FAST BAYESIAN CONSTRAINED NMF FOR POLYPHONIC PITCH TRANSCRIPTION

Nancy Bertin and Emmanuel Vincent
METISS group, IRISA-INRIA
Campus de Beaulieu
35042 Rennes Cedex, France
nancy.bertin@irisa.fr

Roland Badeau
Institut Télécom, Télécom ParisTech
46 rue Barrault
75634 Paris Cedex 13, France
roland.badeau@telecom-paristech.fr

ABSTRACT

This extended abstract presents a submission to the Music Information Retrieval Evaluation eXchange (MIREX) in the Multiple Fundamental Frequency Estimation & Tracking task. This submission is mainly based on our previous work on Nonnegative Matrix factorization (NMF) applied to music transcription. It relies on the harmonicity model we used in our previous participation in MIREX [1] and more recent improvements, including a statistical approach to a temporal continuity constraint and efficient multiplicative update rules.

1. INTRODUCTION

Out of any applicative context, the NMF problem is expressed as follows: given a matrix $\mathbf{V}$ of dimensions $F \times N$ with non-negative entries, NMF is the problem of finding a factorization $\hat{\mathbf{V}} \triangleq \mathbf{WH} \approx \mathbf{V}$, where $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $FK + KN \ll FN$, hence reducing the data dimension. In typical audio applications, the matrix $\mathbf{V}$ is often the magnitude or power spectrogram, f denoting the frequency bin and n the time frame. This factorization is obtained by minimizing a cost function. Multiplicative update rules realizing this minimization may follow a simple heuristics, which can be seen as a gradient descent algorithm with an appropriate choice of the descent step. They are obtained by expressing the partial derivatives of the cost function ∇D as the difference of two positive terms $\nabla^+ D$ and $\nabla^- D$:

$$\begin{cases} w_{fk} \leftarrow w_{fk} \times \dfrac{\nabla^-_{w_{fk}} D(\mathbf{V}|\mathbf{WH})}{\nabla^+_{w_{fk}} D(\mathbf{V}|\mathbf{WH})} \\ h_{kn} \leftarrow h_{kn} \times \dfrac{\nabla^-_{h_{kn}} D(\mathbf{V}|\mathbf{WH})}{\nabla^+_{h_{kn}} D(\mathbf{V}|\mathbf{WH})} \end{cases} \quad (1)$$

NMF has shown to be a way to perform polyphonic pitch transcription with efficiency and few prior knowledge on the signal [2]. However, this lack of constraints may result in ambiguous (e.g. unpitched) components. Here, we focus on two constraints that sound relevant to music transcription: harmonicity of the dictionary components, already used in our previous submission, and temporal smoothness of the decomposition. A Bayesian approach developed in [3, 4] is a way to induce those properties and compute the factorization through an EM-based algorithm. In order to reduce the computational burden, we herein use an alternative approach inspired by multiplicative heuristics [5].

2. CONSTRAINED NONNEGATIVE MATRIX FACTORIZATION

2.1 Harmonicity

Musical notes, excluding transients, are pseudo-periodic. Their spectra consist in regularly spaced frequency peaks. As we wish to use NMF to identify musical notes in a polyphonic recording, we expect that elements in the basis $\mathbf{W}$ follow this harmonic shape. This idea was exploited, for instance, in [6].

In [1], we proposed an alternative model enforcing harmonicity. We impose the basis components to be expressed as the linear combination of fixed narrow-band harmonic spectra (patterns):

$$w_{fk} = \sum_{m=1}^{M} e_{mk} P_{km}(f). \quad (2)$$

For a given component index k, all the patterns $P_{km}(f)$ share the same pitch (fundamental frequency f_0); they are defined by summation of the spectra of a few adjacent individual partials at harmonic frequencies of f_0, scaled by the spectral shape of sub-band k. This spectral envelope is chosen according to perceptual modeling. $\mathbf{E}$ can be interpreted as global frequency envelope coefficients for one component $\mathbf{w}_k$.

