AN ITERATIVE MULTI RANGE NON-NEGATIVE MATRIX FACTORIZATION ALGORITHM FOR POLYPHONIC MUSIC TRANSCRIPTION

Anis Khlif
École des Mines ParisTech, France
anis.khlif@mines-paristech.fr

Vidhyasaharan Sethu
University of New South Wales, Australia
v.sethu@unsw.edu.au

ABSTRACT

This article presents a novel iterative algorithm based on Non-negative Matrix Factorisation (NMF) that is particularly well suited to the task of automatic music transcription (AMT). Compared with previous NMF based techniques, this one does not aim at factorizing the time-frequency representation of the entire musical signal into a combination of the possible set of notes. Instead, the proposed algorithm proceeds iteratively by initially decomposing a part of the time-frequency representation into a combination of a small subset of all possible notes then reinvesting this information in the following step involving a large subset of notes. Specifically, starting with the lowest octave of notes that is of interest, each iteration increases the set of notes under consideration by an octave. The resolution of a lower dimensionality problem used to properly initialize matrices for a more complex problem, results in a gain of some percent in the transcription accuracy.

1. INTRODUCTION

The term Automatic Music Transcription (AMT) refers to the task of designing a system that automatically transposes an acoustic signal into a written format that can be read by a musician e.g. sheet music. In Western music, the basic unit of this transposition is the note, which is partly defined by its duration and its pitch. When more than one note can occur at the same time, the music is said to be polyphonic. Further, each instrument has its own harmonic pattern that is time-dependent for each of its notes. Indeed, the spectral content during the onset part of a note is different from the one during the sustain or fading parts. AMT of polyphonic musics amounts to tracking the fundamental frequencies among a mixture of musical events with possibly overlapping harmonics. Many approaches have been proposed but the results are still unsatisfactory compared to what can be achieved by a human expert [5]. Lately, techniques like NMF [17] [16] [7] and Probabilistic Latent Component Analysis (PLCA) [4] [18] have gained great interest since they have proved very efficient in bringing forward the underlying structure of musical data. Both are conceptually linked and have been shown equivalent under certain formulations [8]. They provide a framework under which the transcription can be formulated as a cost-function minimization problem, which are deeply studied problems and many algorithms exist to solve them. However, these algorithms (such as gradient descent, expectation maximization, alternating least-squares, etc...) suffer from major flaws. They offer no guarantees of finding a global minimum (if any) in general, and can easily get stuck in local ones. On top of this, they are highly sensitive to initial conditions and an improper initialization can lead to bad results [6] [1]. These issues are great liabilities for AMT because the intricate nature of harmonically related sounds results in the existence of many local minima which in turn increases the chance of an incorrect transcription.

In this paper we present an NMF-based algorithm tailored for the task of AMT, showing increased robustness with respect to the issues of finding proper initialization parameters and avoiding irrelevant local minima.

2. THE NMF FRAMEWORK

2.1 General overview

The different steps of the algorithm are presented in Figure 1. First, the time-frequency representation of the signal (spectrogram) is computed by applying a Constant Q Transform (CQT) on successive time windows. Then the proposed IMRNMF algorithm is applied to the spectrogram to produce a matrix representing the activation of each note accross time. This matrix is then post-processed to extract chunks representing potential notes which are then weighted before being truly acknowledged as a note and transcribed.

NMF aims at representing a non-negative signal as an additive synthesis of events taken from a finite dictionary. The original signal is then represented by the activation at each time of a subset of these events. If the signal lends itself to such description, the decomposition will likely be meaningful in the sense that it will bring out some of the underlying structure. In the case of AMT, the decomposition of the music into events that can be assimilated to notes, would be most desirable. A time-frequency repre-

