in Monaural Music Signals", Tampere University of Technology, 2006.

[3] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound Source Separation using shifted Non-negative Tensor Factorisation", IEEE International Conference on Acoustics, Speech and Signal Processing, 2006 (ICASSP2006), Toulouse France.

[4] M.N. Schmidt, M. Mørup, "Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation", Proceedings of the International Conference on Independent Component Analysis, 2006.

[5] J.C. Brown, "Calculation of a Constant Q Spectral Transform" J. Acoust. Soc. Am. 89 425-434. 1991

[6] D. FitzGerald, "Blind Source Separation and automatic transcription of music using tensor decompositions", 6th International Congress on Industrial and Applied Mathematics, Zurich, 2007

[7] T. Virtanen, A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization", Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop, 2006 (extended abstract).

[8] S. Raczynski, N. Ono, and S. Sagayama, "Multipitch Analysis with Harmonic Nonnegative Matrix Approximation", Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria, 2007.

[9] 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

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

[11] T. Virtanen, A. Cemgil and S. Godsill, "Bayesian Extensions to Non-negative Matrix Factorisations for Audio Signal Modelling", Proc. IEEE Conference on Acoustics, Speech and Language Processing, 2008.

[12] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and Inharmonic nonnegative matrix factorisation for polyphonic pitch transcription", Proc. IEEE Conference on Acoustics, Speech and Language Processing, 2008.