Two nonnegative matrix factorization methods for polyphonic pitch transcription

Emmanuel Vincent, Nancy Bertin, Roland Badeau

TWO NONNEGATIVE MATRIX FACTORIZATION METHODS 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 within a music signal. Adaptive signal models such as Nonnegative Matrix Factorization (NMF) appear well suited to this task, since they can provide a meaningful representation whatever instruments are playing. In this paper, we propose a simple transcription method using minimum residual loudness NMF, harmonic comb-based pitch identification and threshold-based onset/offset detection, and investigate a second method incorporating harmonicity constraints in the NMF model. Both methods are evaluated in the framework of MIREX 2007 [1] .

1 INTRODUCTION

Western music signals can be described as a collection of note events defined by several attributes: onset time, duration, pitch, instrument class, 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.

Various approaches have been proposed so far, based on computational auditory models or probabilistic signal models. Successful methods often rely on instrument-specific models, so that their performance decreases for other instruments [3]. By contrast, adaptive signal models, such as independent component analysis, Nonnegative Matrix Factorization (NMF) or adaptive sparse decomposition, can provide a meaningful representation whatever instruments are playing. Early transcription methods based on such models relied on visualization [5] or auditory pitch estimation [1]. More recently, some of us proposed a fully automatic NMF-based method [2].

In the following, we devise an improved variant of this method and investigate a new method incorporating harmonicity constraints in the NMF model. The structure of the rest of the paper is as follows: we describe the proposed methods in Sections 2 and 3, evaluate their performance in Section 4 and conclude in Section 5.

[1] `http://www.music-ir.org/mirexwiki/`

2 BASELINE NMF METHOD

NMF provides an approximate model of a magnitude time-frequency representation as the sum of basis spectra scaled by time-varying amplitudes [5]. Derived transcription methods typically involve four processing steps:

1. magnitude time-frequency representation,
2. approximate decomposition by NMF,
3. pitch identification applied to each basis spectrum,
4. onset detection applied to each amplitude sequence.

The method in [2] addressed these steps using respectively the Short-Time Fourier Transform (STFT), minimum divergence NMF, spectral product-based pitch identification and threshold-based onset detection.

2.1 ERB-scale time-frequency representation

In order to discriminate musical pitches, the time-frequency representation must have a frequency resolution of at least one semitone over the whole frequency range. In the case of the STFT, a large window length is thus needed (64 ms in [2]), inducing both a low temporal onset resolution and a large computation cost.

A representation of smaller size with better temporal resolution in the higher frequency range can be obtained by using a nonlinear frequency scale. We use the auditory-motivated representation proposed in [6] as a front-end for instrument-specific models. We pass the signal through a filterbank of 257 sinusoidally modulated Hanning windows with frequencies linearly spaced between 5 Hz and 10.8 kHz on the Equivalent Rectangular Bandwidth (ERB) scale [11] defined by $f_{\mathrm{ERB}} = 9.26 \log(0.00437 f_{\mathrm{Hz}} + 1)$. We set the length of each filter so that the bandwidth of its main frequency lobe equals four times the difference between its frequency and those of adjacent filters. We then split each subband into disjoint 23 ms time frames and compute the square root of the power within each frame.

2.2 Minimum residual loudness NMF

The standard NMF model can be written as [5]

$$X_{ft} = \left(\sum_{i=1}^{I} H_{it} W_{if} \right) + R_{ft} \qquad (1)$$

where X_{ft} denotes the input magnitude in time-frequency bin (t, f), W_{if} and H_{it} are the basis spectrum and the amplitude sequence of component i, and R_{ft} is the residual. The parameters of the model are adapted by minimizing the residual according to a certain measure. Common measures include the Euclidean norm and a particular divergence [5]. These measures favor a smaller relative residual R_{ft}/X_{ft} in the time-frequency bins of high magnitude X_{ft}. Due to the large amplitude range of music, most components are thus adapted to high energy notes, while low energy notes may be not modeled at all.

We measure instead the (auditory) loudness of the residual by the weighted Euclidean norm defined in [8], which already provided good results for melody transcription [7]. The use of such a measure with NMF was first suggested in the context of source separation in [9]. It associates a larger weight to low energy time-frequency points and accounts for basic auditory masking rules [11]. The parameters of the NMF model are randomly initialized and iteratively estimated using the multiplicative update rules given in [9]. Each basis spectrum is normalized to unit power.

2.3 Harmonic comb-based pitch identification

The basis spectra estimated by NMF can be either pitched or unpitched. A given pitch value may be represented by several pitched basis spectra with harmonic partials at the same frequencies but with different amplitudes. Spectral product-based pitch identification techniques often fail if some partials have zero amplitude.

Harmonic comb-based techniques are more robust to this issue. We use the simple sinusoidal comb [7]

$$f_0^i = \arg\min_{f_0} \sum_{f=1}^{F} W_{if}^2 \left[1 - \cos(2\pi\nu_f/f_0)\right] \qquad (2)$$

where ν_f is the frequency of bin f. The acceptable pitch range is set between 27 Hz and 4.3 kHz, which is the range of the piano. Basis spectra with an estimated pitch outside this range are classified as unpitched and discarded.

2.4 Threshold-based onset/offset detection

A single amplitude sequence is associated to each discrete pitch on the semitone scale by summing the corresponding NMF components and taking the square root of their total power in each time frame. These amplitude sequences are then processed to detect note onsets. A simple threshold-based detection technique was used in [2].

