Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription

Emmanuel Vincent, Nancy Bertin, Roland Badeau

▶ **To cite this version:**

Emmanuel Vincent, Nancy Bertin, Roland Badeau. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. 2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Mar 2008, Las Vegas, United States. pp.109–112. inria-00544183

HAL Id: inria-00544183
https://inria.hal.science/inria-00544183v1

Submitted on 7 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HARMONIC AND INHARMONIC NONNEGATIVE MATRIX FACTORIZATION FOR POLYPHONIC PITCH TRANSCRIPTION

Emmanuel Vincent

METISS group, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuel.vincent@irisa.fr

Nancy Bertin and Roland Badeau

TSI department, ENST-CNRS LTCI
46 rue Barrault, 75634 Paris Cedex 13, France
nancy.bertin@enst.fr

ABSTRACT

Polyphonic pitch transcription consists of estimating the onset time, duration and pitch of each note in a music signal. This task is difficult in general, due to the wide range of possible instruments. This issue has been studied using adaptive models such as Nonnegative Matrix Factorization (NMF), which describe the signal as a weighted sum of basis spectra. However basis spectra representing multiple pitches result in inaccurate transcription. To avoid this, we propose a family of constrained NMF models, where each basis spectrum is expressed as a weighted sum of narrowband spectra consisting of a few adjacent partials at harmonic or inharmonic frequencies. The model parameters are adapted via combined multiplicative and Newton updates. The proposed method is shown to outperform standard NMF on a database of piano excerpts.

Index Terms— Pitch transcription, nonnegative matrix factorization, harmonicity, inharmonicity, spectral smoothness

1. INTRODUCTION

Western music can be described as a collection of note events, each defined by several attributes: onset time, duration, pitch, instrument, playing style, loudness, *vibrato* rate, *etc*. Polyphonic pitch transcription consists of estimating the first three of these attributes. This task lies at the core of many applications, including content-based retrieval and source separation.

Many transcription methods have been proposed in the literature, based on various signal and/or auditory cues [1]. The best results are typically achieved by training signal models on a given database, while performance decreases for instruments or recording conditions absent from this database [2]. This issue can be addressed via transcription methods based on adaptive signal models, such as Nonnegative Matrix Factorization (NMF) or sparse decomposition [3, 4, 5]. These models represent the short-term magnitude spectrum of the signal as the sum of basis spectra scaled by time-varying amplitudes, which are adapted by minimizing the residual. Pitch identification is then applied to each basis spectrum and onset detection to each amplitude sequence. These methods rely on the assumption that the estimated basis spectra are clearly either pitched or unpitched and that pitched spectra involve a single pitch. However this assumption is often violated, particularly over short excerpts where the likelihood that two pitches are always simultaneously active is higher. Models involving a spectral shift invariance constraint implicitly tend to avoid this [6], but do not account for variations of spectral envelope or inharmonicity over the whole pitch range.

In this paper, we propose a family of NMF models with explicit pitch constraints over the basis spectra. We extend our preliminary study [5] about harmonic constraints by considering inharmonic constraints and adaptive tuning. Also, we assess the robustness of the proposed method regarding the choice of the hyper-parameters. We describe the constrained NMF models in Section 2 and the associated transcription algorithms in Section 3. We evaluate their performance on piano excerpts in Section 4 and conclude in Section 5.

2. CONSTRAINED NMF MODELS

As with standard NMF [3], we represent the magnitude time-frequency transform X_{ft} of a signal in frequency bin f and time frame t as the sum of basis spectra scaled by time-varying amplitudes. However, we explicitly associate each basis spectrum with an integer pitch p on the MIDI semitone scale in the range from p_{low} to p_{high}. Each pitch p then corresponds to I_p basis spectra with different spectral envelopes indexed by i. This model can be written as

$$X_{ft} = \sum_{p=p_{\text{low}}}^{p_{\text{high}}} \sum_{i=1}^{I_p} A_{pit} S_{pif} + R_{ft} \qquad (1)$$

where S_{pif} and A_{pit} are the basis spectra and amplitude sequences associated with pitch p and R_{ft} is the residual.

2.1. General formulation of the constraints

In order to ensure that each estimated basis spectrum S_{pif} has the expected pitch value p while retaining the ability of NMF

This work was supported by ANR young researchers project DESAM.

to adapt to the spectral envelopes of various instruments, we constrain the basis spectra as

$$S_{pif} = \sum_{k=1}^{K_p} E_{pik} P_{pkf} \tag{2}$$

where P_{pkf} are K_p narrowband spectra representing adjacent sinusoidal partials at harmonic or inharmonic frequencies and the coefficients E_{pik} model the spectral envelope. Each narrowband spectrum P_{pkf} is defined by summation of the spectra of individual partials of unit amplitude, scaled by the spectral shape of subband k. By contrast with the ad-hoc approach [7] modeling each partial by a single non-zero frequency bin, we compute the exact spectrum of the mth partial given its frequency f_{pm} as $G_f(f_{pm})$ where G_f is the magnitude frequency response associated with bin f of the transform.

