Adaptive harmonic decomposition using shift-invariant PLCA

Benoit Fuentes, Roland Badeau, Gael Richard

ADAPTIVE HARMONIC TIME-FREQUENCY DECOMPOSITION OF AUDIO USING SHIFT-INVARIANT PLCA

Benoit Fuentes, Roland Badeau, Gaël Richard

Institut Télécom,Télécom ParisTech, CNRS LTCI
46, rue Barrault - 75634 Paris Cedex 13 - France
benoit.fuentes@telecom-paristech.fr

ABSTRACT

This paper presents a new algorithm based on shift-invariant probabilistic latent component analysis that analyzes harmonic structures in an audio signal. Each note in a constant-Q transform is modeled as a weighted sum of narrowband parametric spectra, and a positive deconvolution is performed to obtain both pitch and timbre signature. The algorithm has been tested in a task of monopitch and multipitch estimation and shows very promising results.

Index Terms— Pitch estimation, probabilistic latent component analysis, harmonicity.

1. INTRODUCTION

Recently, many methods have been proposed in order to decompose the time-frequency representation of an audio signal into meaningful non-negative components. In the field of music signal processing, such a decomposition is useful to address various applications, such as pitch estimation, automatic transcription, or source separation. To perform this decomposition, the most popular technique is certainly the non-negative matrix factorization (NMF), initially designed for image processing and data mining applications [1], which has then been successfully applied to music analysis [2]. Since the basic NMF does not account for the particular characteristics of audio signals, some variants have been developed in order to model properties like the spectral harmonicity [3], or the time-variability of the fundamental frequency [4] and the spectral envelope [5] of each component. Probabilistic Latent Component Analysis (PLCA) [6] is a framework similar to NMF, which offers a convenient way of constraining the decomposition by introducing appropriate priors. The shift-invariant version of PLCA [7] was later introduced in order to obtain components sharing the same spectral shape but having different pitches.

The main drawback of the above-mentioned decomposition techniques is that they cannot handle components having both time-varying fundamental frequencies and spectral shapes. In this paper, we introduce a new method inspired from the shift-invariant PLCA, which overcomes this limitation. The paper is organized as follows. Section 2 summarizes some useful tools on which our adaptive harmonic decomposition technique relies. This technique is then introduced in section 3. It is evaluated in section 4 in the context of monopitch estimation, and compared to the classical YIN estimator [8]. Finally, the main conclusions of this work are drawn in section 5.

The research leading to this paper was supported by the Quaero Programme, funded by OSEO, French State agency for innovation.

2. SOME USEFUL TOOLS

In this section, we present different tools supporting our work.

2.1. Constant-Q transform

The presented system's input is a time-frequency representation of an audio signal based on the constant-Q transform (CQT) [9]. The CQT is a spectral representation of a temporal signal with a logarithmic frequency scale, contrary to the classical discrete Fourier transform. In this paper, the term CQT denotes the time-frequency representation obtained by applying this transformation at regular time points over an audio signal. With a logarithmic frequency scale, the spacing between two given partials of a harmonic note remains the same, regardless of the pitch of the note played. This feature can be verified on the top part of Fig. 2 where the CQT of three notes played by a harmonica is represented. Due to this characteristic, a change of pitch can be considered as a vertical shift of a spectral template. The mathematical model presented in the next section accounts for this shift-invariance. However, this consideration only stands approximately, since the different notes played by a single instrument may have different spectral shapes. The system we present in section 3 also takes avantage of the shift-invariant characteristic but can handle various spectral shapes.

