# IMPROVING PLCA-BASED SCORE-INFORMED SOURCE SEPARATION WITH INVERTIBLE CONSTANT-Q TRANSFORMS

*J. Ganseman*[*]*, P. Scheunders*

IBBT-Visionlab
University of Antwerp
Universiteitsplein 1
2610 Wilrijk, Belgium

*S. Dixon*

Centre for Digital Music
Queen Mary University of London
Mile End Road
London E1 4NS, UK

## ABSTRACT

Probabilistic Latent Component Analysis is a widely adopted variant of Nonnegative Matrix Factorization for the purpose of single channel audio source separation. It has seen many extensions, including incorporation of prior information derived from music scores. Recent work on the invertibility of the Constant-Q Tranform make that a viable alternative to the Short-time Fourier Transform as underlying data representation. In this paper we assess several implementations for their usability in score-informed source separation. We show that results are comparable to, and in some cases better than, use of the STFT, and that exact transform invertibility is not a significant factor in this application.

***Index Terms***— PLCA, NMF, CQT, STFT, NSGT, BSS_EVAL, PEASS, score informed, source separation

## 1. INTRODUCTION

The constant-Q transform (CQT) [1] is well-suited to the task of music signal analysis. This classical CQT is not invertible, since a DC component cannot be calculated due to the logarithmic frequency spacing, and minimum and maximum frequency bands need to be defined outside which the signal will not be analyzed.

In recent work, several authors developed workarounds and optimizations to construct a CQT that is (approximately) invertible. We cover 3 implementations in this paper for which the source code is available from their respective authors. First, an approximately invertible CQT originating from Schörkhuber and Klapuri [2]. We also consider an alternative implementation by Prado [3]. Finally, a perfectly invertible CQT has been developed recently by Velasco et al. [4], derived from the Non-stationary Gabor Transform, but then non-stationary in terms of frequency. We refer to these implementations as CQT-SK, CQT-P and NSGTF respectively.

Nonnegative Matrix Factorization (NMF) was first applied on spectrograms in [5]. Probabilistic Latent Component Analysis (PLCA) [6] provides a statistical interpretation of NMF using the KL-divergence - numerical equivalence was shown in [7]. In such a probabilistic framework, prior information can be intuitively incorporated through mixing prior probability distributions into the update equations. It has been used succesfully for extraction of a single source from audio given a sample "hummed" recording [8], or to simultaneously extract all instrument parts given the aligned and synthesized score [9]. In independent work, Virtanen et al [10] have incorporated prior Gamma distributions in NMF directly. For similar reasons, Kameoka [11] uses prior distributions to model temporal and harmonic constraints in his Bayesian Harmonic-Temporal Clustering method.

PLCA or versions thereof were applied to magnitude CQT data before, e.g. Benetos et al. [12] used the CQT with shift-invariant (convolutive) PLCA for polyphonic music transcription. Transcription does not require an inverse transform, but source separation does, so the previously mentioned developments on CQT invertibility now allow a wider range of applications. One of the first uses for the purpose of source separation was presented in Fuentes et al. [13], where Prado's CQT implementation [3] was used with an adapted version of PLCA for melody extraction.

Our goal here is to study whether a full score-informed source separation system as in [9] can benefit from using an invertible CQT as underlying signal representation. We briefly elaborate on the CQT implementations and the setup of our tests, describe the score-informed source separation algorithm used, and then proceed to present experimental results using the widely adopted metrics in the BSS_EVAL [14] and PEASS [15] toolkits for evaluation.

[*]The first author is funded by a Specialization Grant from IWT-Flanders, and performed the work while visiting the Centre for Digital Music, Queen Mary University of London. The author wishes to thank Emmanouil Benetos and Holger Kirchhoff for inspiring conversations and constructive comments.

## 2. INVERTIBLE CONSTANT-Q TRANSFORMS

### 2.1. A brief summary

The idea of the CQT resembles that of the wavelet transform, a comparison which is worked out in more detail in [2]. Wavelet transforms are usually applied with a factor 2 dilation (1 bin per octave) or more. For tonal music analysis, we want subdivisions of that that are a multiple of 12 to account for all notes in an equal-tempered musical scale, depending on the frequency detail required.

