Non-Negative Matrix Factorization for Polyphonic Music Transcription

Paris Smaragdis

Mitsubishi Electric Research Lab
201 Broadway
Cambridge, MA 02139 USA
paris@merl.com

Judith C. Brown

Physics Department, Wellesley College
Wellesley, MA 02181, USA and
Media Lab, Massachusetts Institute of Technology
Cambridge, MA 02139, USA
brown@media.mit.edu

ABSTRACT

In this paper we present a methodology for analyzing polyphonic musical passages comprised by notes that exhibit a harmonically fixed spectral profile (such as piano notes). Taking advantage of this unique note structure we can model the audio content of the musical passage by a linear basis transform and use non-negative matrix decomposition methods to estimate the spectral profile and the temporal information of every note. This approach results in a very simple and compact system that is not knowledge-based, but rather learns notes by observation.

1. INTRODUCTION

Polyphonic music transcription has long been an extremely hard problem. Many knowledge-based approaches have been proposed in the past which have predominantly resulted in highly complex systems. In this paper we propose a lighter approach, akin to scene analysis, that is data driven and does not incorporate any prior knowledge of musical structure. It is based on the concept of redundancy reduction [1] an idea that for the last few years has been gaining momentum for many perceptual applications. More recently it has been applied to the polyphonic music transcription problem [2], [3] with very encouraging results. In this paper we pursue the same approach using a different computational angle. We show that it is possible to efficiently perform polyphonic music transcription using non-negative matrix factorization of musical spectra.

1. NON-NEGATIVE FACTORIZATION

1.1. Definitions and cost functions

Non-negative matrix factorization (NMF) was first proposed by Lee and Seung [4] and was inspired from previous work by Paatero [5] on positive matrix factorization. Starting with a non-negative M by N matrix $\mathbf{X} \in \mathbb{R}^{\geq 0, M \times N}$ the goal of NMF is to approximate it as a product of two non-negative matrices $\mathbf{W} \in \mathbb{R}^{\geq 0, M \times R}$ and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$ where $R \leq M$, such that we minimize the error of reconstruction. We do so by minimizing the cost function:

$$C = \| \mathbf{X} - \mathbf{W} \cdot \mathbf{H} \|_{\mathrm{F}}, \tag{1}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, or by using an alternative measure:

$$D = \left\| \mathbf{X} \otimes \ln\left(\frac{\mathbf{X}}{\mathbf{W} \cdot \mathbf{H}}\right) - \mathbf{X} + \mathbf{W} \cdot \mathbf{H} \right\|_F \tag{2}$$

where $\otimes$ is the Hadamard product (an element-wise multiplication of the matrices), and the division is element-wise. Both of these measures equal zero iff $\mathbf{X} = \mathbf{W} \cdot \mathbf{H}$. The cost function in Eq. (2) is somewhat akin to the Kullback-Liebler divergence. Algorithms for finding the appropriate values for **W** and **H** are shown in appendix A.

We could alternatively examine this factorization as a reduced-rank basis decomposition so that $\mathbf{X} \approx \mathbf{W} \cdot \mathbf{H}$, and subsequently $\mathbf{H} = \mathbf{A} \cdot \mathbf{X}$, where $\mathbf{A} \in \mathbb{R}^{\geq 0, R \times M} = \mathbf{W}^{+}$ and the $^{+}$ operator signifies the Moore-Penrose matrix inverse. The latter equation allows us to associate this operation with standard components analysis methods such as Principal Components Analysis (PCA) and Independent Components Analysis (ICA). In fact using the cost function in Eq. (1) we have found that the result of NMF is always a rotation of the equivalent result using PCA (PCA in fact minimizes the same cost function but with an orthogonality constraint). Based on this fact we conjecture that NMF is possibly performing non-negative ICA given that is satisfies the conditions set forth by Plumbley [6].

In more common terms what NMF does is summarize the profiles of the rows of **X** in the rows of **H**, and likewise for the columns of **X** in the columns of **W**. The parameter R that sets the rank of the approximation controls the power of summarization. If $R = M$ then we have an exact decomposition where the contents of **W** and **H** are not particularly informative. As we decrease the value of R the elements of **W** and **H** start to take values that concisely describe the main elements in the composition of **X**. If we choose appropriate values for R then it is possible to extract the major elements of the structure of **X**. A simple example of that is shown in the following section.