2.2. Harmonicity, inharmonicity and tuning

The frequencies of the partials vary depending on the signal. We consider six possible models for these frequencies, based on the combination of three tuning models and two overtone models. The frequency $f_{p,1}$ of the first partial is assumed to be either fixed with standard 440 Hz tuning for $p = 69$

$$f_{p,1} = 440 \times 2^{\frac{p-69}{12}} \tag{3}$$

or shifted by a common tuning factor $q \in [-\frac{1}{2}, \frac{1}{2})$

$$f_{p,1} = 440 \times 2^{\frac{p+q-69}{12}} \tag{4}$$

or shifted by a specific tuning factor $q_p \in [-\frac{1}{2}, \frac{1}{2})$ [8]

$$f_{p,1} = 440 \times 2^{\frac{p+q_p-69}{12}}. \tag{5}$$

In addition, the frequencies f_{pm} of overtone partials are assumed to be either harmonic

$$f_{pm} = m f_{p,1} \tag{6}$$

or inharmonic with inharmonicity factor $b_p \geq 0$ [8]

$$f_{pm} = m f_{p,1} \sqrt{\frac{1 + b_p m^2}{1 + b_p}}. \tag{7}$$

2.3. Spectral smoothness

The width of the subbands affects transcription performance. When the narrowband spectra P_{pkf} contain a single partial, the basis spectra S_{pif} can represent multiples of the expected fundamental frequency, resulting in upper octave errors. When the narrowband spectra contain too many partials, the basis spectra do not adapt well to the spectral envelope of the instruments, leading to note insertion or deletion errors.

We assume uniformly spaced subbands on the Equivalent Rectangular Bandwidth (ERB) scale [9] defined by $f_{\mathrm{ERB}} = 9.26 \log(0.00437 f_{\mathrm{Hz}} + 1)$. The number K_p of subbands is set so that the center of the last subband is below the Nyquist frequency, with a maximum number of $K_{\max}$ subbands. The first subband is centered at $f_{p,1}$ and subsequent subbands are spaced by $B_{\max}/K_{\max}$ ERB, with $B_{\max}$ the maximum total bandwidth. We define the spectral shape of subband k as the symmetric approximation of the response of the gammatone filter of bandwidth $B_{\max}/K_{\max}$ modeling this subband, as given in [9]. Example spectra are depicted in Figure 1.

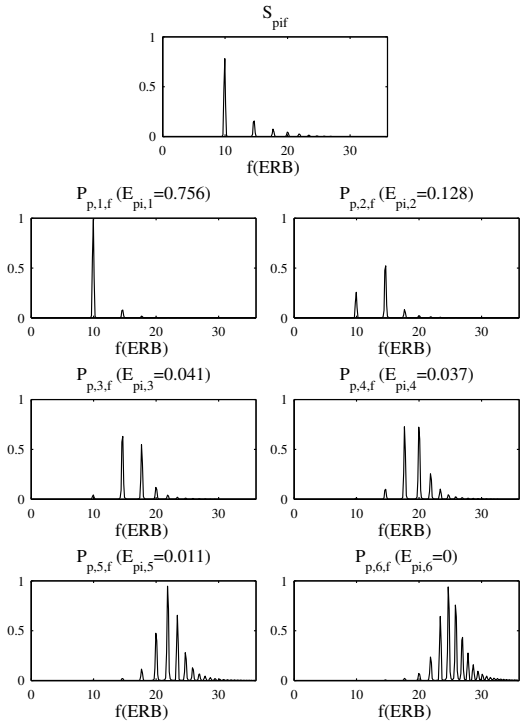

Fig. 1. Harmonic basis spectrum S_{pif} and underlying narrowband spectra P_{pkf} ($p = 69$, $B_{\max} = 18$, $K_{\max} = 6$).

This model appears well-motivated from a cognitive point of view, since the perception of pitch is based on the detection of periodicities within each auditory band [1]. A similar model was used in [10] for the different task of source separation given the fundamental frequencies of all notes.

3. TRANSCRIPTION ALGORITHMS

The above models can be employed for pitch transcription by detecting note events from the amplitude sequences A_{pit}. In practice, different time-frequency representations, parameter adaptation and onset detection algorithms can be chosen.

