MULTIPLE FREQUENCY ESTIMATION FOR PIANO RECORDINGS WITH CONCATENATED REGULARIZED HARMONIC NMF

S. A. Raczyński, S. Sagayma
The University of Tokyo
Graduate School of Information Science and Technology
E-mail: {raczynski,sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

RS1–6. This entry for MIREX's multiple frequency estimation is based on the very common method of Nonnegative Matrix Factorization (NMF). By adding regularizations specific to analysis of harmonic signals, using a harmonic basis and parameterizing the used distortion measure, a more accurate and parameterized algorithm has been developed. Its parameters are optimized with a random optimization algorithm on a dataset of piano recordings synchronized with MIDI data (treated as the groundtruth data). F-measure is maximized for each of the four frequency bands separately, resulting in four sets of parameters. For each parameter set the NMF algorithm is run and their results are concatenated for maximal accuracy. Before performing NMF, the input signal is separated into a harmonic part and a percussive part. The former is used to determine pitches existing in the recording, while the latter for accurate onset detection. The concatenated note activities are thresholded and median-filtered and very short notes are removed according to a note length model trained on the RWC database's piano recordings. Finally, the notes are detected as groups of non-zero note activities and their onsets are associated with the closest onsets detected from the percussive part.

1. DATA FLOW OVERVIEW

Before the analysis is started, the input signal is separated into a harmonic part and a percussive part. The harmonic part is used to determine pitches existing in the recording, while the percussive part is later used for accurate onset detection. The harmonic analysis is done by means of regularized harmonic NMF (described in section 2), which is performed four times, for four disjoint frequency ranges with different sets of parameters. The results are then concatenated to cover the full piano note range. The resulting note activities are thresholded and median-filtered and notes that are too short are removed, according to a note length model, that had been trained on the RWC database's piano recordings. Finally, the notes are detected as series of non-zero note activities and their onsets are snapped to the closest onsets detected from the percussive part.

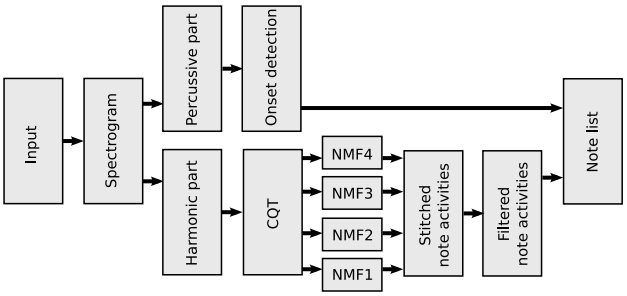


Figure 1: Data flow of the algorithm.

2. NONNEGATIVE MATRIX FACTORIZATION

NMF is a technique of approximating a data matrix $\mathbf{X}$ with a product of two matrices:

$$\mathbf{X} \cong \mathbf{AS} = \widetilde{\mathbf{X}}. \tag{1}$$

Both resulting matrices are nonnegative in terms of their elements and are obtained through minimalization of a Bregman divergence, defined as:

$$D_{\varphi}(\mathbf{X}, \mathbf{AS}) = |\varphi(\mathbf{X}) - \varphi(\mathbf{AS}) - \varphi'(\mathbf{AS})(\mathbf{X} - \mathbf{AS})|. \tag{2}$$

where $\varphi : \mathcal{R} \rightarrow \mathcal{R}$ is a convex generating function with a continuous first derivative. NNMA is an optimization problem with the penalty function being a Bregman divergence between the data $\mathbf{X}$ and its approximation $\mathbf{AS}$ and with a constraint of nonnegativity [1]. It can be easily solved using the Karush-Kuhn-Tucker conditions [2]. By minimizing the divergence between the data $\mathbf{X}$ and its approximation $\mathbf{AS}$ we obtain a pair of multiplicative update rules that, used alternately, leads to the optimal factorization:

$$\mathbf{S} \leftarrow \mathbf{S} \odot \frac{\mathbf{A}^T \left(\mathbf{X} \odot \varphi''(\mathbf{AS})\right)}{\mathbf{A}^T \left(\mathbf{A} \odot \varphi''(\mathbf{AS})\right)}, \tag{3}$$

