Probabilistic model for main melody extraction using constant-Q transform

Benoît Fuentes, Antoine Liutkus, Roland Badeau, Gaël Richard

PROBABILISTIC MODEL FOR MAIN MELODY EXTRACTION USING CONSTANT-Q TRANSFORM

Benoit Fuentes, Antoine Liutkus, Roland Badeau, Gaël Richard

Institut Télécom,Télécom ParisTech, CNRS LTCI
37-39, rue Dareau - 75014 Paris - France
benoit.fuentes@telecom-paristech.fr

ABSTRACT

Dimension reduction techniques such as Nonnegative Tensor Factorization are now classical for both source separation and estimation of multiple fundamental frequencies in audio mixtures. Still, few studies jointly addressed these tasks so far, mainly because separation is often based on the Short Term Fourier Transform (STFT) whereas recent music analysis algorithms are rather based on the Constant-Q Transform (CQT). The CQT is practical for pitch estimation because a pitch shift amounts to a translation of the CQT representation, whereas it produces a scaling of the STFT. Conversely, no simple inversion of the CQT was available until recently, preventing it from being used for source separation. Benefiting from advances both in the inversion of the CQT and in statistical modeling, we show how recent techniques designed for music analysis can also be used for source separation with encouraging results, thus opening the path to many crossovers between separation and analysis.

Index Terms— audio source separation, NTF, PLCA, CQT

1. INTRODUCTION

Source separation has gathered much interest recently and many methods were devised to separate the different *sources* signals mixed together into observable *mixtures*. In the case of audio processing, the sources are the different instruments playing together in a piece of music. The special case of the removal of *voice* in the recordings is of particular interest, because it allows many popular applications from voice enhancement or remixing to automatic karaoke. Unfortunately and because of the extreme diversity of vocal signals, this task also appears to be the most challenging. Many recent studies [1, 2] specifically focus on the separation of singing voice signals from polyphonic mixtures. Popular trends include techniques that decompose Time-Frequency (TF) representations of the mixtures using block models and that perform separation from those decompositions through generalized Wiener filtering [3, 4]. It is noticeable that most existing techniques for source separation make use of an *invertible* TF representation, such as the Short Term Fourier Transform, that permits recovering the waveforms of signals that were separated in the TF domain.

Apart from the separation of the different constitutive sounds from a mixture, other tasks of interest in Music Information Retrieval (MIR) also include the computation of *semantic information* from audio signals. One of the most prominent semantic information related to music is its *score*, i.e. the information concerning the time-varying *pitch* of the musical sounds composing the mixture. Even if the pitch of a sound is a complex notion whose definition is still the matter of some controversy, it has long been shown [5] to be strongly related to the notion of *fundamental frequency* f_0. Indeed, sounds that are perceived as *pitched* often have the remarkable property of being pseudo-periodic. Their pitch is then often considered equivalent on computational grounds to their —time varying— *fundamental frequency*. Hence, the challenging task of tracking multiple fundamental frequencies (also written multiple-f_0) over time has long gathered much attention in the MIR community. Methods ranging from deterministic sinusoidal modeling to fully probabilistic models can both be found in the literature. Among the latter, some recent techniques [6, 7] were proposed that focus on the analysis of the Constant-Q Transform (CQT) of the mixtures. Indeed, an interesting property of the CQT is that a change in fundamental frequency of a signal leads to a translation of its representation, whereas it leads to a more complex *scaling* of the STFT. Consequently, the CQT of a sound event following a complex pitch trajectory can efficiently be modeled as a single pattern merely translated over time. Hence, *translation invariant decompositions* [8], which permit to estimate such translated recurrent patterns in 2D representations, are adequate to model sound events. Unfortunately, there was no available inverse known for the CQT until recently, which prevented the methods derived for the tracking of multiple-f_0 to be used for source separation.

Still, many techniques from both audio source separation and multiple-f_0 estimation make use of the same underlying statistical models borrowed from the machine learning community. Among them, dimension reduction techniques [9, 8, 2, 7] are very popular to decompose the TF representation of mixtures into meaningful constitutive elements, whether it be for separation or for semantic analysis. The idea of jointly estimating the melody of the lead pitched instrument and separating it from the polyphonic mixtures has recently led to promising methods [2]. Still, only a few studies [9, 10] make use of translation invariant representations for the modeling of musical signals. However, even if exact inversion of the CQT is impossible, approximate and efficient algorithms are now publicly available [1] [11] that permit to obtain good performance.