2.2 Temporal continuity

Continuity [7] Another way to induce properties in NMF is to switch to a statistical framework and introduce adequate

prior distributions. Let us consider a complex-valued time-frequency representation $\mathbf{X}$ of the signal, and the following model: $\forall n = 1, ..., N$,

$$\mathbf{x}_n = \sum_{k=1}^{K} \mathbf{c}_{kn} \in \mathbb{C}^F \tag{3}$$

where latent variables $\mathbf{c}_{kn}$ are independent and follow a multivariate complex Gaussian distribution [1] :

$$\mathbf{c}_{kn} \sim \mathcal{N}(0, h_{kn}\text{diag}(\mathbf{w}_k)).$$

The estimation of the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ in a maximum likelihood (ML) sense is performed by maximizing the criterion

$$C_{ML}(\boldsymbol{\theta}) \triangleq \log p(\mathbf{X}|\boldsymbol{\theta}). \tag{4}$$

Then it is easily proved that $C_{ML}(\boldsymbol{\theta}) \stackrel{c}{=} -D(\mathbf{V}|\hat{\mathbf{V}})$, where $\mathbf{V} = |\mathbf{X}|^2$, and the cost function d is the Itakura-Saito (IS) divergence:

$$d_{IS}(a|b) = \frac{a}{b} - \log\frac{a}{b} - 1. \tag{5}$$

Thus ML estimation is equivalent to solving the NMF problem $\mathbf{V} \approx \mathbf{WH}$ (see [3] for a full study and justification of this model).

This approach offers the possibility to switch to maximum a posteriori (MAP) estimation, thanks to Bayes rule:

$$p(\mathbf{W}, \mathbf{H}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{W}, \mathbf{H})p(\mathbf{W})p(\mathbf{H})}{p(\mathbf{X})} \tag{6}$$

Thus, choosing adequate prior distributions $p(\mathbf{W})$ and $p(\mathbf{H})$ is a way to induce desired properties in the decomposition. We decide here to adopt a priori information on $\mathbf{H}$, expressed as a prior distribution $p(\mathbf{H})$, to enforce its temporal continuity. Thanks to Bayes rule (6), we get a maximum a posteriori (MAP) estimator by maximizing the following criterion:

$$\begin{aligned} C_{MAP}(\boldsymbol{\theta}) &\triangleq \log p(\boldsymbol{\theta}|\mathbf{X}) \\ &\stackrel{c}{=} C_{ML}(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \end{aligned}$$

We choose the Markov chain prior structure proposed in [3]:

$$p(h_k) = p(h_{k1}) \prod_{n=2}^{N} p(h_{kn}|h_{k(n-1)}) \tag{7}$$

where $p(h_{kn}|h_{k(n-1)})$ reaches its maximum at $h_{k(n-1)}$, thus favoring a slow variation of h_k in time. We propose to choose

$$p(h_{kn}|h_{k(n-1)}) = \mathcal{IG}(h_{kn}|\alpha_k, (\alpha_k + 1)h_{k(n-1)}) \tag{8}$$

[1] Gaussian distribution: $\mathcal{N}(\mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\det(\pi\boldsymbol{\Sigma})} e^{-(\mathbf{u}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})}$, where the symbol ^{H} denotes the conjugate transpose.

where $\mathcal{IG}(u|\alpha, \beta)$ is the inverse-Gamma distribution [2] with mode $\beta/(\alpha + 1)$ and the initial distribution $p(h_{k1})$ is Jeffrey's non-informative prior: $p(h_{k1}) \propto 1/h_{k1}$. We do not put here any prior on $\mathbf{E}$.