In [2] an optimal approximate inverse CQT is constructed by choosing the kernel atoms, hop sizes and windows for each octave in such a way that overlap-add reconstruction is maximally preserving. Prado [3] uses different filter designs and hop sizes, leading to a smaller coefficient grid. Velasco et al. [4] are able to go further: expanding the notion of orthonormal bases to frames in vector spaces, they use the mathematical property that invertible frame operators can be constructed from a frame and its dual frame as long as it conforms to the *frame condition*. It enables them to construct a perfectly invertible CQT. At the moment of writing, implementations are available from the authors [1].

### 2.2. Preparing the CQT for use with PLCA

The CQT is calculated at different temporal resolutions for different frequencies, resulting in the coefficients of the CQT not being in a rectangular matrix form. NMF or PLCA need rectangular (fixed time resolution) matrices to work with. Whereas [16] notes this too, but proceeds to adapt shift-invariant PLCA to work on STFT data, we increase the temporal resolution of the lower frequencies to match that of the highest frequency bin so we can continue with the CQT.

The CQT-P and CQT-SK implementations contain functions to retrieve such a "rasterized" CQT. With NSGTF a rectangular grid can be obtained by appropriately zero-padding the set of (zero-phase) analysis windows that is initially generated, such that all window lengths equal that of the largest window. This technique is explained in the chapter on spectral interpolation in [17]. The NSGTF implementation processes the Nyquist frequency separately with a window the length of the signal itself. To enable NMF/PLCA decomposition, we discard that band from the forward transform. In the resynthesis phase, we put its coefficients to 0, as that band contains little or no information for audible sounds. We keep the DC bin (0 frequency) as it is also present in the STFT matrix.

[1] CQT-SK: http://www.elec.qmul.ac.uk/people/anssik/cqt/
CQT-P : http://perso.telecom-paristech.fr/~grichard/CQT/demo_cqt_inv.zip
NSGT : http://www.univie.ac.at/nonstatgab/cqt/

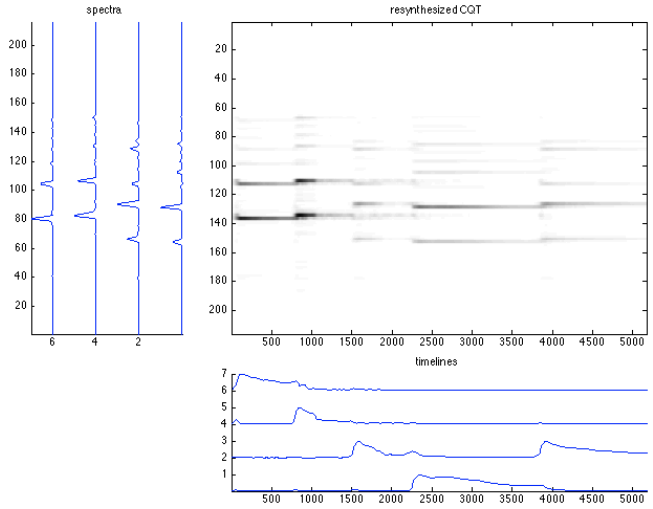


**Fig. 1**. PLCA decomposition of a constant-Q spectrogram

An example of applying PLCA to a CQT representation of a sound is shown in figure 1. It shows the CQT of a short excerpt with 5 piano notes, 4 pitches, also used in [5]. On the left hand side is the dictionary, below are the activations for each dictionary element, resulting from PLCA decomposition.

CQT-P and CQT-SK make a few different implementation choices. Notably, CQT-P uses the Parks-McClellan algorithm to generate a lowpass Chebyshev filter, while CQT-SK relies on a Butterworth lowpass filter. All algorithms calculate a different step-size according to the minimum/maximum frequency parameters and bins per octave that have been given.

Requesting 9 octaves down from 14700Hz at 48 bins per octave and a sampling rate of 44100Hz, using default parameters otherwise, and given the example of figure 1, the matrix sizes and reconstruction errors of the 3 implementations are given in table 1. Note that the original NSGTF implements exact inversion, but in order to obtain a rectangular matrix we needed to leave out the Nyquist component. At resynthesis we put a vector of zeroes in its place, hence we introduced reconstruction error.

| | CQT-P | CQT-SK | NSGTF |
|---|---|---|---|
| Matrix size | $432 \times 826$ | $432 \times 6823$ | $434 \times 1407$ |
| Rec. error | $3.34 \times 10^{-1}$ | $2.09 \times 10^{-2}$ | $1.22 \times 10^{-3}$ |