3.1. ERB-scale time-frequency representation

In the following, we use the ERB-scale representation in [11]. The signal is passed through a bank of F filters, defined as sinusoidally modulated Hanning windows with frequencies

linearly spaced between 0 and 36 ERB. The main-lobe bandwidth of each filter is set to four times the frequency difference between this and adjacent filters. The root-mean-square magnitude of the filtered signals is then computed over disjoint 23 ms time frames. This representation was shown to yield similar transcription performance as the short-time Fourier transform at lower computation cost [5]. The magnitude frequency response G_f of the fth filter can then be analytically computed as a combination of sine cardinal functions.

3.2. Adaptation of the model parameters

The model parameters are adapted by minimizing the residual loudness, as measured by the weighted Euclidean norm [12] $\mathcal{L} = \sum_{f,t} W_{ft} R_{ft}^2$ with perceptual weights W_{ft} depending on X_{ft}. This criterion improves the transcription of low energy notes compared to the usual Euclidean norm [5]. Other criteria, such as those proposed in [13], are not considered.

The amplitudes sequences A_{pit}, the envelope coefficients E_{pik} and the tuning and/or inharmonicity factors q, q_p and b_p are randomly initialized and alternatingly updated until a local minimum of $\mathcal{L}$ is achieved. Since the model is linear in A_{pit} and E_{pik}, these parameters are adapted under nonnegativity constraints via multiplicative updates derived from [14]

$$A_{pit} \leftarrow A_{pit} \frac{\sum_f S_{pif} W_{ft} X_{ft}}{\sum_f S_{pif} W_{ft} \sum_{p',i'} A_{p'i't} S_{p'i'f}} \quad (8)$$

$$E_{pik} \leftarrow E_{pik} \frac{\sum_{f,t} A_{pit} P_{pkf} W_{ft} X_{ft}}{\sum_{f,t} A_{pit} P_{pkf} W_{ft} \sum_{p',i'} A_{p'i't} S_{p'i'f}}. \quad (9)$$

By contrast, the model is nonlinear in q, q_p and b_p, so these parameters can be updated via any Newton-based optimizer given the gradient and Hessian of $\mathcal{L}$. These quantities can be derived from the analytical expressions of the first and second order derivatives of G_f. We choose Matlab's `fmincon` optimizer[1] with a diagonal approximation of the Hessian.

3.3. Threshold-based onset/offset detection

Once the model parameters have converged, note events are transcribed using simple threshold-based activity detection. We associate each pitch p with a summary amplitude sequence $\bar{A}_{pt}$ defined by $\bar{A}_{pt} = [\sum_f (\sum_{i=1}^{I_p} A_{pit} S_{pif})^2]^{1/2}$. A note onset is detected each time $\bar{A}_{pt}$ becomes larger than $A_{\text{thr}} \times \max_{p,t} \bar{A}_{pt}$ for at least 3 consecutive frames, where A_{thr} is a fixed threshold. The same principle is used for note offsets.

4. EVALUATION

The proposed transcription algorithm was applied to a set of 43 Disklavier piano excerpts of 30 s duration involving a total of 9489 notes [2] with a polyphony level of 2.1 on average and 7 at most. The standard NMF algorithm in [12] was also evaluated for comparison using 88 components. A given note was considered to be correctly transcribed if its pitch was equal to the ground truth and its onset time was within 50 ms of the ground truth. Performance was then classically assessed in terms of recall $\mathcal{R}$, precision $\mathcal{P}$ and F-measure $\mathcal{F}$ [15]. The full pitch range of the piano between $p_{\text{low}} = 21$ and $p_{\text{high}} = 108$ was considered. The best onset detection threshold A_{thr} was determined to be -23 dB for all algorithms.

Table 1. Average performance for the transcription of piano excerpts. The computation time is indicated per 30 s excerpt with Matlab 7.5 on a 1.2 GHz dual core laptop.

NMF	Tuning	$\mathcal{P}$ (%)	$\mathcal{R}$ (%)	$\mathcal{F}$ (%)	Time (min)
Standard	N/S	70.9	79.3	74.3	2
Harmonic	Fixed	88.0	87.1	87.3	1
	Common	86.9	87.2	86.8	8
	Pitchwise	84.8	87.3	85.7	30
Inharmonic	Fixed	84.2	86.9	85.3	60
	Common	84.4	86.7	85.2	60
	Pitchwise	84.3	86.9	85.4	80

Average results are given in Table 1 for the following values of the hyper-parameters: $F = 250$, $I_p = 1$, $B_{\max} = 18$, $K_{\max} = 6$. Harmonic NMF with fixed tuning achieved a F-measure of 87%. This is 13% better than standard NMF, 10% better than the piano-specific transcription algorithm proposed in [2] and 2% better than the piano-specific SONIC software[2] for the same data. This is also in line with results recently reported in the literature [2] and at the latest Music Information Retrieval Evaluation eXchange[3], with F-measures between 17% and 83% for similar piano data. Surprisingly, inharmonicity constraints and pitchwise adaptive tuning did not further improve performance, but decreased the F-measure by 2% instead. The estimated inharmonicity and tuning factors appeared accurate for the pitches actually present in the signal but grossly inaccurate for other pitches, resulting in spurious estimated notes. We expect that this could be avoided using a better model of these factors, involving smoothness constraints and maximum inharmonicity values for example.