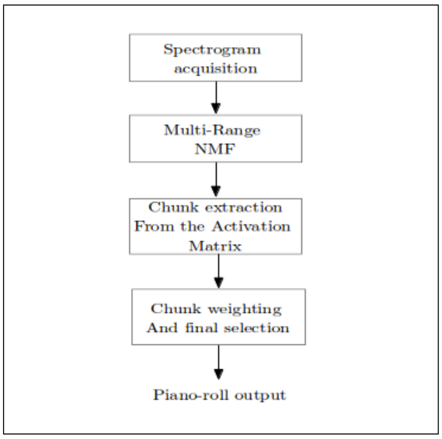
Spectrogram acquisition
Multi-Range NMF
Chunk extraction From the Activation Matrix
Chunk weighting And final selection
Piano-roll output

Figure 1: Overall algorithm.

sentation, like the spectrogram (which is a matrix containing the amplitude spectrum for a sequence of time windows) is an example of additive data where the sources would be constituted by the amplitude spectra of the different notes composing it. As mentioned, we will consider that the time-frequency representation is obtained via a CQT, which allows all frequencies of interest for all notes to be contained in the same number of frequency bins (in the time frequency representation) regardless of the octave or the note unlike the standard Discrete Fourier Transform.

More formally, given the spectrogram $Y \in \mathbb{R}_+^{N \times T}$, where N is the number of frequency bins and T the number of temporal frames, and given $K \leq N, T$; find $W \in \mathbb{R}_+^{N \times K}$ the spectral dictionary matrix, and $H \in \mathbb{R}_+^{K \times T}$ such that

$$Y \approx WH \tag{1}$$

Where, Y denotes the spectrogram obtained from the CQT, and which is decomposed into weighted sums of a finite set of notes whose spectra constitute the columns of W.

This decomposition does not have an exact solution and consequently the typical approach is to find a solution that minimises a suitable cost function, $\mathcal{C}_Y(W, H)$ with the constrainsts that the elements of W and H are positive. Historically, as introduced by Paatero [15] the canonical norm of the matrices difference: $\|Y - WH\|$ was taken as a cost function. Incidently, a factorization is inherently dependent on the cost function used to weight the reconstitution. As a result, the choice of a relevant cost function to increase the accuracy of the decomposition has been largely studied and yielded significant increases in the results. In the next section we review some of the key principles driving current efforts to enhance the transcription through NMF related techniques.

2.2 Achieving a good factorization

The best factorization we could hope for, would express Y as the activation of spectral templates that correspond exactly to the ones of the notes present in the excerpt. That implies especially, that no existing note be expressed as the sum of two or more elements (columns) of the dictionary W, (no false detection), or that no combination of two or more notes be expressed by a single element (no deletion). Such issues are referred to as cross-row talk. A common response to cross-row talk is to try to increase the sparsity of the decomposition matrices, and especially the columns of H. (A vector is said to be sparse when most of its elements are zeros). The energy is concentrated in a few units which are used to represent typical data vectors. Having a control over sparsity provides more robustness in "real-life" situations where the number of sources is not known by advance and a higher rank than needed is fixed for the decomposition matrix.

Controlling the sparsity is mainly achieved by choosing a suitable cost function $\mathcal{C}_Y$ and estimation methods that allow desirable properties to be enforced on W and H. Although, the task of finding a minimum for $\mathcal{C}_Y$ is not easy since the problem is often ill-posed, the reformulation of the factorization problem in terms of approaches such as Convex Quadratic Programming [7] [19] [11] provides elegant frameworks to naturally introduce new cost functions (with regularization parameters), or enforce relevant constraints on W and H.