**Table 1**. Matrix size after forward CQT transforms with the 3 algorithms, and relative reconstruction error after subsequent inverse transform, of the example in figure 1. For reference, a 1024-point STFT with 75% overlap results in a $513 \times 569$ matrix.

| | | BSS_EVAL 3.0 | | | PEASS 2.0 | | | |
|---|---|---|---|---|---|---|---|---|
| | method | SDR (dB) | SIR (dB) | SAR (dB) | OPS (%) | TPS (%) | IPS (%) | APS (%) |
| piano | STFT | $7.69_{\pm 0.45}$ | $10.08_{\pm 0.64}$ | $11.85_{\pm 0.23}$ | $\mathbf{26.26}_{\pm 1.67}$ | $56.14_{\pm 3.31}$ | $24.15_{\pm 3.02}$ | $53.56_{\pm 2.20}$ |
| | CQT-P | $6.24_{\pm 0.37}$ | $11.25_{\pm 0.68}$ | $8.22_{\pm 0.24}$ | $24.72_{\pm 1.31}$ | $\mathbf{57.03}_{\pm 3.29}$ | $\mathbf{33.17}_{\pm 3.91}$ | $52.84_{\pm 1.46}$ |
| | CQT-SK | $\mathbf{10.46}_{\pm 0.62}$ | $14.67_{\pm 0.85}$ | $\mathbf{12.69}_{\pm 0.53}$ | $18.82_{\pm 2.27}$ | $37.67_{\pm 3.58}$ | $15.03_{\pm 3.22}$ | $57.35_{\pm 2.23}$ |
| | NSGTF | $9.94_{\pm 0.62}$ | $\mathbf{14.68}_{\pm 0.92}$ | $11.88_{\pm 0.52}$ | $14.75_{\pm 2.79}$ | $41.95_{\pm 4.51}$ | $11.72_{\pm 2.55}$ | $\mathbf{68.58}_{\pm 2.52}$ |
| violin | STFT | $11.51_{\pm 0.34}$ | $18.49_{\pm 0.46}$ | $12.55_{\pm 0.35}$ | $\mathbf{46.60}_{\pm 1.33}$ | $60.27_{\pm 1.35}$ | $\mathbf{67.27}_{\pm 1.83}$ | $51.20_{\pm 2.28}$ |
| | CQT-P | $12.27_{\pm 0.43}$ | $\mathbf{21.41}_{\pm 0.55}$ | $12.87_{\pm 0.47}$ | $45.31_{\pm 1.24}$ | $64.61_{\pm 2.07}$ | $61.62_{\pm 2.28}$ | $53.49_{\pm 1.35}$ |
| | CQT-SK | $\mathbf{13.31}_{\pm 0.57}$ | $19.98_{\pm 0.62}$ | $\mathbf{14.42}_{\pm 0.64}$ | $34.06_{\pm 4.59}$ | $62.91_{\pm 2.17}$ | $52.84_{\pm 3.22}$ | $60.04_{\pm 1.72}$ |
| | NSGTF | $12.55_{\pm 0.55}$ | $19.24_{\pm 0.88}$ | $13.67_{\pm 0.54}$ | $37.62_{\pm 5.63}$ | $\mathbf{65.94}_{\pm 1.67}$ | $55.81_{\pm 4.08}$ | $\mathbf{63.17}_{\pm 2.13}$ |

**Table 2**. Quality of source separation of a synthetic piano and violin example. Showing mean BSS_EVAL and PEASS metrics calculated over 120 runs, standard deviation shown in subscript. Higher is better for all scores, best scores are boldfaced.

