SHIFT-VARIANT NON-NEGATIVE MATRIX DECONVOLUTION FOR MUSIC TRANSCRIPTION

Holger Kirchhoff, Simon Dixon, Anssi Klapuri

Queen Mary University of London
Centre for Digital Music
Mile End Road, London E1 4NS, UK

ABSTRACT

In this paper, we address the task of semi-automatic music transcription in which the user provides prior information about the polyphonic mixture under analysis. We propose a non-negative matrix deconvolution framework for this task that allows instruments to be represented by a different basis function for each fundamental frequency ("shift variance"). Two different types of user input are studied: information about the types of instruments, which enables the use of basis functions from an instrument database, and a manual transcription of a number of notes which enables the template estimation from the data under analysis itself. Experiments are performed on a data set of mixtures of acoustical instruments up to a polyphony of five. The results confirm a significant loss in accuracy when database templates are used and show the superiority of the Kullback-Leibler divergence over the least squares error cost function.

Index Terms— semi-automatic music transcription, non-negative matrix deconvolution, music signal processing

1. INTRODUCTION

Automatic music transcription refers to the transformation of a piece of music into a musical score or score-like representation. A large amount of research work has been carried out on this task — predominantly during the last two decades — and a wide variety of approaches has been proposed, ranging from heuristic procedures to perceptually-motivated approaches and data-adaptive methods. Despite these extensive endeavours, the performance of current systems for polyphonic music transcription is still not comparable to the transcription accuracy achieved by human experts. A comprehensive overview of computational methods for music transcription can be found in [1].

In this paper we study *semi-automatic* transcription where the user provides some prior information for the transcription process, such as the instrument identities in the target signal or some correct notes (as examples) for each instrument. The method proposed in this work is in line with various data-adaptive approaches based on the *non-negative matrix factorisation (NMF)* method [2]. Smaragdis and Brown [3] were the first to apply the basic NMF algorithm to the task of music transcription. This method approximates a magnitude spectrogram by a sum of instrument sound spectra (basis functions) each of which is weighted by time-variant gain values. Basis functions and gains are estimated from the spectrogram by iteratively and alternately updating randomly initialised matrices.

In the same year, Smaragdis [4] introduced an extension to the basic NMF algorithm which models the magnitude spectrogram of a signal as a convolution of *two-dimensional source spectrograms* with the corresponding gain functions. This method was evaluated on — and is particularly useful for — drum transcription applications where events of an instrument are likely to exhibit a similar frequency content and time evolution.

A modification of the NMF algorithm that aims at transcribing and separating sounds of pitched instruments was described by Fitzgerald et al. [5] and a similar approach based on the related PLCA technique can be found in [6]. Here, the constant-Q magnitude spectrogram of each source is modeled as a convolution of a *single spectral shape* with the corresponding two-dimensional gain function. Thus, besides the estimation of the spectral shapes, corresponding matrices are estimated that contain the gains for each point in time and frequency shift.

Schmitt and Mørup [7] combine both time-extended basis functions and shifts along the frequency axis in their *non-negative matrix factor 2-D deconvolution (NMF2D)* algorithm. The authors evaluate their method on a small excerpt of synthesized chamber music.

The representation of a musical instrument by a fixed spectral shape shifted in frequency is a very coarse characterisation of most real-world musical instruments. Usually, the average spectral shape of an instrument sound is strongly dependent on the fundamental frequency (cf. [8]). We therefore propose a model that estimates a different instrument spectrum for each fundamental frequency under analysis which we call *shift-variant non-negative matrix deconvolution (svNMD)*. The model is fully shift-variant (i. e. a different basis function can be assigned to each frequency shift) but can also be used with partially or fully shift-invariant basis functions. *Partially shift-invariant* basis functions refer to a set of fixed instrument spectra that are shifted within adjacent pitch ranges; *fully shift-invariant* basis functions are shifted over the whole frequency range and were used in the above mentioned publication [5, 6].