The control over sparsity can be explicit. In [12], Hoyer develops algorithm to enforce constant predefined sparsities s_w, s_h over W and H. Such conditions are not realistic in real-life situations for audio data since the degree of polyphony can evolve throughout the excerpt. In [11], Heiler and Schnörr, give a formulation of the factorization as a second order cone programming problem, enabling them to enforce only boundary conditions on the sparsities. In [1], an adaptation of the ALS algorithm called Alternating Hoyer-Constrained Least Squares is proposed. However, this way of enforcing sparsity is often too restrictive in the case of musical data where the degree of polyphony is free to evolve during time, on top of the fact that we do not have prior knowledge on it. Consequently, we would prefer a softer, implicit control over sparsity. In such cases, it is often achieved through cost functions that are expressed in a form where the variation of a parameter provides an input to indirectly affect sparsity. In [7] the cost function is defined by

$$C_y = \frac{1}{2}\|Y - WH\|_2^2 + \lambda_1\|H\|_1 + \frac{\lambda_2}{2}\|H\|_2^2 \tag{2}$$

The coefficient λ_1 weights the importance given to sparse vectors against a good reconstitution, and λ_2 is a Tikhonov regularization parameter. Other successful approaches have considered a class of divergences called β-divergences as cost functions [10], which were successfully applied to AMT in [7]. $d_\beta(Y|W, H)$ is defined by:

$$d_\beta(Y|W,H) = \begin{cases} Y \otimes \log \frac{Y}{WH} - Y + WH & \beta = 1 \\ \frac{Y}{WH} - \log \frac{Y}{WH} - \mathbb{1} & \beta = 0 \\ \frac{1}{\beta(\beta-1)}(Y^\beta + (\beta - 1)(WH)^\beta - \beta Y \otimes (WH)^{\beta-1}) & else \end{cases} \tag{3}$$

Where the divisions, the logarithm and the powers have to be understood element-wise, $\otimes$ is the element-wise product, and $\mathbb{1}$ the matrix containing only ones. The choice of β provides an indirect control over sparsity. It can be noted that in the case of $\beta = 2$ it reduces to the Euclidean distance, and in the case $\beta = 1$ to the KL-div KL divergence, which has been found to promote sparsity [17]. The minimization of both those cost functions can be achieved through multiplicative update rules given in [10] and [7]. This is the cost function which has been adopted in the proposed method.

Finally, all the algorithms mentioned are highly sensitive to initial conditions and perform poorly when dimension and the density of local minima increase. In the case of AMT, initializing the spectral dictionary matrix W so that the elements (columns) are structurally relevant, improves the factorization a great deal. In [16], the columns of W are initialized with one for each note at harmonic positions and zeros elsewhere. It makes W relevant for the transcription and straightforward to associate to a note. While it is not too difficult to see how W can be intitialized, it is much less obvious for H.

In the next section, we present a versatile algorithm to perform the factorization which can be used with any update rules, enhances the sparsity and gives element of answer as to how initialize H leading to increased robustness.

3. THE PROPOSED FACTORIZATION ALGORITHM

3.1 Principle

The proposed algorithm performs an iterative factorization of the spectrogram by initially starting with a single octave of notes prior to incrementing it by an octave in each subsequent iteration. The algorithms performs by starting from the lowest octave, and by including one higher octave at each step until the whole range of note is covered. Let $\mathcal{S} = \{n_0, ..., n_{K-1}\} \in \mathbb{N}^K$ be an interval of integers containing the midi notes considered. The i-th range is the subset of $\mathcal{S}$ defined by $r_i = \{n_0, ..., n_{12i-1}\}$.

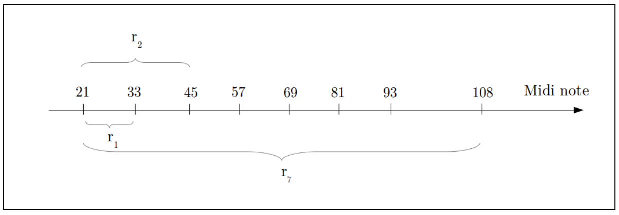

Figure 2: Cutting of the midi scale in ranges.