2.2. Shift-invariant probabilistic latent component analysis

Probabilistic Latent Semantic Analysis [10] (PLSA) is a probabilistic tool for non-negative data analysis: it is used to decompose an observation as the sum of several independent sources. Shift-Invariant PLCA [7] (SI-PLCA) is an extension of PLSA, able to extract shifted patterns in multi-dimentional non-negative data. In our case, SI-PLCA consists in considering the CQT of an audio signal, V_{ft}, as the histogram of the sampling of N independent and identically distributed (i.i.d.) random variables (f_n, t_n), distributed according to the probability density function (PDF) $P(f,t)$. The way $P(f,t)$ is modeled induces the decomposition of V_{ft}. In [11], V_{ft} is decomposed as the weighted sum of several CQTs, representing monopitched sources. Each source is decomposed as a spectrum (called kernel distribution) convolved by time-frequency activations (called impulse distribution). $P(f,t)$ is then defined as:

$$\begin{aligned} P(f,t) &= \sum_{z=1}^{Z} P(z)P(f,t|z) \\ \text{with} \quad P(f,t|z) &= \sum_{i\in\mathbb{Z}} P_K(f-i|z)P_I(i,t|z) \end{aligned} \tag{1}$$

where $P_K(\mu|z)_{\mu\in[\![1;M]\!]}$ and $P_I(i,t|z)_{(i,t)\in\mathbb{Z}\times[\![1;T]\!]}$ are PDFs respectively representing the basic spectra and the time-frequency activations of each source, and $P(z)_{z\in[\![1;Z]\!]}$ is the probability of source z. This model allows defining a simple framework for the NMF problem and adding some interesting priors.

By means of the Expectation-Maximization (EM) algorithm, the parameters $P(z)$, $P_K(\mu|z)$ and $P_I(i,t|z)$ can be randomly initialized and iteratively updated until convergence to a local maximum of the likelihood. Update rules are expressed in [11], and the method which permits to calculate them in [6].

3. ADAPTIVE HARMONIC DECOMPOSITION

In [11], each source represents a specific instrument with its own basic spectrum and activations. It is a powerful method since no prior about the shape of instrument spectra is needed. Indeed, they can be strictly harmonic, or strongly inharmonic such as bell sounds or distorted guitars. But this model does not account for the possible amplitude variations of the partials of one instrument as a function of pitch and time. Instead, the model we propose supposes a perfect harmonicity of the notes but allows representing potential variations of the spectral envelopes. For the sake of clarity, this model is presented in the framework of monophonic signals analysis, but as it will be shown, it can easily be extended to polyphonic signals.

3.1. SI-PLCA model adaptation

As in [3], in order to account for variations of the spectral envelope, a note spectrum is decomposed as a weighted sum of narrowband basic spectra sharing the same pitch but with their energy concentrated at different frequency bands. In the SI-PLCA model, it means that for each source, the kernel distribution P_K is fixed and corresponds to one of those basic spectra, and the corresponding impulse distribution can be decomposed as $P_I(i,t|z) = P_I(i|t,z)P(t|z) = P_{Ih}(i|t)P(t|z)$, *i.e.* the random variable i does no longer depend on z. Besides, in order to consider the possible presence of colored noise in the audio signal, the last source is reserved for noise: its kernel distribution corresponds to a narrowband window and its impulse distribution $P_{In}(i|t,Z)$ is different from $P_{Ih}(i|t)$. Finally, model (1) becomes:

$$\begin{aligned} P(f,t) &= \sum_{z=1}^{Z} P(z) \sum_{i\in\mathbb{Z}} P_K(f-i|z) P_I(i,t|z) \\ &= \sum_{z=1}^{Z} P(z)P(t|z) \sum_{i\in\mathbb{Z}} P_K(f-i|z) P_I(i|t,z) \\ &= \sum_{z=1}^{Z} P(z,t) \sum_{i\in\mathbb{Z}} P_K(f-i|z) P_I(i|t,z) \end{aligned} \tag{2}$$

with:

$$\begin{aligned} &\bullet\ P_I(i|t,z) = P_{Ih}(i|t) \quad \text{if } z < Z \\ &\bullet\ P_I(i|t,Z) = P_{In}(i|t,Z). \end{aligned} \tag{3}$$

Fig. 1 shows an example of a CQT synthesized according to this model, with $Z = 3$ sources.