1.2. NMF on Magnitude Spectra

Let us start with a sound segment:

$$s(t) = \mathrm{g}(\alpha\, t) \sin(\gamma\, t) + \mathrm{g}(\beta\, t) \sin(\delta\, t), \tag{3}$$

where $\mathrm{g}(\cdot)$ is a gate function with a period of 2π and α, β, γ, δ are arbitrary scalars with α and β significantly smaller than γ

0-7803-7850-4/03/$17.00 © 2003 IEEE

and δ. We then compute its L-length time dependent magnitude spectrum $\mathbf{x}(t) = \|\,\mathrm{DFT}(\,[\mathrm{s}(t)\,\dots\,\mathrm{s}(t+L)])\,\|$. The set of all the $\mathbf{x}(t)$ can be packed as columns into a non-negative matrix $\mathbf{X} \in \mathbb{R}^{\geq 0, M \times N}$, where N are the total number of spectra we computed and M (= L/2+1) is the number of their frequencies. The matrix $\mathbf{X}$ (essentially the magnitude spectrogram of $s(t)$) is shown in figure 1.

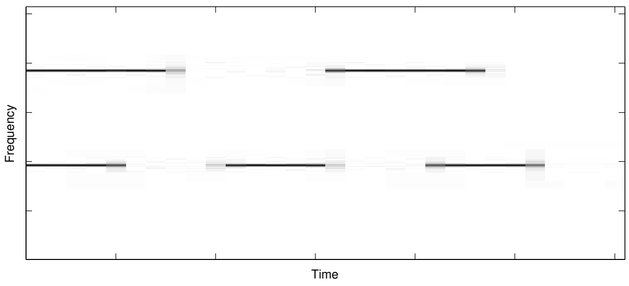


Figure 1. *Time dependent magnitude spectrum of a simple sound scene*

We can now try to perform NMF on our non-negative matrix **X**. Before we do so we note some of its characteristics. There is very little energy except for a few frequency bins, and in these bins we see a very regular pattern. In fact one could say that this is a highly redundant spectrogram (a compression engineer's best case scenario!) By performing NMF on this matrix we see how this redundancy can be taken advantage of to create a more compact and informative description. We optimize the cost function in Eq. (1) with R = 2 and display the results in figure 2.

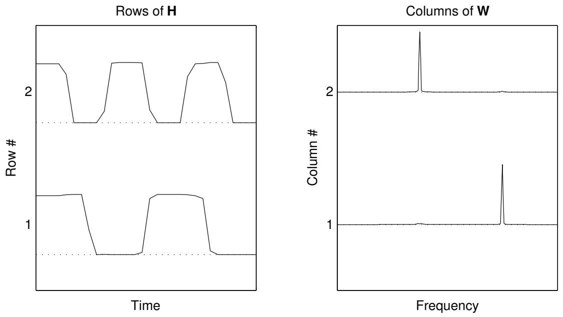


Figure 2. *NMF decomposition of the matrix pictured in figure 1*

By examination of the results we see a very useful behavior. The two rows of **H** contain two time series that best summarize the horizontal structure of **X**. Likewise the columns of **W** do so for the vertical structure of **X**. The pairing of the nth row of **H** with the nth row of **W** describes the two elements that made up this scene.

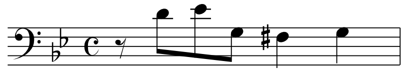
In the remaining sections we will generalize this idea to the case where **X** contains musical spectra. We will show that in this case the elements of **W** and **H** will respectively contain the spectrum and the temporal information of the notes in the analyzed passage. As this example demonstrates the proposed framework is robust enough to deal with multiple overlapping notes, a point to be made clearer using real world data in the following sections.

2. RESULTS ON MUSICAL SPECTRA

In this section we will demonstrate some results from real piano recordings and consider some of the issues that arise when performing NMF on musical spectra. We will first treat the case of isolated notes, discuss some of the issues regarding coinciding notes and conclude with a polyphonic transcription example. All of the examples are from a real piano recording by Keith Jarrett of Bach's fugue XVI in G minor [7] sampled at 44,100kHz and converted to mono by averaging the left and right channels.

2.1. Isolated notes

For the first example we will consider the first few notes of the fugue. There are no major overlapping note regions.



This passage contains five events made up from four different notes. We produce the time dependent magnitude transform spectrum and analyze it using NMF with the cost function in Eq. (1) and R = 4. For the spectrum analysis we used a 4096-point DFT and a Hanning window. The results of the analysis are shown in figure 3.