Only considering the notes lying in this range comes down to focusing on subregions, in terms of frequency and notes, of the decomposition matrices defined as follows.

$$Y^{(i)} \approx W^{(i)} H^{(i)} \tag{4}$$

with:

$$Y^{(i)} = Y_{[l_b, u_b^i], \bullet} \tag{5}$$

$$W^{(i)} = W_{[l_b, u_b^i], [l_s, u_s^i]} \tag{6}$$

The columns of W and the rows of H, indexed by the sources, are restricted to the subset $\{l_s, ..., u_s^i\}$ where l_s denotes the source of the lower note and u_s^i the source of the higher note of the i-th range. We have the following equalities: $l_s = n_0$ and $u_s^i = n_{12i-1}$. The rows of W and Y representing the frequency bins are restricted to the subregion $\{l_b, ..., u_b^i\}$ where l_b designs the lower frequency bin associated with the fundamental frequency of the lower note in the range, and u_b^i the upper bound for the frequency bins associated with the fundamental frequency of the higher note in the i-th range. As previously mentioned, the spectrogram is computed with a CQT, therefore we can note that the semitone resolution, b, i.e., the number of bins associated with a single semitone is a constant. The superscript (i) denotes the restriction of a matrix to the i-th range. With this notation, we can express the boundaries as: $l_b = b(n_0 - 1) + 1$ and $u_b^i = b(n_{12i-1})$. All the temporal frames are considered at each step of the factorization, this is noted $\bullet$.

As it has been said, any multiplicative update rule can be used with this approach. Specifically, in the work reported in this paper, the update rules (8) and (9) for the KL divergence are applied as follows to $H^{(i)}$ and the submatrix $W^{(i)}$.

$$H^{(i)} \leftarrow H^{(i)} \otimes \frac{{}^tW^{(i)}(Y^{(i)} \otimes (W^{(i)}H^{(i)})^{.\beta-2})}{{}^tW^{(i)}(W^{(i)}H^{(i)})^{.\beta-1}} \tag{7}$$

$$W^{(i)} \leftarrow W^{(i)} \otimes \frac{(Y^{(i)} \otimes (W^{(i)}H^{(i)})^{.\beta-2})\,{}^tH^{(i)}}{(W^{(i)}H^{(i)})^{.\beta-1}\,{}^tH^{(i)}} \tag{8}$$

Then $H^{(i+1)}$ is initialized as follows (see 3):

$$\begin{cases} H^{(i+1)}_{[l_s, u_s^i], \bullet} = H^{(i)} \\ \\ H^{(i+1)}_{[u_s^i, u_s^{i+1}], \bullet} = random\ positive\ matrix \end{cases} \tag{9}$$

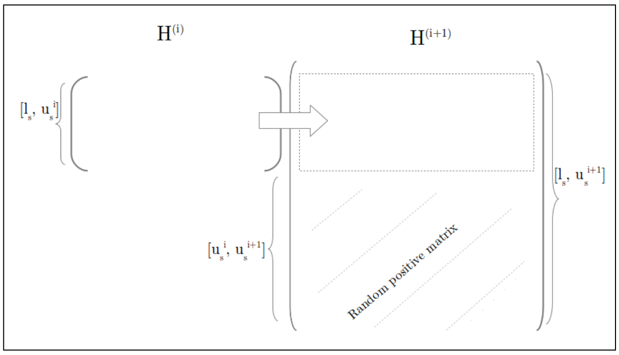

Figure 3: Initialization of $H^{(i+1)}$ from $H^{(i)}$.

Figures depicting the evolution of the activation matrix throughout the different steps are shown in section 5.

3.2 Motivation and advantages

This method has been designed as a way to compensate for some of the weaknesses of NMF applied to AMT, principally being having to use more potential sources than strictly necessary in the decomposition (which can cause confusion in the factorization, hence the necessity of enforcing sparsity), and the high spectral similarity between certain combination of harmonically related notes, which added to the high number of sources is likely to increase the probability of falling into a local minimum. Starting from the lowest octave, helps secure a sound bass-line and avoid confusing notes with weak fundamental with their upper octave counterpart (octave problem); as it is likely to happen in the usual implementation since low notes often have a weak fundamental. Incrementing the set of notes by a single octave is also a step in this direction, in order to limit as much as possible the risks of mistaking a note for one of its harmonically related counterparts. Beside, limiting the number of sources reduces the dimension of the problem and heuristically, the risks of falling into a local minimum. Re-investing knowledge in the next steps of the factorization helps converge toward a better minimum by ensuring convergence on growing subspaces, where confusion is less likely. The resulting activation matrix is much sparser, and much easier to post-process because of the more distinct activation peaks.