3.2. EM algorithm and updates rules

Given a CQT and a fixed set of kernel distributions $P_K(\mu,z)$, our purpose is to find the best set of discrete distributions $\Lambda = \{P_{Ih}(i|t), P_{In}(i|t,Z), P(z,t)\}$ from which useful information can be extracted. For instance, the pitch can be deduced for every t by selecting the value of i which maximizes $P_{Ih}(i|t)$. In our model, variables f and t are the observations, z and i are latent variables and the parameters are contained in the set Λ. The expectation step of the EM algorithm consists of calculating the log-likelihood function (we denote $\bar{x}$ the set of i.i.d. variables $\{x_n\}_{n=1\dots N}$):

$$L\left(\bar{f},\bar{t},\bar{z},\bar{i}\right) = \log\left(P\left(\bar{f},\bar{t},\bar{z},\bar{i}\right)\right) = \log\left(\prod_n P\left(f_n,t_n,z_n,i_n\right)\right)$$

and then calculating its conditional expectation given the observations and the parameters:

$$\begin{aligned} Q_\Lambda &= \mathbb{E}\left[L(\bar{f},\bar{t},\bar{z},\bar{i})|\bar{f};\bar{t};\Lambda\right] \\ &= \sum_{z,i}\sum_{n} P(i,z|f_n,t_n)\log\left(P\left(f_n,t_n,z,i\right)\right) \end{aligned}$$

As V_{ft} is modeled as the histogram of the observations, it is possible to change the summation over n by a summation over f and t, since the number of times the couple (f,t) is observed is known. Moreover, $P\left(f,t,z,i\right) = P(z,t)P_K(f-i|z)P_I(i|t,z)$. That leads to:

$$\begin{aligned} Q_\Lambda = \sum_{f,t,i,z} V_{ft}P(i,z|f,t)\Big[&\ln\left(P(z,t)\right) + \\ &\ln\left(P_K(f-i|z)\right) + \ln\left(P_I(i|t,z)\right)\Big] \end{aligned} \tag{4}$$

with, due to the Bayes' theorem (the notation $\hat{x}$ is used for bound variables):

$$P(i,z|f,t) = \frac{P(z,t)P_K(f-i|z)P_I(i|t,z)}{\sum_{\hat{z}}P(\hat{z},t)\sum_{\hat{i}}P_K(f-\hat{i}|\hat{z})P_I(\hat{i}|t,\hat{z})} \tag{5}$$

where $P_I(i|t,z)$ was defined in equation (3).

In the maximization step, Q_Λ is maximized with respect to (w.r.t.) the model parameters, under the constraint

$$\sum_i P_I(i|t,z) = \sum_{z,t} P(z,t) = \sum_\mu P_K(\mu|z) = 1.$$

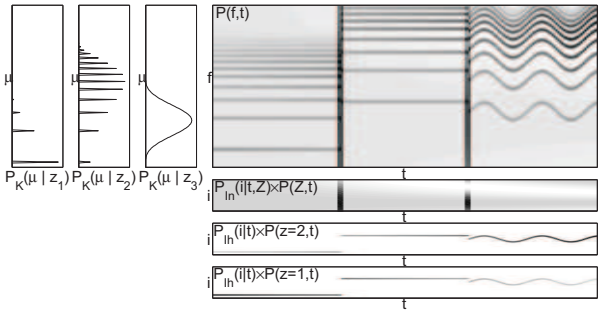

Fig. 1. A synthesized CQT and its decomposition. The designed model allows describing continuous variations of the fundamental frequency and of the spectral envelope of harmonic structures with a few number of kernels. The last kernel distribution is used for the description of smooth non-harmonic structures.