## 3. SCORE-INFORMED PLCA

### 3.1. Probabilistic Latent Component Analysis

PLCA [6] can be seen as a probabilistic interpretation of NMF minimizing a Kullback-Leibler (KL) divergence. The elements of the resulting dictionary $W$ and activation matrix $H$ are scaled such that they sum to 1, due to them being modeled as multinomial probability distributions. Additionally a diagonal "gain" matrix $Z$ is introduced, modeling the energy contribution of each component (or dictionary element) to the mixture separately. The decomposition of a magnitude spectrogram $V$ can then be written as:

$$V \approx WZH \tag{1}$$

or in terms of probability distributions where $z$ is the component index:

$$V \approx \sum_z P(z)P(w|z)P(h|z) \tag{2}$$

PLCA employs an Expectation-Maximization algorithm to obtain a maximum-likelihood estimate of the model parameters. Prior information is introduced using Dirichlet prior distributions. We refer to [8] for the detailed derivation, and just mention the iterative update equations resulting from the model here:

$$P(z|w,h) = \frac{P(z)P(w|z)P(h|z)}{\sum_{z'} P(z')P(w|z')P(h|z')} \tag{3}$$

$$P(w|z) = \frac{\sum_h V_{w,h}P(z|w,h) + \kappa_z \alpha(w|z)}{\sum_{w'}\sum_h V_{w',h}P(z|w',h) + \kappa_z \alpha(w'|z)} \tag{4}$$

$$P(h|z) = \frac{\sum_w V_{w,h}P(z|w,h) + \mu_z \alpha(h|z)}{\sum_w \sum_{h'} V_{w,h'}P(z|w,h') + \mu_z \alpha(h'|z)} \tag{5}$$

$$P(z) = \frac{\sum_w \sum_h V_{w,h}P(z|w,h)}{\sum_{z'}\sum_w \sum_h V_{w,h}P(z'|w,h)} \tag{6}$$

where the prior distributions, characterized by their hyperparameters $\alpha(w|z)$ and $\alpha(h|z)$ for the dictionary and activation priors respectively, are blended into the update equations with respective weights $\kappa_z$ and $\mu_z$. The denominators are merely normalization factors.

### 3.2. Score information and resynthesis

The introduction of prior distributions transforms the learning problem from an unsupervised to a semi-supervised one. Inspired by [8], we chose in prior work [9] to separately synthesize the instruments from an aligned score, then learn components from that to act as prior hyperparameters. When source-specific priors are applied to mutually disjoint subsets of dictionary elements in the mixture analysis, the decomposition converges towards a solution where those subsets will each model one of the sources. Alternatively, in [18] and [19] the score is used to generate a binary mask which forces parameters to be updated using only specific parts of the spectrogram.

At resynthesis, we make one change to the system in [9] here: instead of directly resynthesizing the estimated sources, we use the normalized estimates as a Wiener filter (a time-frequency mask) on the original transform, as in [19]. This decomposes the original spectrogram according to the estimates in a minimum mean squared error sense.

## 4. EXPERIMENTAL RESULTS

For this and all further experiments, the following parameters were used. All transforms use Hann windowing. A 1024-point STFT was calculated with an overlap between frames of 75%. For CQT-SK, CQT-P and NSGTF, we define the minimum and maximum frequency bin centers at $28.7Hz$ and $14700Hz$ respectively, spanning 9 octaves at 48 bins per octave. The relative weight of priors to data ($\kappa_z$ and $\mu_z$) was kept at 3 to 1 in the first 25 iterations, after which PLCA was left to converge for another 25 iterations without priors.

### 4.1. Test on synthetic perfectly aligned data

We test first on a 10-second synthetic piano and violin mixture, prepared using 2 different synthesizers. One of the mixes serves as original mix, the sources made with a the other synthesizer serve as priors. This example data is part of a larger synthetic dataset created for the purpose of testing