In this study, we show how efficient models for music analysis on the CQT can directly be used for source separation. More specifically, we demonstrate that translation invariant models designed specifically to track the lead melody on CQT [6] can be used to recover separated waveforms. Doing so, we pursue seminal work in this direction by FITZGERALD [9], benefiting from both specialized and efficient models for vocal signals and recent statistical in-

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

[1]The invert CQT we used in our algorithm, implemented by J. Prado, is freely available at http://www.tsi.telecom-paristech.fr/aao/en/2011/06/06/inversible-cqt

sights concerning decompositions of the TF representations of non-stationary signals [3, 12].

The article is organized as follows: first, we present the statistical framework we use for modeling both the signal playing the lead melody and the musical background in section 2. In section 3, we focus on model parameters estimation and explain how the separation is performed. Finally, we present an evaluation of the proposed method in section 4 and conclude in section 5.

2. THE PROBABILISTIC MODEL FOR THE CQT OF AN AUDIO SIGNAL

As in [6] or [7], the model that we put forward is based on the Probabilistic Latent Component Analysis (PLCA) [13]. The absolute value of the CQT X_{ft} of an audio signal x, denoted $V_{ft} = |X_{ft}|$, is modeled as the histogram of N random variables $(f_n, t_n) \in \mathbb{Z} \times [\![1; T]\!]$, representing time-frequency bins, independently distributed according to the discrete probability distribution $P(f, t)$ (we suppose that $V_{ft} = 0$ for $f \notin [\![1; F]\!]$). The distribution $P(f, t)$ is structured according to the desired decomposition of V_{ft}, and the model parameters can be estimated by means of the Expectation-Maximization (EM) algorithm. Since our purpose is to separate the lead melody from the accompaniment, let $P(f, t)$ be equal to[2]:

$$P(f,t) = P(c_1)P_m(f,t|c_1) + P(c_2)P_a(f,t|c_2), \qquad (1)$$

where the probability distributions $P_m(f,t|c_1)_{(f,t)\in\mathbb{Z}\times[\![1;T]\!]}$ and $P_a(f,t|c_2)_{(f,t)\in\mathbb{Z}\times[\![1;T]\!]}$ respectively represent the CQTs of the two sources: the main melody (when the hidden variable c is equal to 1) and the accompaniment (when $c = 2$). $P(c)_{c=1,2}$ corresponds to the normalized global energy of each source. We now present the model of CQT for those two components.

2.1. The accompaniment model

For the accompaniment, the model used is the classic PLCA [13], equivalent to the Non-Negative Matrix Factorization [14]. Each column of a CQT $P_a(f,t|c_2)_f$ is modeled as a weighted sum of R basis spectra $P(f|r,c_2)_f$ as described in Fig. 1:

$$P_a(f,t|c_2) = \sum_r P(r,t|c_2)\, P(f|r,c_2). \qquad (2)$$

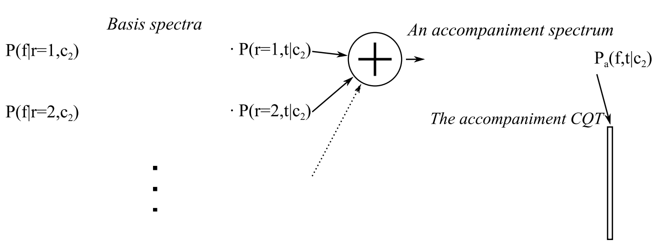

Fig. 1. Spectrum model for the accompaniment at time t: the classic PLCA.

[2] For the sake of simplicity, the notation c_k is used for $c = k$.

2.2. The lead melody model

In order to account for the non-stationary nature of many musical instruments (and especially the human voice), in terms of both pitch and spectral envelope, the model used for the lead melody is based on the one that we presented in [6], which allows simultaneously considering those two characteristics. At time t, the melody spectrum, represented by $P_m(f,t|c_1)$ is decomposed as a weighted sum of Z fixed narrow-band harmonic spectral kernels, denoted $P_K(\mu|z,c_1)_{(\mu,z)\in[\![1;F]\!]\times[\![1;Z]\!]}$, spectrally convolved by a time-frequency impulse distribution $P_I(i,t|c_1)_{(i,t)\in\mathbb{Z}\times[\![1;T]\!]}$ (when $c = 1$, f is then defined as the sum of the two random variables μ and i). The parameters have the following characteristics:

- all kernels share the same fundamental frequency, but have their energy concentrated at a given harmonic,
- the weights applied to the kernels, denoted $P(z|t,c_1)$, define the spectral envelope of the spectrum,
- each column of the impulse distribution $P_I(i,t|c_1)_i$ is unimodal, and its mode corresponds to the pitch of the melody.