Inferring the parameters in this model may be done by two different approaches: EM-like derivation of update rules, or multiplicative heuristics rules. [4] Here, we directly use the update rules (1) with

$$-C^{MAP}(\boldsymbol{\theta}) = D_{IS}(\mathbf{V}|\mathbf{WH}) - \sum_{k=1}^{K} \log(p(h_k)),$$

where the contribution of the prior can be seen as a penalty term. Updates for $\mathbf{E}$ are unchanged and we obtain new update rules for $\mathbf{H}$. However, first simulation experiments showed that under this update scheme, the criterion was not always monotonically decreasing. Then, we propose to raise the ratio in (1) to a certain power $\eta \in]0, 1[$, whose role is similar to the step size in usual gradient descents. We then obtain the following update rules: for $n = 2 \ldots N-1$,

$$h_{kn} \leftarrow h_{kn} \times \left(\frac{\sum_{f=1}^{F} \frac{v_{fn} w_{fk}}{\hat{v}_{fn}^2} + \frac{(\alpha_k+1)h_{k,n-1}}{h_{kn}^2}}{\sum_{f=1}^{F} \frac{w_{fk}}{\hat{v}_{fn}} + \frac{1}{h_{kn}} + \frac{\alpha_k+1}{h_{k,n+1}}} \right)^{\eta} \tag{9}$$

Similar updates are determined for the boundaries of the Markov chain ($n = 1$ and $n = N$).

3. APPLICATION TO POLYPHONIC PITCH TRANSCRIPTION

3.1 Transcription system

NMF provides an approximate model of a magnitude time-frequency representation as the sum of basis spectra scaled by time-varying amplitudes [2]. Derived transcription methods typically involve four processing steps:

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

In this submission, we use the same framework as in our previous one [1]:

1. We pass the signal through a lterbank of 257 sinusoidally modulated Hanning windows with frequencies linearly spaced between 5 Hz and 10.8 kHz on the Equivalent Rectangular Bandwidth (ERB) scale. We set the length of each lter so that the bandwidth of its main frequency lobe equals four times the difference between its frequency and those of adjacent lters. We then split each subband into disjoint 23 ms time frames and compute the square root of the power within each frame.

[2] Inverse-Gamma: $\mathcal{IG}(u|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} u^{-(\alpha+1)} e^{-(\beta/u)}$, $u \geq 0$.

2. Pitch of component number k is set to the associated pitch in the fixed narrowband spectra P_{kf}

3. 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. We use a simple threshold-based detection technique described in [1]. Notes shorter than 50 ms are removed.

3.2 Choice of the parameters

The parameters were optimized on a set of 60 piano excerpts recorded on a DisKlavier or obtained by high quality software synthesis. The number of components K was set to 88, with 88 semitone-spaced fundamental frequencies assuming 440 Hz tuning, and one spectral envelope component per fundamental frequency. The onset detection threshold A, the maximal number of bands M and the shape parameters α_k were respectively set to -50 dB, 10 and 20 for all components. The descent step η is 0.4. More results on these data are available in [4].

4. CONCLUSION

Despite good results on piano data [4,5], the submitted system performed poorly on MIREX 2009 database. Several explanations may be considered:

- Woodwind instruments and piano produce considerably different sounds, which may put our assumptions on harmonicity and smoothness in default;
- 440 Hz tuning is a limitation; although our experiments on different (adaptive) tuning previously failed to enhance performance, this should be investigated on these new data;
- The shape parameter α, which is here fixed, may be not adapted to the temporal characteristics of the sounds analyzed in this contest; learning this paramater rather than fixing it could be an option to avoid this risk;
- Due to computational cost issues, we had to limit the number of iterations, and to take this constraint into account in the choice of step descent η, which also may have had an influence on the results.

Since the results are very disappointing compared to previous results, the possibility of an implementation problem is also not excluded.

5. REFERENCES

[1] E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 109–112, Las Vegas, Nevada, USA, March 30 - April 4, 2008.

[2] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, pages 177–180, New Paltz, New York, USA, October 19-22 2003.

[3] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[4] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *To appear in IEEE Trans. on Audio, Speech and Language Processing*, 2008.

[5] N. Bertin, R. Badeau, and E. Vincent. Fast bayesian nmf algorithms for enforcing harmonicity and smoothness in polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA09)*, New Paltz, NY, oct 2009.

[6] S.A. Raczyński, N. Ono, and S. Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR07)*, Vienna, Austria, September 2007.

[7] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, mar 2007.