Additional experiments with different hyper-parameter values showed that performance did not significantly change when considering $I_p > 1$ basis spectra per pitch and that F, $B_{\max}$ and $K_{\max}$ had roughly independent effects on performance, which are illustrated in Figure 2. Harmonic NMF is seen to be more robust than standard NMF with respect to the chosen number of frequency bins F. Also, it performs well for a maximum basis spectrum bandwidth $B_{\max}$ between about 13 and 23 ERB, corresponding to a range of 6 to 20 modeled partials, and it is robust to an over-estimation of the maximum number $K_{\max}$ of subbands per basis spectrum.

[1] This optimizer is based on a subspace trust region. For more details, see www.mathworks.com/access/helpdesk_r13/help/toolbox/optim/fmincon.html

[2] http://lgm.fri.uni-lj.si/sonic.html

[3] http://www.music-ir.org/mirex2007/

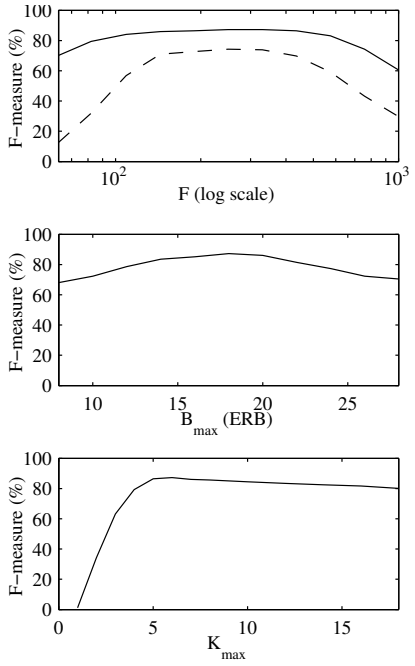

Fig. 2. Performance of standard NMF (dashed) and harmonic NMF with fixed tuning (plain) as a function of the number of frequency bins F, the maximum bandwidth $B_{\max}$ of the basis spectra and the maximum number $K_{\max}$ of subbands.

5. CONCLUSION

We proposed a family of NMF-based polyphonic pitch transcription methods using harmonic or inharmonic constraints on the basis spectra with fixed or adaptive tuning and variable spectral smoothness. These methods outperformed standard NMF and achieved a transcription accuracy comparable to the state-of-the-art on a set of piano excerpts. In the future, we plan to improve onset detection within the NMF framework by modeling the amplitude sequences as weighted sums of delayed amplitude sequences. We will also investigate the use of harmonic spectra learned on isolated notes as the basis for the definition of narrowband harmonic spectra.

6. REFERENCES

[1] A. Klapuri and M. Davy, *Signal processing methods for music transcription*, Springer, New York, NY, 2006.

[2] J.P. Bello, L. Daudet, and M.B.Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2242–2251, 2006.

[3] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[4] S.A. Abdallah and M.D. Plumbley, "Unsupervised analysis of polyphonic music using sparse coding," *IEEE Trans. on Neural Networks*, vol. 17, no. 1, pp. 179–196, 2006.

[5] E. Vincent, N. Bertin, and R. Badeau, "Two nonnegative matrix factorization methods for polyphonic pitch transcription," in *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.

[6] M. Kim and S. Choi, "Monaural music source separation: nonnegativity, sparseness and shift-invariance," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2006, pp. 617–624.

[7] S.A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2007, pp. 381–386.

[8] L.I. Ortiz-Berenguer, F.J. Casajús-Quirós, M. Torres-Guijarro, and J.A. Beracoechea, "Piano transcription using pattern recognition: aspects on parameter extraction," in *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2004, pp. 212–216.

[9] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psycho-acoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1805–1808.

[10] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1757–1760.

[11] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[12] E. Vincent and M.D. Plumbley, "Low bitrate object coding of musical audio using Bayesian harmonic models," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1273–1282, 2007.

[13] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[14] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, 2001, pp. 556–562.

[15] C.J. van Rijsbergen, *Information retrieval, 2nd Edition*, Butterworths, London, UK, 1979.