Finally, the whole melody model can be written as:

$$P_m(f,t|c_1) = \sum_{i,z} P_I(i,t|c_1)P(z|t,c_1)P_K(f-i|z,c_1). \qquad (3)$$

This model is illustrated in Fig. 2.

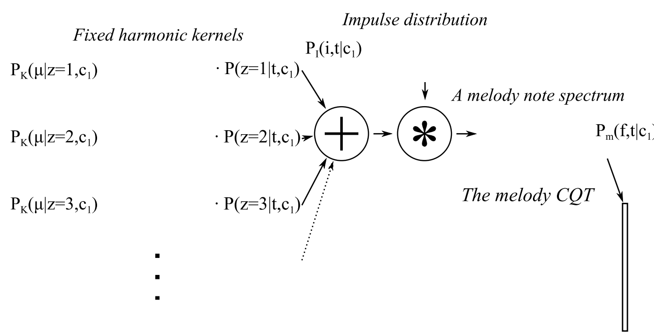

Fig. 2. Harmonic spectrum model for the main melody at time t. Each kernel has its main energy concentrated on a given harmonic (multiple of a reference fundamental frequency), and the rest of the energy is shared between adjacent partials.

3. PARAMETERS ESTIMATION AND ALGORITHM

3.1. EM algorithm and parameter updates

As described in [13], the EM algorithm defines update rules for the parameters, such that the log-likelihood L of the variables f_n and t_n observed via the histogram V_{ft}, which is proved equal to $L = \sum_{f,t} V_{ft} \ln(P(f,t))$, increases at every iteration.

First, the a posteriori distributions of the latent variables i, z, r, c are computed using Bayes' theorem, in the "expectation step":

$$P(i,z,c_1|f,t) = \frac{P(c_1)P_I(i,t|c_1)P(z|t,c_1)P_K(f-i|z,c_1)}{P(f,t)}, \qquad (4)$$

$$P(r,c_2|f,t) = \frac{P(c_2)P(r,t|c_2)P(f|r,c_2)}{P(f,t)}, \qquad (5)$$

$P(f,t)$ being defined by equations (1), (2) and (3).

Then the expectation of the log-likelihood of the complete data (including observed and latent variables) is maximized in the "maximization step":

$$P(c_1) \propto \sum_{f,t,z,i} V_{ft} P(i,z,c_1|f,t), \quad (6)$$

$$P_I(i,t|c_1) \propto \sum_{f,z} V_{ft} P(i,z,c_1|f,t), \quad (7)$$

$$P(z|t,c_1) \propto \sum_{f,i} V_{ft} P(i,z,c_1|f,t), \quad (8)$$

$$P(c_2) \propto \sum_{r,f,t} V_{ft} P(r,c_2|f,t), \quad (9)$$

$$P(r,t|c_2) \propto \sum_{f} V_{ft} P(r,c_2|f,t), \quad (10)$$

$$P(f|r,c_2) \propto \sum_{t} V_{ft} P(r,c_2|f,t). \quad (11)$$

The algorithm consists in initializing the parameters, then iterating equations (4), (5), the various update rules (equations (6), (7), (8), (9), (10) and (11)) and finally the normalization of all parameters so that the probabilities sum to one. Ideally, for a given time t, $P_I(i,t|c_1)_{i\in\mathbb{Z}}$ would be unimodal, the value of the mode would correspond to the pitch of the lead melody, and the coefficients $P(z|t,c_1)$ would describe its spectral envelope. This is however not guaranteed, since other notes from the accompaniment could be modeled by $P_m(f,t|c_1)$. Next section shows how to overcome this flaw.

3.2. Viterbi algorithm and second round of the EM algorithm

To ensure that each column of the impulse distribution $P_I(i,t|c_1)_i$ has a unique maximum, corresponding to the melody pitch, the same Viterbi algorithm used by Durrieu [2, p. 570] is applied on the estimated impulse distribution. The best pitch path (which makes a compromise between high energy and smooth trajectory) is found, and $P_I(i,t|c_1)$ is set to zero for the couples (i,t) which are further than one semi-tone from it. This step is illustrated in Fig. 3. The EM algorithm is then applied again for a few iterations in order to let parameters converge to a new solution.

3.3. Silence detection

The model presented above does not take possible silences in the melody into account. In order to avoid the presence of energy in the estimated melody source when the lead instrument or voice is shut down, we use a simple silence detector. The temporal energy signal of the estimated melody, defined as $E_m(t) = \sum_i P_I(i,t|c_1)$, is filtered with a $1/10$Hz cut-off frequency low-pass filter, and a threshold manually set at -12dB is applied. When the melody is considered off, we set $P_a(f,t|c_2) = P(f,t)$, which means that $P(c_2) = 1$ and $P(c_1) = 0$.