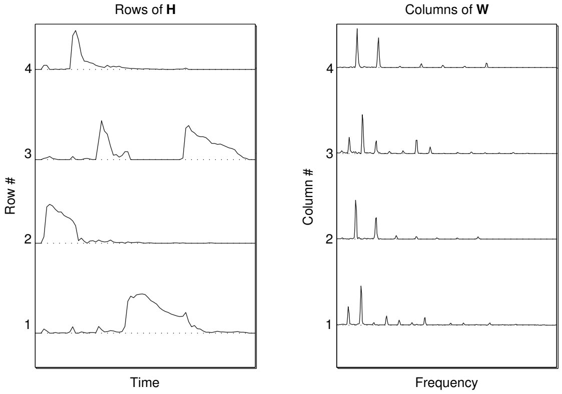


Figure 3. *NMF decomposition of the fugue bar 1*

Upon examination of the results it is clear that the rows of **H** each correspond to the temporal activity of the four notes, whereas the columns of **W** contain their respective spectra. The lowest significant frequency peaks from each of the columns of **W** are at 193.7Hz, 301.4Hz, 204.5Hz and 322.9Hz which correspond to the passage notes $\mathrm{F}^{\sharp}_{3}$, D_4, G_3 and $\mathrm{E}^{\flat}_{4}$. Deviations from the actual note frequencies are due to frequency quantization by the DFT. In general individual peaks will not help in note identification, instead we use the harmonic profile that all peaks collectively outline.

At this point it is worthwhile to ask what happens when R does not equal the number of present notes. For a smaller R we do not have enough expressive power to describe the scene and we are bound to have an incomplete analysis. To be on the safe side we can always choose a large enough R. In this case depending on the algorithm we use we can have one of many outcomes. NMF optimizing the cost function in Eq. (1) will produce the least desirable results distributing the energy of the most dominant notes across columns and rows of **W** and **H**. To counter this effect we can modify the cost function as:

$$C_2 = \| \mathbf{X} - \mathbf{W} \cdot \mathbf{H} \|_F + \lambda \|\mathbf{H}\|_F \qquad (4)$$

This cost function ensures that we not only accurately approximate our input, but that we do so with as low an energy representation as possible. This is similar to the cost function introduced by Hoyer [8]. The parameter λ weighs the importance of a good reconstruction as apposed to low energy. This forces the optimization to not distribute notes across many rows of **H** since that way it introduces more energy. With this cost function, extra components end up as lower energy noisy signals, as compared to the note components. However, the drawback of this approach is that we need an appropriate choice for λ.

Perhaps the most convenient algorithm when we are unsure of the value of R is NMF using the cost function in Eq. (2). In this case the extra components end up as very spiky rows of **H** that have a corresponding column of **W** that is a low energy wideband spectrum. An example of this is shown in figure 4 where $R = 5$ and the input was the four note passage we used before.

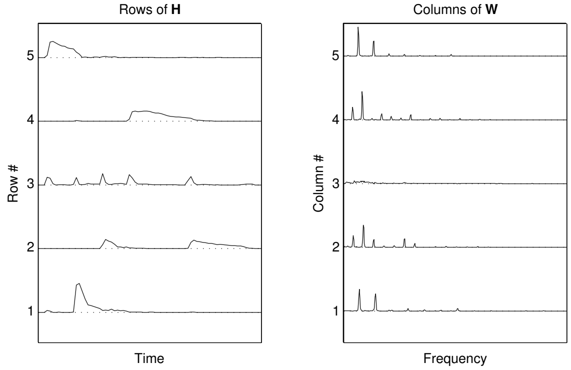
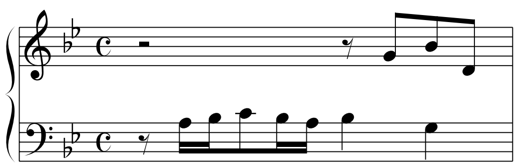


Figure 4. *NMF decomposition of fugue bar 1 with more components than notes*

In general identification of the non-note components is very easy and has not posed a problem in our analyses.

2.2. Coinciding notes

We now proceed to a simple example of polyphonic transcription. We analyze the second bar of our fugue. It is:



It exhibits ten events using seven different notes. Towards the end we have two notes (G_3 and B^{b}_4) sounding at the same time.

We perform NMF with the cost function in Eq. (2) and the same parameters we used before except for R that was set to 7. The results are shown in figure 5.