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{(\mathbf{X} \odot \varphi''(\mathbf{AS}))\mathbf{S}^T}{(\mathbf{AS} \odot \varphi''(\mathbf{AS}))\mathbf{S}^T}. \tag{4}$$

We chose to narrow the possible choice of generating functions to a simple family of divergences by defining:

$$\varphi''(x) = x^{-r}, \tag{5}$$

This family of divergences is virtually identical to the beta divergence proposed by Kompass in [3] with only few small differences. Four important divergences belong to that family: Euclidean distance for $r = 0$, the KL- and I-divergence for $r = 1$ and the Itakura-Saito divergence for $r = 2$, all of which are commonly used in NMF.

3. HARMONIC BASIS MATRIX

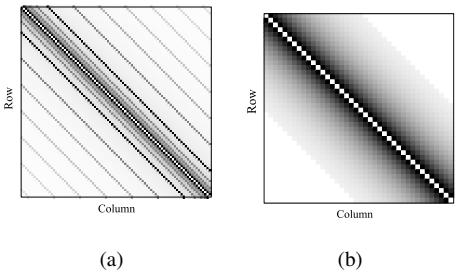
Using an unconstrained basis matrix poses a series of problems. Basis vectors need to be analyzed and assigned to a particular pitch, prior to the analysis of the note activity matrix, which introduces additional errors to the process. However, because note events do not occur sparsely and independently, and their spectra change greatly over time, using an unconstrained basis usually results in basis vectors that do not even have a harmonic structure, making the pitch estimation difficult or impossible. Furthermore, results for an unconstrained basis are very different each time the algorithm is run, and thus very difficult to compare and evaluate. That is why we firmly believe that a harmonic basis matrix with vectors constrained to harmonic structures strictly corresponding to notes (of, for instance, the diatonic scale) is a must when it comes to multipitch analysis. Analysis of the note activity matrix in this case is straightforward, as each row contains amplitudes of a single note.

Basis harmonicity can be achieved in three ways. We can either: use a fixed harmonic basis vectors (i.e. only use eq. 3), use a basis matrix pretrained on solo instrument data, or adapt the harmonic structure to the data. In the first approach we use an artificial harmonic spectra with partials' amplitudes decreasing exponentially with frequency. It would seem like an oversimplification, but, as we will see later, this method yields very good results, especially when additional penalties are used, and the overfitting present in the other two methods is avoided. In the second approach, we use averaged note spectra obtained from the recordings of piano taken from the RWC database, which gives better results than the first method, but the performance drops slightly when different instrument is used.

In the third approach, proposed by us in [4], we use an artificial harmonic basis from the first method and adapt it in a way that changes only the partials' amplitudes, leaving the overall harmonic structure intact. This can be easily achieved without modifying the existing algorithm, because zero-valued elements of basis vectors remain at zero throughout the learning process due to its multiplicative nature. We could therefore initialize the basis to have zeros everywhere but at the positions of fundamental frequencies of notes from a specific range of the 12-TET (Twelve-tone Equal Temperament) scale and at their harmonics, thus constraining the solution space to harmonic factorizations only.