3.4. Time-frequency masking

Once all parameters have been estimated, one can proceed to the unmixing process by means of time-frequency masking. The time-frequency masks, M_m and M_a, which respectively correspond to the main melody and the accompaniment, are defined as:

$$M_m(f,t) = \frac{P(c_1)P_m(f,t|c_1)}{P(f,t)}, \quad (12)$$

$$M_a(f,t) = \frac{P(c_2)P_a(f,t|c_2)}{P(f,t)}. \quad (13)$$

The temporal signals of the two sources can then be estimated by applying the masks on the CQT X_{ft} of the input signal, and calculating the invert CQT:

$$\hat{s}_m = \mathrm{CQT}^{-1}(M_m(f,t)X_{ft}), \quad (14)$$

$$\hat{s}_a = \mathrm{CQT}^{-1}(M_a(f,t)X_{ft}). \quad (15)$$

Fig. 4 illustrates the result obtained by applying our algorithm to the CQT X_{ft} of an input mixture audio file. It shows the estimated CQT's of the two sources ($|M_m X_{ft}|$ and $|M_m X_{ft}|$).

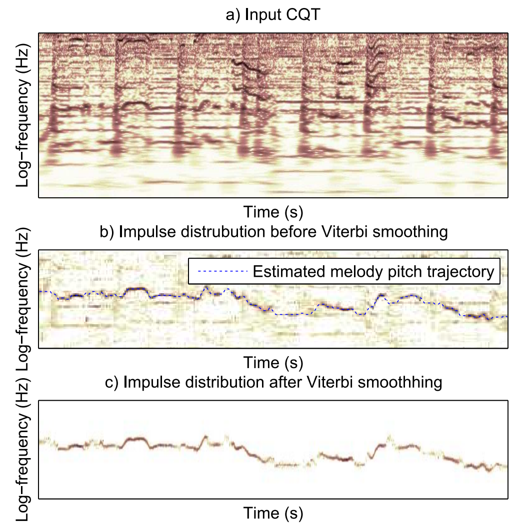

Fig. 3. Illustration of the Viterbi algorithm: the model parameters of the input CQT (a) are estimated in a first round, and the melody pitch trajectory is estimated from the impulse distribution (b). The bottom figure (c) shows the final estimated impulse distribution.

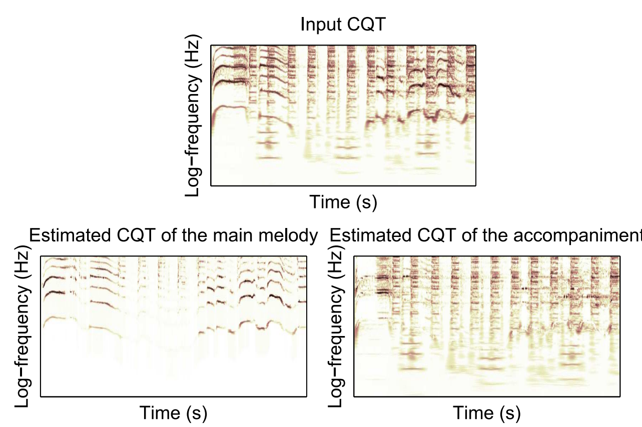

Fig. 4. Input CQT and estimated CQT of each source.

4. EVALUATION

The proposed separation system was tested on twelve monophonic excerpts, which last from 11 to 45 s (340 s in total), sampled at 44.1kHz from the Quaero [3] source separation corpus. The excerpts featured different kinds of music, including rock, reggae or bossa. For each of these files, the ground truth melody and accompaniment signals are known for evaluation, but the input signals in our system are the mixtures. All CQTs are calculated using 36 bins per octave and a step size of 4ms. After estimating the sources of each audio file, the quality of the result is quantified through the BSSEval toolbox [15], which gives three different metrics (in dB): the Source to Distortion Ratio (SDR), the Source to Artifact Ratio (SAR) and the Source to Interference Ratio (SIR). Whereas the SDR measures the global quality of the separation, the SIR and SAR respectively measure the amount of energy from the other sources and the amount of separation/reconstruction artifacts. In order to assess the quality of the proposed method, we compared it with the system proposed in [2] (using the code available on the author's website[4]). Furthermore, we evaluated the quality of the time-frequency filtering using CQT instead of classic STFT by calculating the results obtained by using the idealized (oracle) masks:

$$\hat{s}_m = \mathrm{TF}^{-1}\left(\frac{|\mathrm{TF}(s_m)|^2}{|\mathrm{TF}(s_m)|^2 + |\mathrm{TF}(s_a)|^2}\mathrm{TF}(s_m + s_a)\right), \quad (16)$$

$$\hat{s}_a = \mathrm{TF}^{-1}\left(\frac{|\mathrm{TF}(s_a)|^2}{|\mathrm{TF}(s_m)|^2 + |\mathrm{TF}(s_a)|^2}\mathrm{TF}(s_m + s_a)\right), \quad (17)$$

where the operator TF is either the STFT or the CQT, and s_m (resp. s_a) is the ground truth melody (resp. accompaniment) source. All results are shown in Fig. 5. We can see that oracle performances