It leads to the following update rules:

$$P(z,t) = \frac{\sum_{f,i} V_{ft} P(i,z|f,t)}{\sum_{\hat{z},\hat{t},f,i} V_{f\hat{t}} P(i,\hat{z}|f,\hat{t})} \quad (6)$$

$$P_{Ih}(i|t) = \frac{\sum_{f,z<Z} V_{ft} P(i,z|f,t)}{\sum_{\hat{i},f,z<Z} V_{ft} P(\hat{i},z|f,t)} \quad (7)$$

$$P_{In}(i|t,Z) = \frac{\sum_{f} V_{ft} P(i,Z|f,t)}{\sum_{\hat{i},f} V_{ft} P(\hat{i},Z|f,t)}. \quad (8)$$

After initializing the model parameters, we iterate the above equations until convergence. Ideally, for a given t_0, the impulse distribution $P_{Ih}(i|t_0)$ would be an unimodal probability distribution, the pitch would be given by the value of its mode and the spectral envelope would be parametrized by the coefficients $\{P(z,t_0)\}_{z=1\dots Z}$. However, in absence of any constraint, the parameters do not necessarily converge to the desired solution. In practice, we notice that we can find maxima on $P_{Ih}(i|t_0)$ for i corresponding to the accurate pitch and all its higher harmonics. Since we would like to keep only the maximum of lowest frequency, we employ an asymmetric minimum variance prior as described in the next section.

3.3. Asymmetric minimum variance prior

Let $\boldsymbol{\theta^t}$ be the vector of coefficients $\theta_i^t = P_{Ih}(i|t)$ for a given t. In order to constrain $\boldsymbol{\theta^t}$ to be unimodal and to avoid upper-harmonic errors, we use an asymmetric minimum variance prior, forcing $\boldsymbol{\theta^t}$ to have both low variance and low mean. We first introduce an adequate measure, depending on a parameter $\alpha > 0$ which defines the strength of the asymmetry:

$$\begin{aligned}\mathrm{avar}_\alpha(\boldsymbol{\theta^t}) &= \sum_i \left(e^{\alpha i} - e^{\alpha \sum_{\hat{i}} \hat{i}\theta_{\hat{i}}^t}\right)\theta_i^t \\ &= \left(\sum_i e^{\alpha i}\theta_i^t\right) - e^{\alpha \sum_i i\theta_i^t} \quad \text{since } \sum_i \theta_i^t = 1. \quad (9)\end{aligned}$$

It can be proven, due to the strict convexity of the exponential function, that $\mathrm{avar}_\alpha(\boldsymbol{\theta^t}) \geq 0$ and $\mathrm{avar}_\alpha(\boldsymbol{\theta^t}) = 0 \Leftrightarrow \exists\, i_0 \mid \forall i, \theta_i = 1$ if $i = i_0$ and 0 otherwise. To bias $\mathrm{avar}_\alpha(\boldsymbol{\theta^t})$ during training, a prior distribution is introduced for the set of parameters Λ as $P(\Lambda) = (1/\sigma)\prod_t e^{-\beta\, \mathrm{avar}_\alpha(\boldsymbol{\theta^t})}$, where $\beta > 0$ is a parameter indicating the strength of the prior and σ a normalizing coefficient. The maximization step is now replaced by a maximum a posteriori (MAP) step, *i.e.* instead of maximizing Q_Λ we maximize $Q_\Lambda + \log(P(\Lambda))$ w.r.t. the model parameters and under the same constraints as before. Updates rules for $P(z,t)$ and $P_{In}(i|t,Z)$ do not change, but maximizing w.r.t. $P_{Ih}(i|t)$ leads to the equation:

$$\theta_i^t = \frac{\omega_i^t}{\beta\left(e^{\alpha i} - \alpha i e^{\alpha \sum_{\hat{i}} \hat{i}\theta_{\hat{i}}^t}\right) + \rho^t} \quad (10)$$

where $\omega_i^t = \sum_{f,z<Z} V_{ft} P(i,z|f,t)$, $\theta_i^t = P_{Ih}(i|t)$, and ρ^t is an additional coefficient which insures that $\sum_i \theta_i^t = 1$. There is no closed-form solution for θ_i^t, but numerical simulations showed that the fixed point Algorithm 1 always converges to a solution. Fig. 2 illustrates the effect of using the prior.