An additional advantage of the proposed method is that it allows for different treatments on the parts of the spectrogram that are factorized. For instance, it allows for the definition of octave-based tolerance thresholds in terms of amplitude or spatial repartition (peaks with a maximum value under a threshold or ranging on less than a given number of frames will be discarded). Various works in the fields of psychoacoustics and acoustic signal processing showed that such treatment is of the utmost importance in order to reliably weight and perform competitive selection between acoustic events distributed across a large frequency span and with different amplitudes [13] [20] [14].

4. BACK-END TRANSCRIPTION

The back-end transcription limits itself to the mere detection of activation events in H, since the initialization of W made straightforward the association between events and notes. H having previously been normalized we applied a threshold-based onset detection, allowing to debit activation matrix rows into chunks that can further along be weighted and sieved before being labelled as note. Those chunks are bits of the activation matrix defined by: the midi note (the row number), the onset time and the offset time. The computation of the onsets is performed by applying an adaptive thresholding on the first order differential vector of each row of H as suggested in [2]. The thresholding value is based on the mean of the half-wave rectified first order differential signal on the 100 neighbouring frames. The onset is defined as the first frame for which the amplitude is superior than 0.2 times the thresholding value (it has experimentally been found as a good value).

A score on the chunks was defined in order to perform a post-selection of the chunk and screen out the ones that are very likely false positives. This cost function is based on features of the chunks considered as indicators of the probability of this chunk to represent a true positive. This features are: the length of the chunk l, the maximum value of the amplitude within this chunk m, the value of the first order differential of the signal at the onset time (representing the steepness of the onset) d, and the energy of the signal e within the chunk against the cumulated energy of the signal of lower harmonics during the same time range e_l. The score of a chunk is defined as:

$$S = (1 - \exp(\frac{-l}{c_1}))(1 - \exp(\frac{-m}{c_2}))$$
$$(1 - \exp(\frac{-d}{c_3}))(1 - \exp(\frac{-e}{c_1 e_l})) \quad (10)$$

where c_1, c_2, c_3, and c_4 are arbitrary constants. For the tests we used $(c_1, c_2, c_3, c_4) = (8, 0.1, 0.03, 0.66)$ and only chunks with a score higher than 0.2 were kept. These values were experimentally determined as reasonable and were kept fixed for the totality of our test. No music-specific fine-tuning was performed.

5. EXPERIMENTAL RESULTS

Tests were performed on the MAPS ENSTDkCl database [9] which is composed exclusively of piano recordings with a wide variety of polyphony, genre, tempo, and rhythm. The set of notes taken into account ranges between the midi notes 21 and 108. The spectrogram is computed by a CQT algorithm with sixty bins per octave to be robust to frequency shifts around the theoretical peak position. beta divergence cost function, with $\beta = 1$ (KL-divergence) was chosen for all matrix factorisations. The matrix W is kept fixed during the step-by-step factorization then, an additional standard NMF is performed with initialization from previous results. Our Iterative Multi Range Non-negative Matrix Factorization (IMRNMF) system is compared against an NMF-based system without the range-by-range factorization but the same back-end transcription algorithm; and the winning algorithm of the MIREX 2013 competition in Multi-F0 note tracking and Multi-F0 note estimation based on Shift Invariant Probabilistic Latent Component Analysis (SI_PLCA) [3]. The matrix W is initialized offline using the array provided with the SI_PLCA source code which consists of pre-extracted and pre-shifted spectral templates for various instruments. An onset-based metric is used with a 50 ms tolerance.