The remainder of this paper is organised as follows: Section 2 describes the algorithmic foundations of the svNMD algorithm. In Sect. 3, we consider different types of user information and explain how they can be used for semi-automatic transcription. Section 4 presents the evaluation procedure and discusses the results and Sect. 5 summarises and concludes the work.

2. SHIFT-VARIANT NON-NEGATIVE MATRIX DECONVOLUTION

In the proposed model, each instrument is represented by a set of basis functions corresponding to different fundamental frequencies in the constant-Q spectrogram. The model for the magnitude spec-

This work was funded by a Queen Mary University of London CDTA studentship.

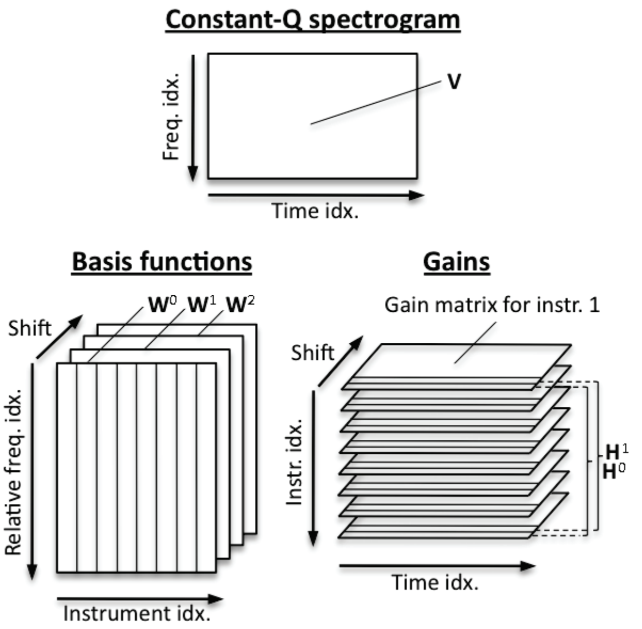

Fig. 1. Graphical illustration of matrices $\mathbf{V}$, $\mathbf{W}^{\phi}$ and $\mathbf{H}^{\phi}$

trogram $\mathbf{V}$ is given by

$$\mathbf{V} \approx \mathbf{\Lambda} = \sum_{\phi=0}^{\Phi-1} \overset{\phi\downarrow}{\mathbf{W}^{\phi}}\, \mathbf{H}^{\phi}, \tag{1}$$

where $\mathbf{V} \in \mathcal{R}_{+}^{N\times M}$ denotes the constant-Q magnitude spectrogram (with N frequency bins and M frames) and $\mathbf{\Lambda} \in \mathcal{R}_{+}^{N\times M}$ the approximation by the model. $\mathbf{W}^{\phi} \in \mathcal{R}_{+}^{N\times d}$ contains in its columns the spectra (basis functions) of the d instruments for a particular frequency shift ϕ. The operator $\phi\!\downarrow$ denotes a downward shift of the matrix elements by ϕ rows while the upper ϕ rows are filled with zeros. $\mathbf{H}^{\phi} \in \mathcal{R}_{+}^{d\times M}$ contains the gains of each basis function for each instrument at shift ϕ. Figure 1 illustrates the matrices $\mathbf{V}$, $\mathbf{W}^{\phi}$ and $\mathbf{H}^{\phi}$ graphically.

This model is similar to the NMF2D algorithm in [7], but modifies it in two aspects:

1. The temporal extent τ of the basis functions is set to a value of 0 (equivalent to a single audio frame), so that all temporal information is contained in the gain matrices. This way, the sum over the time instances τ can be abandoned.

2. We extend Schmidt and Mørup's model to work with different basis functions for each frequency shift and each instrument. We therefore use $\mathbf{W}^{\phi}$ instead of $\mathbf{W}^{\tau}$ to indicate the basis functions for the different shifts ϕ.

