2007-01-01

Blind Source Separation and Automatic Transcription of Music Using Tensor Decompositions

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

Blind Source Separation and Automatic Transcription of Music Using 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.

Blind source separation and automatic transcription of music using tensor decompositions

Derry FitzGerald[*1]

[1] Dept. of Electronic Engineering, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland

Recent advances in the use of tensor decompositions for the analysis of music are described. In particular, the use of such decompositions for sound source separation and the automatic transcription of music are explored.

1 Background

Recently, factorisation-based approaches to sound source separation and automatic transcription of pitched musical instruments, such as non-negative matrix factorisation (NMF) [1], have received much attention [2, 3]. NMF attempt to factorise a data matrix $\mathbf{X}$ into matrix factors $\mathbf{A}$ and $\mathbf{S}$ such that $\mathbf{X} \approx \mathbf{AS}$, where $\mathbf{X}$ is an n x m matrix, $\mathbf{A}$ is an n x r matrix, and $\mathbf{S}$ is an r x m matrix, with r smaller than n or m. A wide variety of cost functions exist for tackling this problem, although in the context of music, functions which encourage sparse solutions are prefered. This is because musical signals containing only pitched instruments are harmonic in nature, and outside these harmonic regions there is little or no frequency activity. Further, a given note is not played all the time and so its activations are sparse.

In the context of single channel sound source separation, the single channel is typically transformed into a time-frequency representation such as a magnitude or power spectrogram. On factorising the spectrogram, the columns of $\mathbf{A}$ contain frequency basis functions, and $\mathbf{S}$ contains a set of amplitude envelopes associated with the frequency basis functions. A problem with using these methods for sound source separation is that each frequency basis function typically describes only a single note. As most musical signals involve changes in pitch, this restriction means that some form of grouping must be carried out after using the above methods to obtain separated sources which change in pitch [2, 3]. However, it has been observed that it is difficult to obtain the correct grouping of basis functions,resulting in serious limitations on the usefulness of these algorithms to-date. Further, using a single basis function to represent a note does not reflect the fact that the spectral content of real instruments evolves over time. It can therefore be seen that extended models are needed in order to overcome these problems.

2 Shifted 2D Non-negative Tensor Factorisation

The approach taken was the extension of NMF approaches with additional constraints and parameters to deal more realistically with musical signals. In particular, NMF was extended to a non-negative tensor factorisation (NTF) approach, allowing the analysis of multi-channel signals and the concept of shift invariance was also incorporated. If a time-frequency representation with log-frequency resolution, such as the Constant Q transform [4], is used then nearby notes of an instrument can be modelled as translations of another note. Therefore, an instrument can be modelled by a single basis function, and different notes modelled as translations of the instrument basis function. This is done by incorporating shift invariance in the frequency basis functions. Similarly, extending NTF to be invariant to time shifts allowed the algorithm to model the temporal evolution of instrument spectra as each instrument basis function will have an associated spectrum for each translation in time[5].

For the remainder of this report, the following conventions are used. Tensors are denoted by calligraphic uppercase letters, such as $\mathcal{T}$. $\langle \mathcal{AB} \rangle_{\{a,b\}}$ denotes contracted tensor multiplication of $\mathcal{A}$ and $\mathcal{B}$ along the dimensions a and b of $\mathcal{A}$ and $\mathcal{B}$ respectively. For simplicity of notation, we also adopt the convention that : k refers to the tensor slice associated with the k^{th} source or instrument. The shift invariant 2D non-negative tensor factorisation model can then be defined as:

$$\mathcal{X} \approx \sum_{k=1}^{K} \langle \mathcal{G}_{:k} \langle \langle \mathcal{T} \mathcal{A}_{:k} \rangle_{\{3,1\}} \langle \mathcal{S}_{:k} \mathcal{P} \rangle_{\{3,1\}} \rangle_{\{2:4,1:3\}} \rangle_{\{2,2\}} \tag{1}$$

where, in the context of musical signals, $\mathcal{X}$ is a tensor of size r x n x m,containing the magnitude spectrograms of each channel of the signal. $\mathcal{G}$ is a tensor of size r x K, containing the gains of each of the K sources in each of the r channels. $\mathcal{T}$ is an n x z x n translation tensor, which translates the instrument basis functions in $\mathcal{A}$ up or down in frequency, thereby approximating different notes played by a given source. $\mathcal{A}$ is a tensor of size n x K xp, where p is the number of translations across time. $\mathcal{S}$ is a tensor of size z x K x m containing the activations of the translations of $\mathcal{A}$ which indicate when a given note played by a given instrument occurs, thereby generating a transcription of the signal. $\mathcal{P}$ is an m x p x m translation tensor which translates

* Corresponding author: e-mail: derry.fitzgerald@cit.ie, Phone: +00353 21 432 6881,

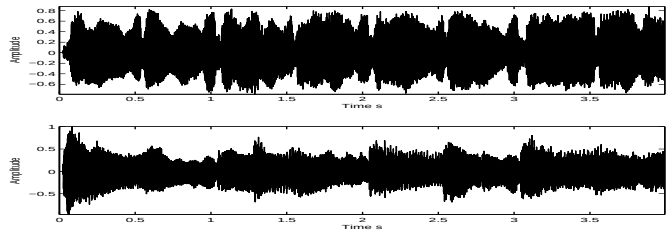

Fig. 1 Stereo mixture of flute, viola and piano.