The transcription is performed on the first 30 seconds of each track in the database. The thresholding and weighing constants used in the back-end transcription as well as in the IMRNMF are kept fixed during the whole test independently of the extract being processed, and even better results can be achieved with a case-by-case fine tuning of these constants, based on parameters such as genre and tempo. Below are shown illustrative examples of the evolution of the activation matrix on the MAPS _MUS-schu_143_3_ENSTDkCl track.

Method	Accuracy	F-measure
NMF	0.38	0.55
SI_PLCA	0.37	0.53
IMRNMF	**0.52**	**0.69**

Table 1: Comparative results on the MAPS ENSTDkCl database.

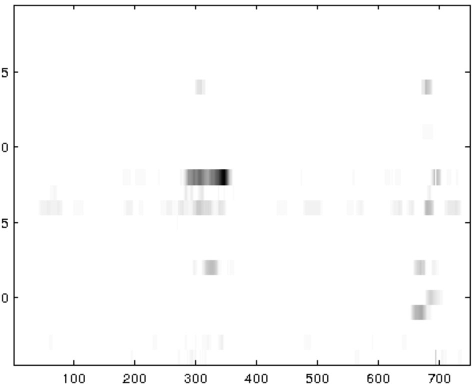

Figure 4: $H^{(2)}$ after the first 2 steps

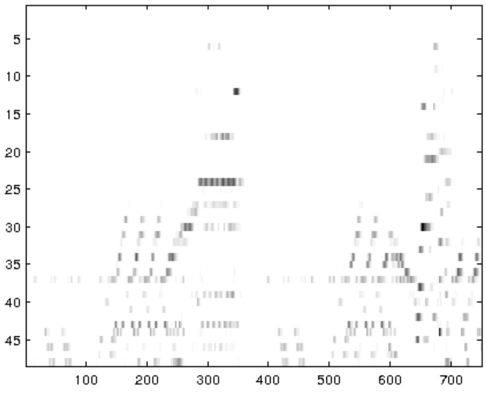

Figure 5: $H^{(4)}$ after the first 4 steps

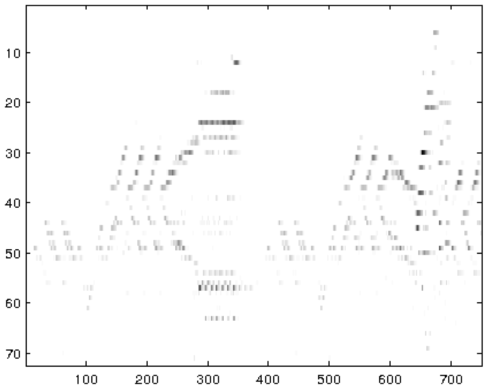

Figure 6: $H^{(6)}$ after the first 6 steps

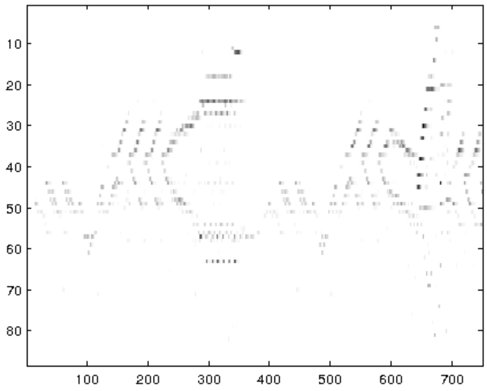

Figure 7: Final output of the factorization

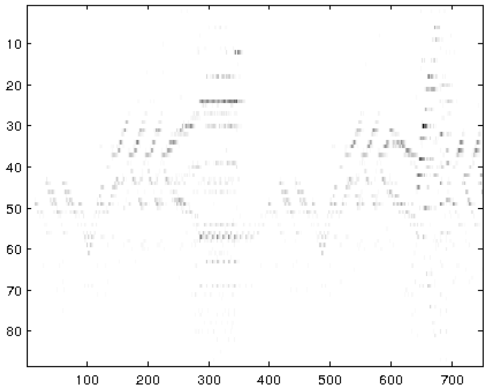

Figure 8: H obtained with SI_PLCA

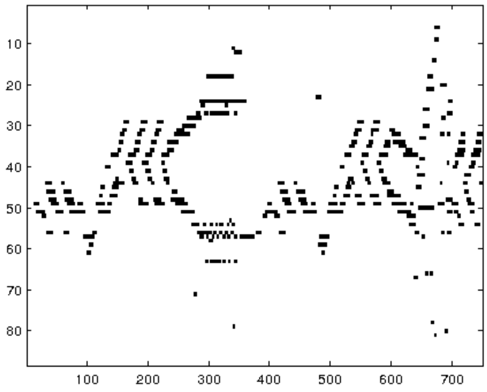

Figure 9: Backend transcription output

6. CONCLUSION AND FUTURE WORK

A novel Iterative Multi-Range Non-negative Matrix Factorisation (IMRNMF) based algorithm for automatic music transcription is presented in this paper. At the cost of increased computational requirements, though still perfectly accessible, the proposed system leads to an increase in transcription accuracy compared to the top-performing existing algorithms. This increase may be better explained by the increased sparsity of the activation matrix. The improved sparsity is most likely due to the proposed algorithm finding better local minima to the cost function when compared to the traditional NMF. While a number of parameters in the proposed systems are empirically determined at this stage (thresholding constants, weighting parameters, chunk-wise cost function in the final decision process...), a more data-driven approach to estimating them may lead to even better performance and will be addressed in future work.

Acknowledgements. Many thanks to Emmanouil Benetos for providing the sources of the Shift-Invariant Probabilistic Latent Component Analysis algorithm.

7. REFERENCES

[1] Russell Albright, James Cox, David Duling, Amy N Langville, and C Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical report, Tech. rep. 919. NCSU Technical Report Math 81706. http://meyer. math. ncsu. edu/Meyer/Abstracts/Publications. html, 2006. url: http://citeseerx. ist. psu. edu/viewdoc/download, 2006.