A similar framework could be achieved by concatenating all stacked matrices $\mathbf{W}^{\phi}$ and $\mathbf{H}^{\phi}$ and using those in combination with the basic NMF algorithm [2]. The advantage of the proposed framework, however, is its clear and meaningful structure, and the property that all partials are at the same frequency positions which is achieved by the use of the shift-parameter and a logarithmic frequency axis. This property facilitates the interpolation of missing templates which is a crucial requirement for semi-automatic transcription (cf. Sect. 3).

In order to derive the iterative update functions for both $\mathbf{W}^{\phi}$ and $\mathbf{H}^{\phi}$ we calculate the gradient of the two commonly used cost functions *least squares error (LS)* and *generalised Kullback-Leibler divergence (KL)*:

$$C_{LS} = \|\mathbf{V} - \mathbf{\Lambda}\|_F^2 = \sum_i \sum_j \left([\mathbf{V}]_{i,j} - [\mathbf{\Lambda}]_{i,j} \right)^2 \tag{2}$$

$$C_{KL} = \sum_i \sum_j [\mathbf{V}]_{i,j} \log \left(\frac{[\mathbf{V}]_{i,j}}{[\mathbf{\Lambda}]_{i,j}} \right) - [\mathbf{V}]_{i,j} + [\mathbf{\Lambda}]_{i,j} \,. \tag{3}$$

In eq. (2), $\|\cdot\|_F$ denotes the Frobenius norm.
The update functions based on gradient descent are given by

1. Least squares error:

$$\mathbf{W}^{\phi} \leftarrow \mathbf{W}^{\phi} \bullet \frac{\overset{\phi\uparrow}{\mathbf{V}} \left[\mathbf{H}^{\phi}\right]^T}{\overset{\phi\uparrow}{\mathbf{\Lambda}} \left[\mathbf{H}^{\phi}\right]^T}, \quad \mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\left[\overset{\phi\downarrow}{\mathbf{W}^{\phi}}\right]^T \cdot \mathbf{V}}{\left[\overset{\phi\downarrow}{\mathbf{W}^{\phi}}\right]^T \cdot \mathbf{\Lambda}} \tag{4}$$

2. Generalised Kullback-Leibler divergence:

$$\mathbf{W}^{\phi} \leftarrow \mathbf{W}^{\phi} \bullet \frac{\overset{\phi\uparrow}{\left(\frac{\mathbf{V}}{\mathbf{\Lambda}}\right)} \left[\mathbf{H}^{\phi}\right]^T}{\mathbf{1} \cdot \left[\mathbf{H}^{\phi}\right]^T}, \quad \mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\left[\overset{\phi\downarrow}{\mathbf{W}^{\phi}}\right]^T \cdot \frac{\mathbf{V}}{\mathbf{\Lambda}}}{\left[\overset{\phi\downarrow}{\mathbf{W}^{\phi}}\right]^T \cdot \mathbf{1}} \tag{5}$$

$\mathbf{1} \in \mathcal{R}^{N\times M}$ is a matrix of ones with the same dimensions as $\mathbf{V}$ and $\mathbf{\Lambda}$. In these equations, $\bullet$ denotes elementwise multiplication and all divisions are also elementwise. A detailed derivation of equations (4) and (5) can be found in [9].

The model is highly underdetermined as the number of parameters to estimate in the model is larger than the number of elements in the input spectrogram $\mathbf{V}$. Therefore, the application of the update equations (eq. (4) and (5)) in a completely unsupervised manner will not yield any useful results. A useful estimation can only be achieved when a certain amount of prior information is provided.

3. USE OF PRIOR INFORMATION FOR MODEL INITIALISATION

The method introduced in Sect. 2 in its given form cannot be applied for completely unsupervised automatic music transcription. The aim of this paper is to study its use for *semi-automatic* music transcription in which the user provides some prior information about the instruments in the polyphonic mixture under analysis. We consider two different types of information from the user:

1. The user provides information about the instrument types contained in the mixture.

2. The user transcribes some notes for each instrument in the mixture.

In the first case, a transcription can be achieved by initialising the basis functions by a set of instrument sound spectra learned from an instrument database and updating the gain matrices of the model while keeping the basis functions fixed.