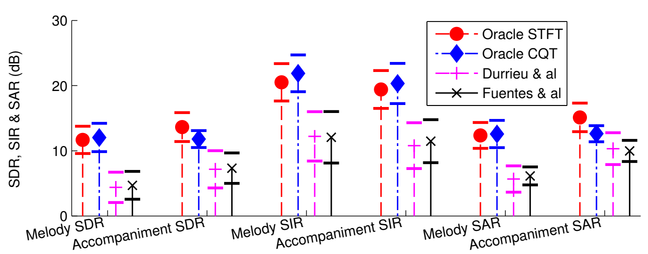

Fig. 5. SDR, SAR and SIR of the melody and accompaniment estimates using four different systems. The error bars represent the means (circle, diamond, plus and cross markers) plus/minus standard deviations (horizontal lines on each side) of BSSEval results. Higher values mean better separation.

with the CQT are very similar to the ones with the STFT, despite the fact that the CQT is only approximately invertible. We can also observe that the measures given by our algorithm are very closed to those given by Durrieu's method. This proves that our model is well adapted the task of melody extraction, and that using the CQT for source separation is relevant. Sound excerpts and a full implementation in Matlab of this separation technique are freely available on our website [5].

[3] http:///www.quaero.org

[4] http://www.durrieu.ch/research/jstsp2010.html

[5] http://www.tsi.telecom-paristech.fr/aao/en/2012/01/16/fuentes_icassp2012

5. CONCLUSION

In this study, we proposed a system that accurately separates the main melody from the accompaniment in a music audio file. The CQT of an audio signal is modeled as the sum of two CQTs, one for each source, and some analysis tools have been proposed in order to estimate both of them. Then the separation is performed by time-frequency masking. The very good results that we obtained show that the models designed for music analysis on the CQT can easily be applied to source separation. In the future, we would like to model rhythmic patterns, in order to separate the percussive part in addition to the lead melody and accompaniment in a musical excerpt. An additional interesting work would be to integrate the silence detection into the model, instead of performing it as a post-processing step.

6. REFERENCES

[1] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *Proc. of the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, September 2008.

[2] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.

[3] A.T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill, "Prior structures for Time-Frequency energy distributions," in *Proc. of WASPAA*, NY, USA, October 2007, pp. 151–154.

[4] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *Signal Processing, IEEE Transactions on*, vol. 59, no. 7, pp. 3155 –3167, July 2011.

[5] A.S. Bregman, *Auditory Scene Analysis, The perceptual Organization of Sound*, MIT Press, 1994.

[6] B. Fuentes, R. Badeau, and G. Richard, "Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA," in *Proc. of ICASSP*, Prague, Czech Republic, May 2011, pp. 401–404.

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

[8] M. Mørup, M. N. Schmidt, and L. K. Hansen, "Shift invariant sparse coding of image and music data," Tech. Rep., DTU Informatics, Technical University of Denmark, Lyngby, Denmark, 2008.

[9] D. Fitzgerald, M. Cranitch, and E. Coyle, "Sound source separation using shifted non-negative tensor factorisation," in *Proc. of ICASSP*, Toulouse, France, May 2006, vol. 5, pp. 653–656.

[10] D. Fitzgerald, M. Cranitch, and E. Coyle, "Resynthesis methods for sound source separation using shifted non-negative factorisation models," in *Proc. of ISSC*, Derry, Ireland, September 2007.

[11] C. Schoerkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. of the 7th Sound and Music Computing Conference*, Barcelona, Spain, July 2010.

[12] O. Dikmen and A. T. Cemgil, "Unsupervised single-channel source separation using Bayesian NMF," in *Proc. of WASPAA*, NY, USA, October 2009, pp. 93–96.

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

[14] 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.

[15] E. Vincent, C. Févotte, and R. Gribonval, "Performance measurement in blind audio source separation," *Audio, Speech and Language Processing, IEEE Trans. on*, vol. 14, no. 4, pp. 1462–1469, 2006.