[2] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *Speech and Audio Processing, IEEE Transactions on*, 13(5):1035–1047, 2005.

[3] Emmanouil Benetos, Srikanth Cherla, and Tillman Weyde. An efficient shift-invariant model for polyphonic music transcription. In *6th International Workshop on Machine Learning and Music*, 2013.

[4] Emmanouil Benetos and Simon Dixon. A shift-invariant latent variable model for automatic music transcription. *Computer Music Journal*, 36(4):81–94, 2012.

[5] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: Breaking the glass ceiling. In *ISMIR*, pages 379–384. Citeseer, 2012.

[6] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

[7] Arnaud Dessein, Arshia Cont, Guillaume Lemaitre, et al. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *ISMIR-11th International Society for Music Information Retrieval Conference*, pages 489–494, 2010.

[8] Chris Ding, Tao Li, and Wei Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927, 2008.

[9] Valentin Emiya, Roland Badeau, and Bertrand David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1643–1654, 2010.

[10] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[11] Matthias Heiler and Christoph Schnörr. Learning sparse representations by non-negative matrix factorization and sequential cone programming. *The Journal of Machine Learning Research*, 7:1385–1407, 2006.

[12] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[13] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 6, pages 3089–3092. IEEE, 1999.

[14] Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[15] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[16] Stanisław A. Raczyński, Nobutaka Ono, and Shigeki Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *in ISMIR 2007, 8th International Conference on Music Information Retrieval*, pages 381–386, 2007.

[17] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180. IEEE, 2003.

[18] Emmanuel Vincent, Nancy Bertin, and Roland Badeau. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 109–112. IEEE, 2008.

[19] Rafal Zdunek and Andrzej Cichocki. Nonnegative matrix factorization with quadratic programming. *Neurocomputing*, 71(10):2309–2320, 2008.

[20] Ruohua Zhou. *Feature extraction of musical content for automatic music transcription*. PhD thesis, EPFL, 2006.