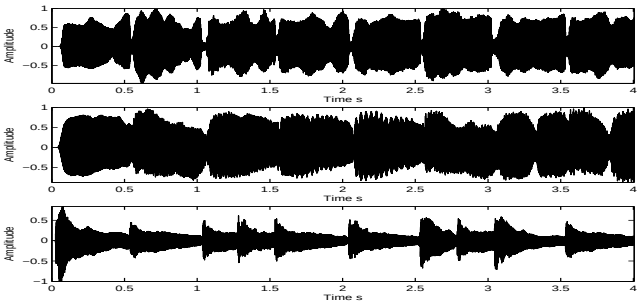

Fig. 2 Original waveforms of flute, viola and piano respectively.

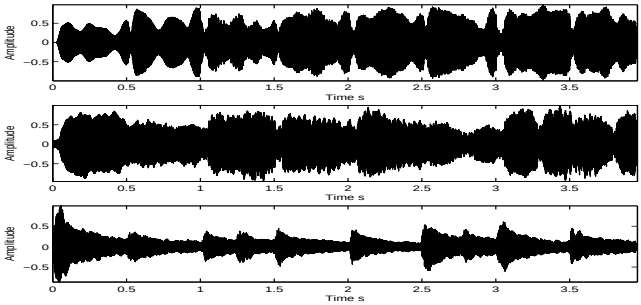

Fig. 3 Separated waveforms of flute, viola and piano respectively.

the time activation functions contained in $\mathcal{S}$ across time, and so allow time-varying source or instrument spectra. A problem with the algorithm is that for signal resynthesis, the recovered spectrograms must be mapped from the log-frequency to the linear-frequency domain. As the mapping matrix is not square, no true inverse exists. However, a way of overcoming this is to incorporate the mapping into the model by replacing $\mathcal{T}$ with $\langle \mathcal{C}\mathcal{T} \rangle_{\{2,1\}}$ where $\mathcal{C}$ maps from log to linear domain.

Separated source spectrograms can then be recovered by contracted product multiplication of the tensor slices of a given source. The algorithm was implemented in Matlab with the tensor classes for Matlab [6], using the generalised Kullback-Liebler divergence as a cost function. The method's utility can be seen in that, apart from separating the instruments, it generates an instrument model, a transcription of the notes played by each instrument, and an estimate of the mixing matrix.

The algorithm was tested on both synthetic signals and real world recordings and was found to be capable of separating simple music such as three to four instruments playing melodies, or a lead instrument accompanied by a single instrument, such as fiddle accompanied by guitar. Fig. 1 shows a three instrument stereo mixture, with the original signals in fig. 2, and the separated sources in fig. 3. It can be seen that the algorithm has successfully separated the sources. However, performance degrades with increasing numbers of instruments. Further, the algorithm was found to be sensitive to the number of allowable shifts in frequency. To obtain good results, it was found that the number of allowable shifts had to be close to the pitch range of the instrument with the widest pitch range.

A simplified version of the algorithm, where the instrument basis functions are fixed pre-learned harmonic spectra is capable of giving a reliable transcription of all notes played in a piece of music [7]. This simplified algorithm can be implemented on a frame-by-frame basis and is sufficiently fast to run in real time, thereby allowing real time transcription of audio signals, although at the cost of not being able to allocate the notes played to their instruments. Further, the simplified algorithm is not sensitive to the allowable number of shifts, and gives good results for polyphonies of up to six.

Acknowledgements This research was funded by the Embark Initiative and by Enterprise Ireland.

References

[1] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization", Adv. Neural Info. Proc. Syst. 13, 556-562 (2001).
[2] M. Casey and A. Westner "Separation of Mixed Audio Sources By Independent Subspace Analysis" in Proc. Of ICMC 2000, pp. 154-161, Berlin, Germany.
[3] T. Virtanen, "Sound Source Separation Using Sparse Coding with Temporal Continuity Objective", Proc. of International Computer Music Conference (ICMC2003), Singapore, 2003.
[4] J. Brown, "Calculation of a Constant Q Spectral Transform" J. Acoust. Soc. Am. 89 425-434, 1991
[5] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted 2D Non-negative Tensor Factorisation", Proceedings of the Irish Signals and Systems Conference, Dublin, June 2006.
[6] Tensor Classes for Matlab, available at http://csmr.ca.sandia.gov/ tgkolda/TensorToolbox/
[7] D. FitzGerald, M. Cranitch, and E. Coyle, "Generalised Prior Subspace Analysis for Polyphonic Pitch Transcription", Proceedings of the 8th International Conference on Digital Audio Effects (DAFX05), Madrid September 2005.