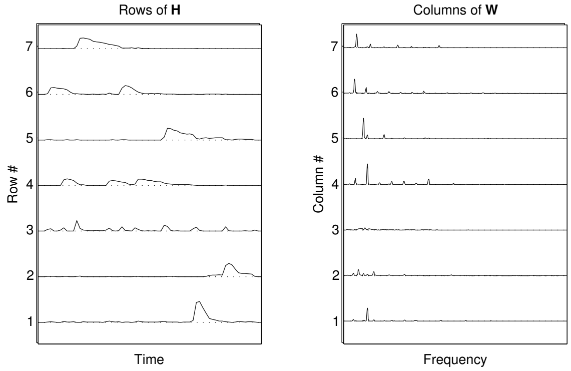


Figure 5. *NMF decomposition of fugue bar 2*

To our distress we see that although the events were very nicely transcribed we have a non-note component, and we see that the two simultaneous notes B^{b}_4 and G_3 were consolidated as one component. The reason for this unexpected result is rather easy to track. As we stated before, this is a transcription technique based on the system's accumulated experience from the presented input and not on predefined knowledge. Due to this all unique events are understood to be a new component. It is important to understand that this method will not extract notes, but rather unique events. We do not provide enough knowledge to purposely extract notes; but rather the algorithm has to examine the data to discover what appear to be unique events which should correspond to notes. The reason why in this case we end up with two notes being identified as one component is because they always occur at the same time. As far as the data goes the pair of B^{b}_4 and G_3 is a unique event and it should be one component not two.

The way to alleviate this problem is to present enough data so that all notes are exposed as either isolated events, or as parts of different polyphonic groups so as to highlight their individuality. In our case by analyzing the first two bars we provide an extra instance of the note G_3 which is enough to break up our previously fused component into two note components. The results are shown in figure 6. Upon examination it is easy to see that we have perfectly transcribed the input.

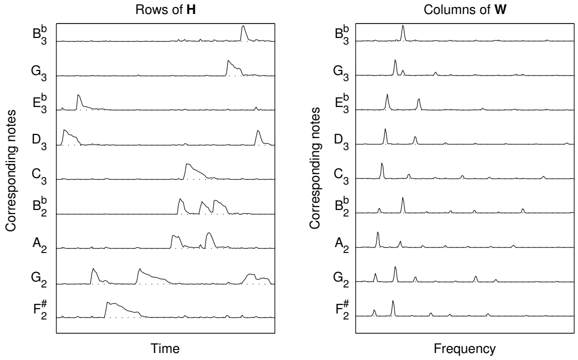


Figure 6. *NMF analysis of the first two fugue bars*

The same procedure scales well to additional bars of the fugues where the polyphonic element becomes more pronounced. Figure 7 displays the results for the first six bars of the fugue with $R = 27$. Upon convergence we removed 2 components that were clearly not notes and they are not displayed. Correlating the results with the score of the fugue we can verify that the musical structure has been almost correctly identified. The only mistakes in the transcription are the consolidation of E^{b}_{4} and E^{b}_{5} into one component and not tracking the note $F^{\sharp}_{5}$ (possibly because it only appears once as part of a chord).

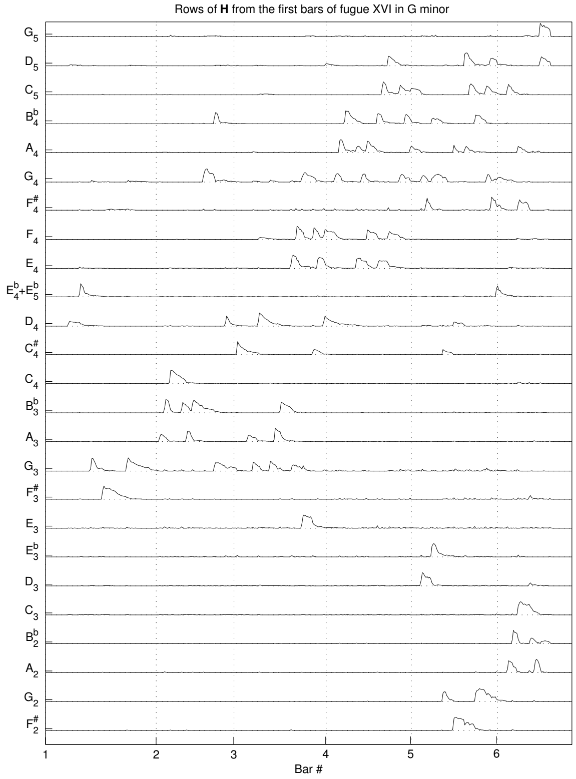


Figure 7. *Results from the first six bars of the fugue. Pictured are the only rows of **H**.*

3. CONCLUSIONS

We have presented a polyphonic music transcription system that is based on non-negative matrix factorization of magnitude spectra. We've demonstrated that this approach can produce good results without a prohibitive computational requirement or cumbersome system design. One of the shortcomings of this approach is that it requires music passages from instruments with notes that exhibit a static harmonic profile. In future work we anticipate addressing this issue with alternative decomposition methods that have more expressive power than linear transforms.