4. REGULARIZATIONS

NNMA can be extended to include additional penalties on both matrices. In this case, instead of minimizing a Bregman divergence, the following objective function is minimized



(a) (b)

Figure 2: Circulant weighting decorrelation matrices: (a) a matrix that penalizes cross-correlation between activities of close notes and between notes in a common harmonic relation (octave 1:n, major third 5:4 and perfect fifth 3:2), (b) a matrix that encourages temporal smoothness; an exponential smoothness profile was used.

$$D_\varphi(\mathbf{X}, \mathbf{AS}) + \alpha(\mathbf{A}) + \beta(\mathbf{S}), \tag{6}$$

where α and β are the penalty functions. The update rules for NNMA become

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{(\mathbf{X} \odot \varphi''(\mathbf{AS}))\mathbf{S^T}}{(\mathbf{AS} \odot \varphi''(\mathbf{AS}))\mathbf{S^T} + \nabla_\mathbf{A}\alpha(\mathbf{A})}, \tag{7}$$

$$\mathbf{S} \leftarrow \mathbf{S} \odot \frac{(\mathbf{X} \odot \varphi''(\mathbf{AS}))\mathbf{S^T}}{(\mathbf{AS} \odot \varphi''(\mathbf{AS}))\mathbf{S^T} + \nabla_\mathbf{S}\beta(\mathbf{S})}. \tag{8}$$

This allows the user to have greater impact on the resulting factorization. However, caution must be exercised when designing these additional penalty functions, as they might cause the solution to become negative and make the algorithm unstable. Nevertheless, in our experience, using only penalties with positive derivative led to a stable algorithm. Among the note activity matrix penalties used most successfully by us are: the sparseness and the cross-correlation penalties, and the time smoothness objective.

To obtain sparser note activities we employ the l_p-norm with $p < 2$:

$$\beta_1(\mathbf{S}) = \mu_1|\mathbf{S}^p|, \tag{9}$$

$$\nabla_\mathbf{S}\beta_1(\mathbf{S}) = \mu_1 p\mathbf{S}^{p-1}. \tag{10}$$

The cross-correlation penalty can be used to decrease the crosstalk between activities of different notes. The penalty function is defined as:

$$\beta_2(\mathbf{S}) = \mu_2 \sum_{i,j,k} W_{i,j}S_{i,k}S_{j,k} = \mu_2|\mathbf{W} \odot (\mathbf{SS}^T)|, \tag{11}$$

where $\mathbf{W}$ is a weighting matrix. In order to penalize only cross-correlation between different notes, we set $W_{i,i} = 0$. Also, the weights will usually only depend on the interval between the notes and the weighting matrix will become circulant. In this case we simply get:

$$\nabla_\mathbf{S}\beta_2(\mathbf{S}) = 2\mu_2\mathbf{WS} \tag{12}$$

By using this penalty we can also decrease the number of the most common pitch detection errors: octave errors

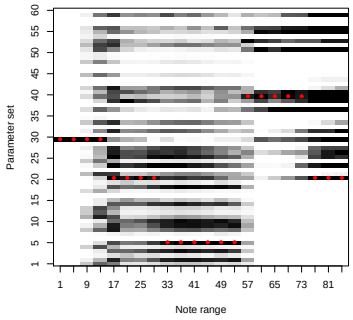


Figure 3: F-measure measured for different parameter sets for different note ranges. Red dots mark the parameter sets chosen for particular note ranges.

(by increasing all weights $W_i \equiv 0 \pmod{12}$), major third errors (by increasing all $W_i = 4$) and perfect fifth errors (by increasing all $W_i = 7$). An example of a weighting matrix constructed in this manner is presented in Fig. 2a.

A very similar penalty can be used to encourage temporal smoothness in a way quite similar to the one presented in [5], but using less complicated penalty function:

$$\beta_3(\mathbf{S}) = -\mu_3 \sum_{i,j,k} V_{i,j} S_{k,i} S_{k,j} = -\mu_3 |\mathbf{V} \odot (\mathbf{S}^T \mathbf{S})|, \quad (13)$$

where $\mathbf{V}$ is a weighting matrix. As with the note decorrelation penalty, using a circulant matrix with nullified main diagonal leads to a simple derivative:

$$\nabla_{\mathbf{S}} \beta_3(\mathbf{S}) = -2\mu_3 \mathbf{S}\mathbf{V}. \quad (14)$$