We use the same principle but with a different threshold defined as A times the maximum observed amplitude over all pitches and all time frames. Notes shorter than 50 ms are removed.

3 HARMONIC NMF METHOD

The above transcription method is based on the assumption that the basis spectra estimated by NMF are clearly either pitched or unpitched and that pitched spectra involve a single pitch. In practice, the lack of constraints in the NMF model often leads to violations of this assumption. One way of enforcing it is to incorporate harmonicity constraints in the NMF model by associating a fixed fundamental frequency f_0^i to each basis spectrum and constraining it as

$$W_{if} = \sum_{k=1}^{K_i} E_{ik} P_{ikf} \qquad (3)$$

where P_{ikf} is a fixed narrowband spectrum consisting of a few adjacent partials at harmonic frequencies of f_0^i. The weights E_{ik} model the spectral envelope. Since this model is linear, the minimization of the residual loudness can still be addressed using multiplicative updates. The estimated basis spectra are then guaranteed to be pitched with known fundamental frequencies, while the ability of NMF to adapt to the spectral envelope of various instruments is retained.

In the following, we assume that the bands k are linearly spaced on the ERB scale with a step of N ERB. The first band is centered at f_0^i and the number K_i of bands is set so that the center of the last band is below the Nyquist frequency, with a maximal number of $K_{\max}$ bands. We define P_{ikf} as the product of a harmonic spectrum with unit amplitude partials by the frequency response of the gammatone filter [11] of bandwidth N modeling band k. A similar model was used in [10] for source separation given the fundamental frequencies of all notes, but with separate adaptation of the spectral envelopes on each time frame.

4 EVALUATION

4.1 Choice of the parameters

The two proposed transcription methods were applied to a set of 43 Disklavier piano excerpts of 30 s duration [2], containing 34 different pitches each on average. The number of components I was set to multiples of 34 for the baseline method and to multiples of 88 for the harmonic method, with 88 semitone-spaced fundamental frequencies assuming 440 Hz tuning and one or more spectral envelope components per fundamental frequency. The onset detection threshold A, the bandwidth N and the maximal number of bands $K_{\max}$ were varied manually by increments of 1 dB, 0.25 ERB and 10 respectively. The accuracy of the estimated pitches and onsets was then assessed by the F-measure, with a ± 50 ms tolerance for onsets.

The best results were obtained with $A = -22$ dB, $I = 68$ for the baseline method and $I = 88$, $N = 1.75$ and $K_{\max} = 10$ for the harmonic method. These settings resulted in average F-measures of 73% and 84%, the latter being only 1% below that of the piano-specific SONIC software [2]. By comparison, our previous method resulted in an average F-measure on the order of 50% [2].

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

Additional experiments showed that the use of the residual loudness measure, that of harmonic comb-based pitch identification and the new definition of the onset detection threshold resulted in similar performance increases. The ERB-scale time-frequency representation did not significantly change performance compared to the STFT, but reduced the computation time by about 35%.

4.2 Results

Both methods were also evaluated within the MIREX 2007 evaluation framework for Multiple Fundamental Frequency Estimation & Tracking, using $I = 88$ and the above optimal settings for other parameters. The test data consisted of 10 piano excerpts and 28 excerpts with two to four instruments taken from a multitrack woodwind quintet recording or synthesized from MIDI, with 30 s duration each.

Using the above performance criterion (Task II), our methods scored 4th and 2nd among 11 methods, with average F-measures of 45.3% and 52.7%. Using a different criterion measuring the pitch accuracy over 10 ms frames (Task I), our methods scored 8th and 5th among 16 methods, with average accuracies of 46.6% and 54.3%. By comparison, the best method achieved a F-measure of 61.4% and an accuracy of 60.5%. This suggests that the accuracy of the pitches estimated via our harmonic NMF-based method is close to the state-of-the-art, while that of the estimated onsets could be further improved.

Interestingly, this method relies on similar principles as other top-scoring ones, including spectral smoothness, bandwise power compression and power-based pitch salience measurement. Also it performed slightly better than a concurrent harmonic NMF-based method [4] representing each partial by a single non-zero frequency bin and constraining all notes to have the same partial amplitudes, regardless of their fundamental frequency.

5 CONCLUSION

We proposed two NMF-based polyphonic pitch transcription methods using either unconstrained basis spectra or harmonically constrained spectra represented as weighted sums of narrowband spectra consisting of a few adjacent partials. The latter provided a pitch accuracy close to the state-of-the-art. In the future, we plan to improve the onset detection accuracy within the NMF framework by representing the amplitude sequences as weighted sums of delayed amplitude sequences learned on isolated note signals. 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] S.A. Abdallah and M.D. Plumbley. Unsupervised analysis of polyphonic music using sparse coding. *IEEE Trans. on Neural Networks*, 17(1):179–196, 2006.

[2] N. Bertin, R. Badeau, and G. Richard. Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 65–68, 2007.

[3] M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Trans. on Multimedia*, 6(3):439–449, 2004.

[4] 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.

[5] 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)*, pages 177–180, 2003.

[6] E. Vincent. Musical source separation using time-frequency source priors. *IEEE Trans. on Audio, Speech and Language Processing*, 14(1):91–98, 2006.

[7] E. Vincent and M.D. Plumbley. Predominant-F0 estimation using Bayesian harmonic waveform models. In *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, 2005.

[8] 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*, 15(4):1273–1282, 2007.

[9] T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2004.

[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)*, volume 2, pages 1757–1760, 2002.

[11] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models, 2nd Edition*. Springer, Heidelberg, 1999.