| | | BSS_EVAL 3.0 | | | PEASS 2.0 | | | |
|---|---|---|---|---|---|---|---|---|
| | method | SDR (dB) | SIR (dB) | SAR (dB) | OPS (%) | TPS (%) | IPS (%) | APS (%) |
| bassoon | STFT | $3.29_{\pm 0.11}$ | $7.60_{\pm 0.15}$ | $\mathbf{6.00}_{\pm 0.09}$ | $\mathbf{32.91}_{\pm 0.38}$ | $\mathbf{39.26}_{\pm 1.00}$ | $50.06_{\pm 1.03}$ | $\mathbf{41.65}_{\pm 0.53}$ |
| | CQT-P | $4.57_{\pm 0.14}$ | $15.47_{\pm 0.36}$ | $5.06_{\pm 0.12}$ | $30.85_{\pm 0.43}$ | $31.69_{\pm 0.99}$ | $\mathbf{50.73}_{\pm 1.41}$ | $37.49_{\pm 0.58}$ |
| | CQT-SK | $\mathbf{4.73}_{\pm 0.17}$ | $14.98_{\pm 0.39}$ | $5.30_{\pm 0.14}$ | $30.75_{\pm 0.41}$ | $29.53_{\pm 0.99}$ | $49.71_{\pm 0.96}$ | $37.08_{\pm 0.57}$ |
| | NSGTF | $4.04_{\pm 0.10}$ | $\mathbf{15.61}_{\pm 0.25}$ | $4.47_{\pm 0.09}$ | $30.23_{\pm 0.53}$ | $30.32_{\pm 0.86}$ | $45.10_{\pm 1.34}$ | $36.59_{\pm 0.67}$ |
| clarinet | STFT | $3.39_{\pm 0.10}$ | $8.06_{\pm 0.14}$ | $5.83_{\pm 0.10}$ | $\mathbf{21.68}_{\pm 0.53}$ | $22.62_{\pm 0.70}$ | $\mathbf{16.54}_{\pm 0.72}$ | $40.32_{\pm 0.79}$ |
| | CQT-P | $3.45_{\pm 0.20}$ | $8.90_{\pm 0.26}$ | $5.44_{\pm 0.17}$ | $19.16_{\pm 0.63}$ | $\mathbf{22.98}_{\pm 0.83}$ | $13.28_{\pm 0.84}$ | $\mathbf{40.52}_{\pm 0.98}$ |
| | CQT-SK | $\mathbf{5.35}_{\pm 0.14}$ | $\mathbf{11.28}_{\pm 0.19}$ | $\mathbf{6.94}_{\pm 0.13}$ | $20.76_{\pm 0.53}$ | $21.07_{\pm 0.73}$ | $13.46_{\pm 0.71}$ | $38.48_{\pm 1.21}$ |
| | NSGTF | $3.51_{\pm 0.18}$ | $9.24_{\pm 0.32}$ | $5.35_{\pm 0.14}$ | $19.60_{\pm 0.47}$ | $21.13_{\pm 0.75}$ | $13.71_{\pm 0.55}$ | $36.98_{\pm 0.88}$ |
| flute | STFT | $6.71_{\pm 0.12}$ | $\mathbf{15.22}_{\pm 0.18}$ | $7.50_{\pm 0.12}$ | $27.65_{\pm 0.49}$ | $31.28_{\pm 1.17}$ | $\mathbf{48.52}_{\pm 1.05}$ | $37.46_{\pm 0.74}$ |
| | CQT-P | $6.59_{\pm 0.12}$ | $14.50_{\pm 0.13}$ | $7.51_{\pm 0.13}$ | $27.01_{\pm 0.56}$ | $45.45_{\pm 1.13}$ | $48.09_{\pm 0.87}$ | $44.48_{\pm 0.45}$ |
| | CQT-SK | $\mathbf{7.13}_{\pm 0.16}$ | $14.73_{\pm 0.16}$ | $\mathbf{8.10}_{\pm 0.16}$ | $26.55_{\pm 0.42}$ | $40.46_{\pm 0.90}$ | $42.71_{\pm 1.15}$ | $44.05_{\pm 0.57}$ |
| | NSGTF | $6.88_{\pm 0.14}$ | $14.11_{\pm 0.21}$ | $7.95_{\pm 0.14}$ | $\mathbf{29.68}_{\pm 0.43}$ | $\mathbf{52.79}_{\pm 0.97}$ | $39.13_{\pm 0.98}$ | $\mathbf{51.45}_{\pm 0.35}$ |
| horn | STFT | $3.17_{\pm 0.11}$ | $9.77_{\pm 0.16}$ | $4.68_{\pm 0.09}$ | $17.09_{\pm 0.71}$ | $25.45_{\pm 1.10}$ | $\mathbf{18.23}_{\pm 1.01}$ | $45.38_{\pm 0.63}$ |
| | CQT-P | $5.96_{\pm 0.19}$ | $11.73_{\pm 0.37}$ | $\mathbf{7.58}_{\pm 0.16}$ | $16.31_{\pm 0.77}$ | $35.46_{\pm 1.83}$ | $11.70_{\pm 0.89}$ | $53.82_{\pm 0.57}$ |
| | CQT-SK | $5.55_{\pm 0.23}$ | $\mathbf{12.64}_{\pm 0.26}$ | $6.72_{\pm 0.22}$ | $17.38_{\pm 0.89}$ | $27.93_{\pm 1.36}$ | $11.55_{\pm 0.94}$ | $51.83_{\pm 0.68}$ |
| | NSGTF | $\mathbf{6.03}_{\pm 0.11}$ | $12.59_{\pm 0.19}$ | $7.35_{\pm 0.10}$ | $\mathbf{19.15}_{\pm 0.63}$ | $\mathbf{38.99}_{\pm 1.37}$ | $10.54_{\pm 0.76}$ | $\mathbf{57.05}_{\pm 0.50}$ |
| oboe | STFT | $-4.25_{\pm 0.08}$ | $-0.58_{\pm 0.09}$ | $1.50_{\pm 0.09}$ | $\mathbf{20.88}_{\pm 0.56}$ | $34.13_{\pm 0.76}$ | $\mathbf{16.04}_{\pm 0.74}$ | $52.00_{\pm 0.42}$ |
| | CQT-P | $\mathbf{-3.00}_{\pm 0.15}$ | $\mathbf{0.97}_{\pm 0.18}$ | $\mathbf{1.77}_{\pm 0.09}$ | $\mathbf{20.88}_{\pm 0.33}$ | $\mathbf{40.01}_{\pm 1.41}$ | $15.03_{\pm 0.68}$ | $\mathbf{55.13}_{\pm 0.38}$ |
| | CQT-SK | $-3.98_{\pm 0.11}$ | $-0.37_{\pm 0.14}$ | $1.69_{\pm 0.08}$ | $19.86_{\pm 0.41}$ | $31.70_{\pm 1.27}$ | $12.38_{\pm 0.61}$ | $53.46_{\pm 0.49}$ |
| | NSGTF | $-4.74_{\pm 0.14}$ | $-1.22_{\pm 0.16}$ | $1.48_{\pm 0.12}$ | $20.05_{\pm 0.47}$ | $36.12_{\pm 1.13}$ | $12.30_{\pm 0.60}$ | $55.44_{\pm 0.41}$ |