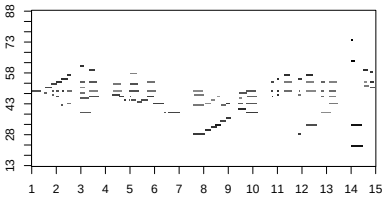
As mentioned before, using regularizations with negative derivative may lead to instability, so we used $\exp(\beta_3(\mathbf{S}))$ in place of 14, which should lead to equivalent solutions thanks to monotonicity of the exponential function. An example of weighting matrix $\mathbf{V}$ is depicted in Fig. 2b.

5. NMF STITCHING

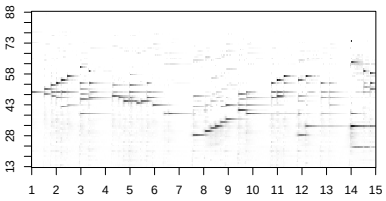
We have noticed that different NMF parameters yield better results for different frequency ranges. This lead to the idea of stitching results of differently parameterized algorithms to get the most optimal results. In our approach we used 4 parameter sets for 5 note ranges (see Fig. 3).

6. HARMONIC FILTERING

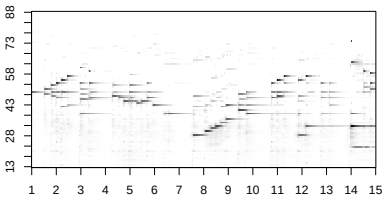
Before the multiple frequency estimation procedure, the input signal was separated in to harmonic and percussive parts using the algorithm presented in [6]. In case of piano recordings, this effectively separates the onset noise from the temporarly smooth notes. The former cas be used for an accurate onset detection, while the latter for multiple frequency estimation unaffected by the often troublesome onset noise.



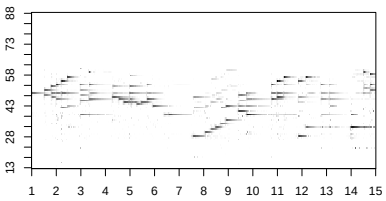
(a) MIDI reference



(b) Regular harmonic NMF

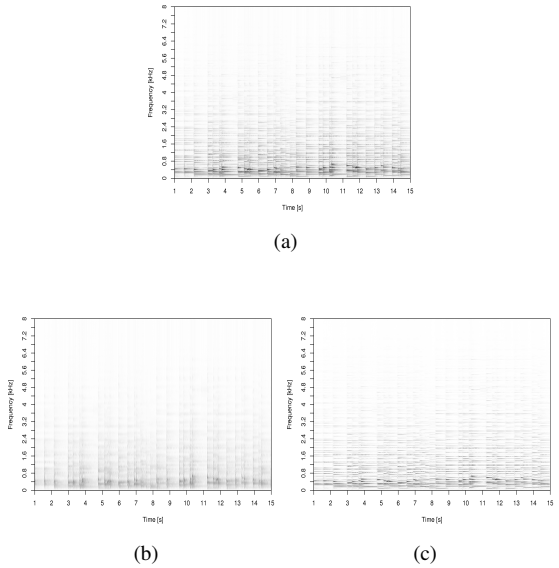


(c) Best parameter set



(d) Stitched NMF

Figure 4: Note activities obtained for different methods with MIDI reference.



(a)

(b) (c)

Figure 5: Separation of power spectrogram (a) into harmonic (b) and percussive (c) parts.