In the second case, the basis functions for the user-annotated pitches can be learned from the data under analysis itself by first initialising the gain matrices at the annotated pitches and learning the spectra of these by updating the basis functions only. Once the instrument sound spectra are learned, the transcription can be obtained as in the first case by randomly initialising the gain matrices and applying the update functions for $\mathbf{H}^{\phi}$. In a practical application for user-assisted transcription, the user would only be required to label a small number of notes for each instrument and it can be assumed that the transcription performance depends on the number and type of notes the user selects. In this study, however, we are only interested in the upper limit of performance that can be achieved by the proposed svNMD procedure. This upper limit is given when the user provides information about all notes of all instruments during the basis function learning process. An investigation of the number and type of notes required is deferred to a subsequent study.

3.1. Learning the basis functions

In order to study the first case of semi-automatic transcription mentioned in the beginning of this section, instrument spectra were learned from training data that was completely separate from the data used for testing. The training audio files were taken from the RWC database [10] for each instrument identified by the user as being present in the target (test) data. Each of these audio files contains monophonic recordings of the instruments playing a chromatic scale over their whole compass. These recordings were manually annotated and the annotations were stored in MIDI format. To learn the basis function sets from this training data, the above-described update rules were used, fixing the gains $\mathbf{H}^{\phi}$ to contain ones at the frequency bins and time frames corresponding to the notes in the training data annotation and zeros elsewhere. The exact frequency bins were determined by finding the maximum within the frequency region of a semitone in the corresponding constant-Q spectrogram of each note at each time frame. The constant-Q analysis covered the frequency range from C2 (~ 65 Hz) to C8 (~ 4.2 kHz) with a frequency resolution of 48 bins per octave. The time resolution was 4.1 ms. Matrices $\mathbf{W}^{\phi}$ were randomly initialised and 10 iterations of the update functions were computed.

For the second case mentioned in the beginning of this section, the same learning procedure was applied to derive basis functions from the target (test) data under analysis. A set of basis functions for each instrument was learned both from the *monophonic* phrases used in creating the mixtures (see Sect. 4.1) and from the *polyphonic* instrument mixtures of the test set itself.

To summarise, we study the performance of basis function sets learned from three different sources:

1. recordings of the corresponding instruments in the *training data* (RWC database),
2. *monophonic* phrases used in creating the target (test) mixtures,
3. the target *polyphonic* instrument mixtures.

3.2. Learning the gain matrices

We are interested in the transcription accuracy that can be achieved by each of the basis function sets introduced in the previous section in combination with the svNMD method. In a practical application of semi-automatic transcription, no prior information about the gains of each spectrum would be available apart from the few user-labelled notes. Therefore, we initialised the gain matrices $\mathbf{H}^{\phi}$ with random non-negative values. The gains were learned by initialising the matrices $\mathbf{W}^{\phi}$ with one of the three basis function sets discussed above, and applying the update functions only for the gain matrices $\mathbf{H}^{\phi}$. This was done for all combinations of basis function sets (cf. Sect. 3.1) and cost functions (cf. Sect. 2). Again, a fixed number of 10 iterations was computed.

4. EVALUATION

4.1. Test data

A test set was constructed based on monophonic musical phrases from 12 different acoustical instruments (flute, oboe, clarinet, bassoon, alto sax, horn, trumpet, trombone, tuba, violin, viola and violoncello). Most of the phrases were previously used in [11]. Each of the signals had a length of approximately 30 s. From these phrases, random mixtures of 2, 3, 4 and 5 instruments were generated by summing the amplitude-normalised signals. 50 mixtures were generated for each polyphony level.