Algorithm 1 Fixed-point method

for all t **do**
  $\boldsymbol{\theta^t} \leftarrow \frac{\boldsymbol{\omega^t}}{\sum_i \omega_i^t}$
  loop
    · $m^t \leftarrow \sum_i i\theta_i^t$
    · $\forall i, c_i \leftarrow \beta\left(e^{\alpha i} - \alpha i e^{\alpha m^t}\right)$
    · find ρ^t such that $\sum_i \frac{\omega_i^t}{c_i+\rho^t} = 1$ and $\forall i, \frac{\omega_i^t}{c_i+\rho^t} \geq 0$ (there is a unique solution, that can be calculated with any numerical root finder algorithm)
    · $\forall i, \theta_i^t \leftarrow \frac{\omega_i^t}{c_i+\rho^t}$
  end loop
end for

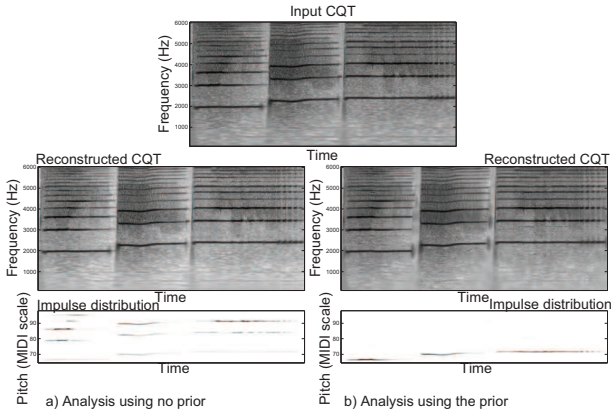

Fig. 2. Illustration of the use of the asymmetric minimum variance prior. If the reconstructed CQT remains almost unchanged, the impulse distribution becomes unimodal at each frame. The input signal corresponds to the recording of three notes played by a harmonica.

3.4. Polyphonic signals extension

So far, only the monophonic case has been considered, but the proposed model can easily be extended to polyphonic signals. To do so, we need to consider a polyphonic signal as a sum of monophonic signals, called channels. In the SI-PLCA model, it means that a new hidden variable c, corresponding to a channel, is introduced in the calculation of $P(f,t)$. All channels share the same kernel distributions and the same noise impulse distribution. Equation (2) then becomes:

$$P(f,t) = \sum_{c=1}^{C} P(c) \sum_{z=1}^{Z} P(z,t|c) \sum_{i\in\mathbb{Z}} P_K(f-i|z,c) P_I(i|t,z,c) \quad (11)$$

with

- $P_K(f-i|z,c) = P_K(f-i|z)$
- $P_I(i|t,z,c) = P_{Ih}(i|t,c) \quad \text{if } z < Z$
- $P_I(i|t,Z,c) = P_{In}(i|t,Z)$.

From this equation, similar update rules can be derived. Fig. 3 represents the CQT of J.S. Bach's first prelude played by a synthesizer (the notes are played with a slight vibrato) and the corresponding time-frequency activations found by our algorithm.

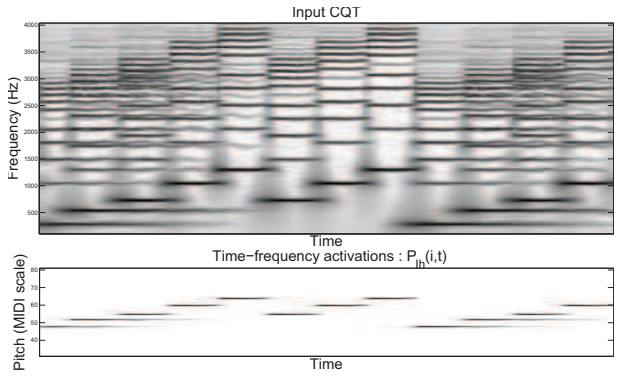

Fig. 3. Illustration of the algorithm for polyphonic signals with $C = 3$ channels. The time-frequency activations are defined as $P_{ih}(i,t) = \sum_{c=1}^{C} \sum_{z=1}^{Z-1} P_{Ih}(i|t,c)P(z,t|c)P(c)$.