The separation was done by using the following update rules:

$$\widehat{\mathbf{W}} = \mathbf{W}^{\gamma}, \tag{15}$$

$$\widehat{\mathbf{P}} = \mathbf{P}^{\gamma}, \tag{16}$$

$$\widehat{\mathbf{H}} = \mathbf{H}^{\gamma}, \tag{17}$$

$$\alpha = \frac{\sigma_P^2}{\sigma_H^2 + \sigma_P^2}, \tag{18}$$

$$\begin{aligned} \Delta_{\omega,\tau} \quad \leftarrow \quad & \alpha \left(\frac{\widehat{H}_{\omega,\tau-1} - 2\widehat{H}_{\omega,\tau} + \widehat{H}_{\omega,\tau+1}}{4} \right) \\ & -(1-\alpha) \left(\frac{\widehat{P}_{\omega-1,\tau} - 2\widehat{P}_{\omega,\tau} + \widehat{P}_{\omega+1,\tau}}{4} \right) \end{aligned}, \tag{19}$$

$$\widehat{H}_{\omega,\tau} \leftarrow min(\widehat{W}_{\omega,\tau}, max(\widehat{H}_{\omega,\tau} + \Delta_{\omega,\tau}, 0)), \tag{20}$$

$$\widehat{P}_{\omega,\tau} \leftarrow \widehat{W}_{\omega,\tau} - \widehat{H}_{\omega,\tau}. \tag{21}$$

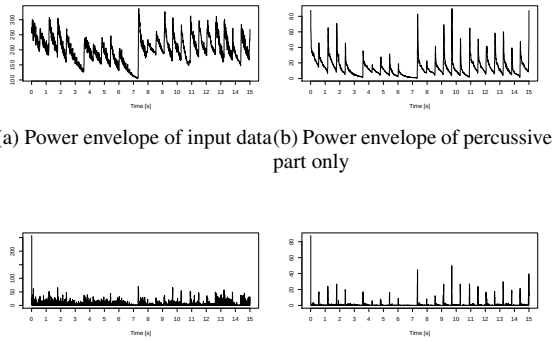
7. ONSET DETECTION

Note detection was performed on the percussive part of the input data. Energy was summed over all frequencies of the power spectra and the resulting power envelope was differentiated and thresholded to give the possible onset positions. The threshold was determined to yield maximal recall, assuring that the correct onset positions were among the detected ones.

8. POSTPROCESSING

A simple postprocessing is applied to the stitched note activity matrix: thresholding, median filtering and short note removal. Threshold has been trained on the Chopin piano database, i.e. it was chosen to minimize the precision with minimal decrease in recall. Its value was determined to be 0.095. Median filter's length was chosen to be between 3 and 5 frames. After that notes that were too short were removed. Minimal note length was determined based on a note envelope model trained on the Chopin piano database.

Finally, notes were detected as group of nonnegative note activities and their onsets were moved to the closest detected onsets, if there were any closer than 100 ms. If there was no onset closer than 500 ms, the note was removed. If a detected onset was close enough to a sufficiently low local minimum in a detected note, the note was split at this position into 2 notes.



(a) Power envelope of input data (b) Power envelope of percussive part only

(c) Rectified difference of power envelope for input data (d) Rectified difference of power envelope for percussive part only

Figure 6: Harmonic filtering helps to greatly improve onset detection accuracy.

9. REFERENCES

[1] I. Dhillon and S. Sra. Generalized Nonnegative Matrix Approximations with Bregman Divergences. *Advances in Neural Information Processing Systems*, 18:283, 2006.

[2] S. Sra and I.S. Dhillon. Nonnegative matrix approximation: Algorithms and applications. Technical report, Technical report, Dept. of Computer Sciences, The Univ. of Texas at Austin, 2006.

[3] R. Kompass. A generalized divergence measure for nonnegative matrix factorization. *Neural computation*, 19(3):780–791, 2007.

[4] S. Raczyński, N. Ono, and S. Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria*, 2007.

[5] K.W. Wilson, B. Raj, and P. Smaragdis. Regularized Non-negative Matrix Factorization with Temporal Dependencies for Speech Denoising.

[6] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proc. of the 9th Int. Conf. on Music Information Retrieval*, pages 139–144, 2008.