Pitch, onset time and offset time of each note were manually annotated for each monophonic phrase and stored as MIDI files. The ground truth for each instrument mixture was created by combining the ground truth annotations of the instruments contained in the mixture.

4.2. Transcription accuracy

For the evaluation of the svNMD method we did not apply the commonly used accuracy measures of note transcription such as precision, recall and F-measure. These measures would require further processing steps — such as thresholding the gains and f0-tracking — which would have an impact on the results. Instead, we are only interested in the evaluation of the svNMD method in combination with different basis functions sets. Therefore, we measure how well the gain matrices estimated by the proposed method agree with the ground truth annotations.

We first sum the gains of all instruments in the mixture to get an overall gain matrix $\mathbf{G}$, each element of which is given by

$$[\mathbf{G}]_{\phi,n} = \sum_{i} \left[\mathbf{H}^{\phi}\right]_{i,n}. \tag{6}$$

The accuracy is then computed as the ratio between the energy of the fundamental frequencies in $\mathbf{G}$ and the overall energy of $\mathbf{G}$:

$$Acc = \frac{\sum_n \sum_{\phi \in \mathcal{F}_n} ([\mathbf{G}]_{\phi,n})^2}{\sum_n \sum_{\phi'} ([\mathbf{G}]_{\phi',n})^2}, \tag{7}$$

where $\mathcal{F}_n$ denotes the set of annotated pitches in frame n.

An accuracy of 1 corresponds to a perfect pitch detection and complete suppression of the harmonics above the fundamental frequencies since all energy in $\mathbf{G}$ is concentrated in the fundamental frequencies. Smaller accuracies indicate that there is a certain amount of energy elsewhere in $\mathbf{G}$.

4.3. Results

The results of the experiments are displayed in Fig. 2. The upper three panels show the results for the KL-divergence, the lower panels those of the least-squares cost function. From left to right, the results for the different dictionaries can be compared. Within each plot, the accuracies of all files in the test set for each polyphony are displayed as boxplots.

When comparing the different dictionaries, it is obvious that basis functions learned from the RWC database lead to lower accuracies than basis functions learned from the recordings under analysis themselves – even when the basis functions are learned from *polyphonic* target data (as opposed to monophonic training data). This confirms that there is a significant loss of accuracy when independent training data from the same instrument type is used. Depending on