**Table 3**. Quality of source separation results of a woodwind quintet. Showing mean BSS_EVAL and PEASS metrics calculated over 45 runs, standard deviation shown in subscript. Higher is better for all scores, best scores shown boldfaced.

score-informed source separation algorithms and alignment algorithms [20]. Testing with synthesized data enables us to ignore a component of score-informed source separation systems that can heavily influence results: the alignment of score to audio, which is a research problem of its own. We used 25 components per source here, and analyzed the entire spectrogram at once. The test was repeated 120 times.

Table 2 shows the resulting averaged BSS_EVAL and PEASS metrics and standard deviations. CQT-SK outperforms the other data representations in most BSS_EVAL metrics, but this comes at the cost of using much more memory as table 1 made clear. NSGTF still scores better in the BSS_EVAL metrics than the STFT. Looking at the PEASS metrics however, the situation is rather different, with the STFT gaining the highest OPS (Overall Perceptual Score). Remark also that NSGTF-extracted piano boasts the highest Signal-to-Interference Ratio (SIR), which measures the leakage of other sources into the extracted source, but the Interference-related Perceptual Score (IPS) is the lowest.

### 4.2. Test on real-world manually aligned data

Results with a 45 second recording of a real woodwind quintet, originating from the MIREX 2007 evaluation and also used in [12], are shown in table 3. The audio was processed in 2 second blocks, and a rather high 20 components per source were sought in each block in the separation process. Mean BSS_EVAL and PEASS scores and standard deviations for the 5 constituent sources are shown, computed over 45 runs of the algorithm.

Here, CQT-SK has an edge over the others in most BSS_EVAL metrics for clarinet and flute. NSGTF does better in most PEASS scores for flute and horn. For the lower-pitched instruments bassoon and horn, most STFT BSS_EVAL scores are noticeably lower. Remarkable is also that the low SDR and SIR scores for the STFT-extracted bassoon appear in conjunction with relatively higher perceptual scores. The oboe is difficult to extract with our relatively unconstrained PLCA-based method as it doubles the clarinet.

## 5. CONCLUSIONS AND FUTURE DIRECTIONS

From the limited data that we have, we need to draw conclusions with the necessary caveats. One of the main problems in score-informed source separation remains the lack of fully annotated multitrack recordings that can serve as ground truth. In our examples it seems slightly beneficial to use a Constant-Q Transform for PLCA-based source separation. Given the BSS_EVAL metrics in table 3 we suspect a possible correlation between improved extraction of lower-pitched instruments, and the CQT emphasis on the lower frequencies as compared to the STFT. However, more testing with a diverse dataset is necessary to confirm or reject this hypothesis.