4. APPENDIX A: ALGORITHMS FOR NMF

There are many ways to minimize the cost function in Eqs. (1), (2) and (4). Due to space constraints we will show only two that we used in this paper.

The first method is for steepest descent on the cost functions from Eqs. (1) and (4). We need to optimize on two variables, **W** and **H**, so we need to find their derivatives with respect to C and C_2. Applying matrix differential calculus we can easily derive:

$$\Delta\mathbf{H} \propto (\mathbf{W}\cdot\mathbf{H} - \mathbf{X})\cdot\mathbf{H}^{T} \text{ and } \Delta\mathbf{W} \propto \mathbf{W}^{T}\cdot(\mathbf{W}\cdot\mathbf{H} - \mathbf{X}) \qquad (5)$$

We can simultaneously update **H** and **W** and ensure their non-negativity by forcing them to be non-negative at the end of each iteration[1]. To add the low energy constraint from the cost function in Eq. (4) we have to modify $\Delta\mathbf{H}$ to:

$$\Delta\mathbf{H} \propto (\mathbf{W}\cdot\mathbf{H} - \mathbf{X})\cdot\mathbf{H}^{T} + \lambda 2\mathbf{H} \qquad (6)$$

The second algorithm we consider is for the cost function in Eq. (2) and is more complicated to derive. Lee and Seung introduce and prove [1] a multiplicative algorithm that exhibits rapid convergence and is defined as:

$$\mathbf{H}_{am} \leftarrow \mathbf{H}_{am} \frac{\sum_{i} \mathbf{W}_{ia}\mathbf{X}_{im} / (\mathbf{W}\cdot\mathbf{H})_{im}}{\sum_{k} \mathbf{W}_{ka}}$$
$$\mathbf{W}_{ia} \leftarrow \mathbf{W}_{ia} \frac{\sum_{m} \mathbf{H}_{am}\mathbf{X}_{im} / (\mathbf{W}\cdot\mathbf{H})_{im}}{\sum_{v} \mathbf{H}_{av}} \qquad (7)$$

where the notation $\mathbf{A}_{ij}$ means the element of **A** at row i and column j.

5. REFERENCES

[1] Barlow, H.B. "Sensory mechanisms, the reduction of redundancy, and intelligence„. In *Symposium on the Mechanization of Thought Processes. National Physical Laboratory Symposium No. 10.* (1959)

[2] Smaragdis, P. "Redundancy reduction for computational audition, a unifying approach„. *Ph.D. dissertation*, MAS department, Massachusetts Institute of Technology, (2001).

[3] Plumbley, M.D., S.A. Abdallah, J.P. Bello, M.E. Davies, G. Monti and M.B. Sandler. "Automatic music transcription and audio source separation„. In *Cybernetics and Systems*, **33**(6), pp 603-627, (2002).

[4] Lee, D.D. and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization„. In *Nature* **401**, pp788-791, (1999).

[5] Paatero, P. "Least squares formulation of robust non-negative factor analysis„. In *Chemometrics and Intelligent Laboratory Systems* **37**, pp23-35, (1997).

[6]Plumbley, M.D. "Conditions for non-negative independent component analysis„. In *IEEE Signal Processing Letters*, **9**(6), pp177-180, (2002).

[7] Jarrett, K. "J.S. Bach, Das Wohltemperierte Klavier, Buch I„ *ECM Records*, CD 2, Track 8 (1988).

[8] Hoyer, P. "Non-negative sparse coding„ In *Neural Networks for Signal Processing XII*, Martigny, Switzerland, (2002).

[1] We can alternatively optimize $\|\mathbf{X} - (\mathbf{W}\otimes\mathbf{W})\cdot(\mathbf{H}\otimes\mathbf{H})\|_{F}$ which ensures matrix non-negativity. However the extra computational cost and computational complexity are not worth the admittedly less elegant non-negativity enforcement we introduce above.