4. EVALUATION

To check the relevance of the model, the algorithm [1] has been evaluated on a task of monopitch estimation, on 3307 isolated notes from the Iowa database [12]. This database includes recordings of several instruments, playing over their full range of notes, and with various play modes and nuances. For each signal, the CQT with 3 bins/semitones from $f = 27.5$Hz to $f = 6000$Hz and with a step size of 20ms is calculated. Then the CQT is analyzed by the algorithm using $Z = 15$ kernel distributions and the pitch is estimated for each time frame. To illustrate the shape of the kernels, Fig. 4 shows two of them. The method is compared with the YIN algorithm [8] using 100ms time frames (we used the code available on the authors' websites). Results are shown in Fig. 5.

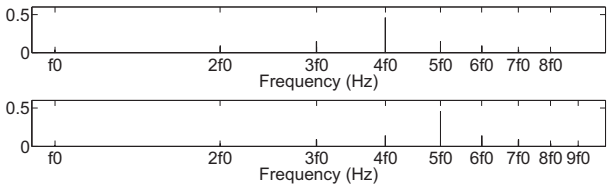

Fig. 4. Two of the parametric kernel distributions used in the algorithm. Partials are located at multiples of a fundamental frequency f_0. The maximum of the z^{th} kernel corresponds to the z^{th} harmonic of the modeled spectrum. Each kernel has a maximum of 9 partials.

5. CONCLUSION

We have presented an adaptive harmonic model for musical signal analysis that could be used in various applications, such as monopitch or multipitch estimation. A new prior to constrain impulse distributions to be unimodal has been introduced. This method is promising and in future work, we plan to include some temporal constraints and take the reverberation into account in order to improve the generality of the model. An other outlook is to implement a learning process for the kernel distributions, which should make this algorithm more robust to real musical signals.

[1]The Matlab code is available at http://perso.telecom-paristech.fr/~fuentes/shared_code/ICASSP_2011_fuentes.zip.

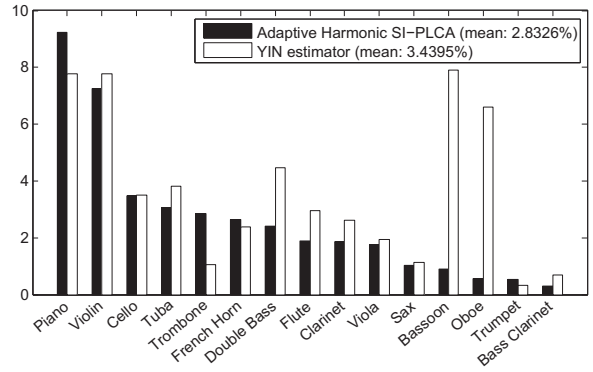

Fig. 5. Simulation results: averaged error rates for each instrument of the database in a task of monopitch estimation.

6. REFERENCES

[1] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negativity matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[2] C. Févotte, N. Bertin, and J-L Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[3] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[4] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *Proc. of DAFx-10*, Graz, Austria, September 2010, pp. 109–112.

[5] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non stationary audio events," *IEEE Transactions on Audio Speech and Language Processing*, pp. 109–112, to be published.

[6] M.V. Shashanka, *Latent variable framework for modeling and separating single-channel acoustic sources*, Ph.D. thesis, Boston University, Boston, MA, USA, August 2007.

[7] P. Smaragdis, B. Raj, and M.V. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proc. of ICASSP*, Las Vegas, Nevada, USA, April 2008, pp. 2069–2072.

[8] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.

[9] J. Brown, "Calculation of a constant Q spectral transform," *JASA*, vol. 89, no. 1, pp. 425–434, January 1991.

[10] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, Sweden, July 1999, pp. 289–296.

[11] G.J. Mysore and P. Smaragdis, "Relative pitch estimation of multiple instruments," in *Proc. of ICASSP*, Taipei, Taiwan, April 2009, pp. 313–316.

[12] "University of Iowa musical instrument sample database," http://theremin.music.uiowa.edu/index.html.