Overall, CQT-SK tends to perform better in BSS_EVAL metrics, but it never gets the highest PEASS scores and has a large memory footprint. NSGTF does better in that respect and boasts the smallest reconstruction error, but that seems not to be a major factor in this application. The STFT still provides the most compact representation. Considering the metrics, we find that high BSS_EVAL scores do not necessarily imply high PEASS scores, and PEASS scores fluctuate more than the BSS_EVAL scores. During experimentation we found the metrics to be sensitive to the gain of the extracted sources. In practice one might want to choose a specific transform depending on the source one tries to extract and the metric that is the most important for the application.

Listening to the results, audible differences can be heard e.g. in the nature of the leakage from other sources into each extracted source: STFT-based extraction seems to contain more isolated bursts with a fluttering quality to it, while in CQT-based extraction this kind of interference seems to occur in larger, smoother patches. We invite readers to judge these and other qualitative differences for themselves on *http://www.eecs.qmul.ac.uk/˜jga/eusipco2012.html*. MATLAB source code is available on the same page.

Many parameters that may affect results remain untested here: window used, number of components, number of iterations, ... For improvements in the score-informed source separation itself, including additional constraints derived from symbolic data as in [19] has shown significant benefits [21]. Another logical next step to take is using a convolutive or shift-invariant NMF method [7] [12], which is well-suited to work on the CQT and can reduce the number of components.

## 6. REFERENCES


[1] J.C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America (JASA)*, vol. 89, no. 1, pp. 425–434, January 1991.

[2] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. 7th Sound and Music Computing Conference (SMC 2010)*, Barcelona, Spain, July 2010.

[3] J. Prado, "Calcul rapide de la transformée à Q constant," Tech. Rep. 2011D006, Telecom-Paristech, Paris, France, May 2011.

[4] G. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible constant-Q transform with non-stationary Gabor frames," in *Proc. 14th Int. Conf. Digital Audio Effects (DAFX 2011)*, Paris, France, September 2011, pp. 93–99.

[5] 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)*, October 2003, pp. 177–180.

[6] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Neural Information Processing Systems (NIPS), Advances in models for acoustic processing workshop*, Whistler, BC, Canada, December 2006.

[7] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, vol. 2008.

[8] P. Smaragdis and G. Mysore, "Separation by 'humming': User-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2009.

[9] J. Ganseman, G.J. Mysore, P. Scheunders, and J.S. Abel, "Source separation by score synthesis," in *Proc. Int. Computer Music Conference (ICMC)*, New York, NY, USA, June 2010.

[10] T. Virtanen, A.T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, april 2008, pp. 1825 –1828.

[11] H. Kameoka, *Statistical Approach to Multipitch Analysis*, Ph.D. thesis, University of Tokyo, 2007.

[12] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a convolutive probabilistic model," in *Proc. 8th Sound and Music Computing Conference (SMC 2011)*, Padova, Italy, July 2011.

[13] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard, "Probabilistic model for main melody extraction using constant-Q transform," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012, pp. 5357–5360.

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

[15] V. Emiya, E. Vincent, N. Harlander, , and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[16] R. Hennequin, R. Badeau, and B. David, "Scale-invariant probabilistic latent component analysis," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oktober 2011, pp. 129–132.

[17] Julius O. Smith, *Spectral Audio Signal Processing*, http://ccrma.stanford.edu/˜jos/sasp/.

[18] Y. Han and C. Raphael, "Informed source separation of orchestra and soloist using masking and unmasking," in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, Makuhari, Japan, September 2010.

[19] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011, pp. 45–58.

[20] J. Ganseman, G.J. Mysore, P. Scheunders, and J.S. Abel, "Evaluation of a score-informed source separation system," in *Proc. 11th Int. Soc. Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, August 2010, pp. 219–224.

[21] J. Fritsch, J. Ganseman, and M. D. Plumbley, "A comparison of two different methods for score-informed source separation," in *Proc. 5th Int. Workshop on Machine Learning and Music*, Edinburgh, Scotland, UK, June 2012.