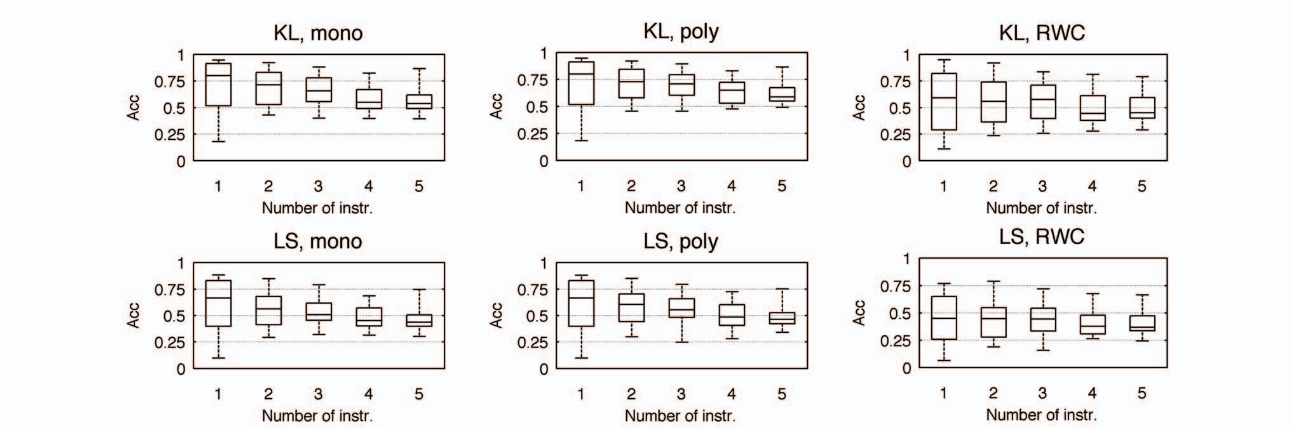

Fig. 2. Transcription accuracy of the the svNMD method with different initialisations of the basis functions. Results for the KL-divergence and least squares cost function can be found in the upper and lower rows, respectively. The different sets of basis functions are displayed from left to right. In each plot, the accuracies of the files of each test set are displayed as boxplots. The edges of the boxes mark the lower and upper quartile (median indicated in the middle) and the whiskers extend to the minimum and maximum data points.

the cost function and the number of instruments in the mixture, the median of the accuracies is between 23% and 48% higher when the basis functions are estimated from the polyphonic mixture instead of using generic basis functions from a database.

Results obtained by the generalised Kullback-Leibler divergence cost function (KL) clearly yield better results than using the least-squares cost function (LS). Without exception, accuracies for the least squares error are significantly lower; the deviation of the medians varies between about 18% and 34%.

5. CONCLUSION

We presented a shift-variant non-negative matrix deconvolution method that chararacterises each instrument by a different basis function at each possible fundamental frequency bin. It can be assumed that this gives a more accurate representation of real-world musical instruments than shift-*invariant* NMF and PLCA methods and has the potential to lead to better approximations of the magnitude spectrogram. The number of model parameters is an order of magnitude higher than the information contained in the input spectrogram, therefore the model is not intended for unsupervised learning, but is well-suited as a framework for semi-automatic music transcription. A MATLAB implementation of the svNMD algorithm can be found at `http://code.soundsoftware.ac.uk/projects/svnmd`.

We experimentally investigated the use of the model for user-assisted music transcription. Two different types of user information were investigated that led to different sets of basis functions. Basis functions were derived from the data under analysis itself and from a database of musical instrument sounds. The results confirmed that significantly higher transcription accuracies can be obtained when instrument templates are learned from the data under analysis than using generic basis functions learned from training data.

6. REFERENCES

[1] Anssi Klapuri and Manuel Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, 2006.

[2] D.D. Lee and H.S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[3] P. Smaragdis and J.C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2003, IEEE, pp. 177–180.

[4] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *5th International Conference on Independent Component Analysis and Blind Signal Separation*, Granada, Spain, September 2004, pp. 494–499.

[5] D. Fitzgerald, M. Cranitch, and E. Coyle, “Shifted non-negative matrix factorisation for sound source separation,” in *IEEE Workshop on Statistical Signal Processing*. IEEE, 2005, pp. 1132–1137.

[6] P. Smaragdis and B. Raj, “Shift-invariant probabilistic latent component analysis,” Tech. Rep. TR2007-009, Mitsubishi Research Laboratories, 2007.

[7] M. Schmidt and M. Mørup, “Nonnegative matrix factor 2-D deconvolution for blind single channel source separation,” in *6th International Conference on Independent Component Analysis and Blind Source Separation*, Chareston, SC, USA, March 2006, pp. 700–707, Springer.

[8] S. Handel, “Timbre perception and auditory object identification,” in *Hearing*, B. C. J. Moore, Ed. 1995, Academic Press.

[9] H. Kirchhoff, S. Dixon, and A. Klapuri, “Derivation of update equations for shift-variant non-negative matrix deconvolution (svNMD),” Tech. Rep. C4DM-TR-01-12, Queen Mary University of London, 2012, `http://www.eecs.qmul.ac.uk/~holger/C4DM-TR-01-12`.

[10] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical, and jazz music databases,” *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, pp. 287–288, October 2002.

[11] K.D. Martin, *Sound-source recognition: a theory and computational model*, Ph.D. thesis, Massachusetts Institute of Technology, MA, USA, June 1999.