2008-01-01

Musical Source Separation using Generalised Non-negative Tensor Factorisation Models

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

Derry Fitzgerald
Cork Institute of Technology

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

Musical Source Separation using Generalised Non-Negative Tensor Factorisation models

Derry FitzGerald DERRY.FITZGERALD@CIT.IE
Matt Cranitch MATT.CRANITCH@CIT.IE
Dept. of Electronic Engineering, Cork Institute of Technology, Rossa Avenue, Cork, Ireland

Eugene Coyle EUGENE.COYLE@DIT.IE
School of Electrical Engineering Systems, Dublin Institute of Technology, Kevin Street, Dublin, Ireland

Keywords: Musical Sound Source Separation, Tensor Factorisation

Abstract

A shift-invariant non-negative tensor factorisation algorithm for musical source separation is proposed which generalises previous work by allowing each source to have its own parameters rather a fixed set of parameters for all sources. This allows independent control of the number of allowable notes, number of harmonics and shifts in time for each source. This increased flexibility allows the incorporation of further information about the sources and results in improved separation and resynthesis of the separated sources.

1. Introduction

In recent years, machine learning techniques such as shift-invariant non-negative matrix and tensor factorisation have received much attention as a means of separating sound sources from single and multichannel mixtures (Mørup M., 2006; Virtanen, 2006). Using shift-invariance in frequency allows a pitched musical instrument to be modelled as a single frequency basis function which is translated up and down in frequency to model different notes played by an instrument, while shift-invariance in time allows the temporal evolution of a sources timbre to be captured.

More recently, improvements have been made to these models by imposing harmonicity constraints on pitched instruments (FitzGerald et al., 2008). This was done by using an additive synthesis model where a pitched instrument is modelled by a set of harmonic weights. A source-filter model was also incorporated to allow the timbre of instruments to change with pitch, resulting in improved separations. Further, the use of harmonicity constraints restricts the decompositions sufficiently to allow simultaneous separation of pitched and unpitched instruments. However, for best results, the model now requires an estimate of the pitch of the lowest note played by each pitched instrument.

However, a limitation of shift-invariant models to date is that all the sources have to have the same parameters. For example, the number of allowable notes or shifts in frequency is the same for all pitched instruments. However, the range of notes played in a given piece will vary with the instrument. It can be seen that being able to set the number of notes for each instrument individually will reduce the possibilities for error in the separation. This would be of particular use in score-assisted separation, such as proposed in (Woodruff et al., 2006). Further, the number of harmonic weights required to model a given source varies. For example, modelling a flute will typically require less harmonics than a piano or violin. Therefore, the ability to vary the parameters for each source individually would be advantageous, and would help to improve the separations obtainable using shift-invariant factorisation models.

2. Generalised Factorisation Models

In the following, $\langle \mathcal{AB} \rangle_{\{a,b\}}$ denotes contracted tensor multiplication of $\mathcal{A}$ and $\mathcal{B}$ along the dimensions a and b of $\mathcal{A}$ and $\mathcal{B}$ respectively. Outer product multiplication is denoted by $\circ$. Indexing of elements within a tensor is notated by $\mathcal{A}(i, j)$ as opposed to using subscripts. This notation follows the conventions used in the Tensor Toolbox for Matlab, which was used to implement the following algorithm (Bader & Kolda, 2007). For ease of notation, as all tensors are now instrument-specific, the subscripts are implicit in all

tensors within summations.

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

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

with $\mathcal{R} = \langle \mathcal{F}\mathcal{H} \rangle_{\{2,1\}}$ and where the first right-hand side term models pitched instruments, and the second unpitched or percussion instruments. K denotes the number of pitched instruments and L denotes the number of unpitched instruments.

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

For 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 evolution of the timbre of the unpitched instrument with time 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 translate the activations in $\mathcal{C}$ in time.

A suitable metric for measuring reconstruction of the original data is the generalised Kullback-Leibler divergence proposed for use in non-negative matrix factorisation by (Lee & Seung, 1999):

$$D\left(\mathcal{X} \parallel \hat{\mathcal{X}}\right) = \sum \mathcal{X} \log \frac{\mathcal{X}}{\hat{\mathcal{X}}} - \mathcal{X} + \hat{\mathcal{X}} \quad (2)$$

Using this measure, iterative update equations can be derived for each of the model variables.

The model has been used to separate stereo mixtures containing both pitched and unpitched instruments, and the ability to set the parameters independently has been observed to improve the separation results obtained in comparison to the situation where the same parameters are used for all instruments.

3. Conclusions

An algorithm was proposed which generalises previous work in shift-invariant non-negative tensor factorisation algorithms by allowing each source to have its own parameters, such as note range, number of harmonics and time shifts. This increased flexibilty has been observed to result in improved musical source separation performance for mixtures of pitched and unpitched instruments.

Acknowledgements

This research was supported by Enterprise Ireland.

References

Bader, B., & Kolda, T. (2007). Matlab tensor toolbox version 2.2.

FitzGerald, D., Cranich, M., & Coyle, E. (2008). Extended non-negative tensor factorisation models for musical sound source separation. *to appear in Computational Intelligence and Neuroscience.*

Lee, D., & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorisation. *Nature*, *401*, 788–791.

Mørup M., Schmidt, M. (2006). *Sparse non-negative tensor 2d deconvolution (sntf2d) for multi channel time-frequency analysis* (Technical Report). Technical University of Denmark.

Virtanen, T. (2006). *Sound source separation in monaural music signals.* Doctoral dissertation, Tampere University of Technology.

Woodruff, J., Pardo, B., & Dannenberg, R. (2006). Remixing stereo music with score-informed source separation. *Proceedings of the 7th International Conference on Music Information Retrieval.*